TØI report
692/2003

# Topics in meta-analysis

A literature survey

Peter Christensen



ISSN 0802-0175

ISBN 82-480-0396-5

Oslo, December 2003

# Preface

This report presents results from a literature survey of meta-analysis, concentrating on three topics, heterogeneity, publication bias and the assessment and the incorporation of the quality of studies included in meta-analyses.

This work is part of a Strategic Institute Program (SIP) on meta-analysis. It has been funded by the Research Council of Norway, with additional funding from the Institute of Transport Economics.

The literature survey has been carried out by Peter Christensen who has also written this report. Rune Elvik has been project manager. He has read and approved the report. Secretary Laila Aastorp Andersen has been responsible for the final layout of the report.

Many people, too numerous to mention, have responded to e-mail queries. Sue Duval, Stephen Sharp and Julian Higgins have been particularly helpful, both by answering questions and by sending material. Our sincere thanks to them.

Oslo, December 2003
Institute of Transport Economics

| | |
|---|---|
| *Sønneve Ølnes* | *Marika Kolbenstvedt* |
| Acting Managing Director | Head of Department |

*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 $\theta$ 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 $\sigma_i^2$ of the estimate is known an optimal estimate for the true effect $\theta$ 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 $\sigma_i^2$ 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 $\sigma_i^2$ is known. When the studies included are fairly large this is not a serious error. Because the parameter $\theta$ 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 $\theta_i$ in the ith study is randomly selected from a normal distribution of studies with mean $\theta$. 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 $Var(u)=\tau^2$.

$\sigma_i^2$ is the within-study variance and $\tau^2$ 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 $\tau^2$.

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 $\tau^2$ is less reasonable since the number of studies in a meta-analysis is usually fairly small. Normally the uncertainty of the estimate for $\tau^2$ will be considerable. All the same, the most common estimate for $\tau^2$ and the most common form for random effects analysis does not take the uncertainty of the estimate of $\tau^2$ 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 $\tau^2$ 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 $\tau^2$ 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, $\tau^2$ and the weighted mean $\hat{\theta}$ are computed simultaneously. Equations for the solution are given in Hardy and Thompson (1996).

$\hat{\theta} = \frac{\sum \frac{\hat{\theta}_i}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}}{\sum \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}}$ and $\hat{\tau}^2 = \frac{\sum \frac{(\hat{\theta}_i - \theta)^2 - \sigma_i^2}{(\hat{\sigma}_i^2 + \hat{\tau}^2)^2}}{\sum \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)^2}}$. 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 $\tau^2$, $\hat{\theta}$ can be solved for. Using this value in the other equation a new value of $\tau^2$ 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 $\tau^2$ as the value of $\theta$ varies and $\hat{\theta}(\tau^2)$ is the MLE of $\theta$ as $\tau^2$ varies. A confidence interval for $\tau^2$ 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 $\tau^2$ is only known after a regression analysis has been done. A method for estimating the regression equation and $\tau^2$ 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.

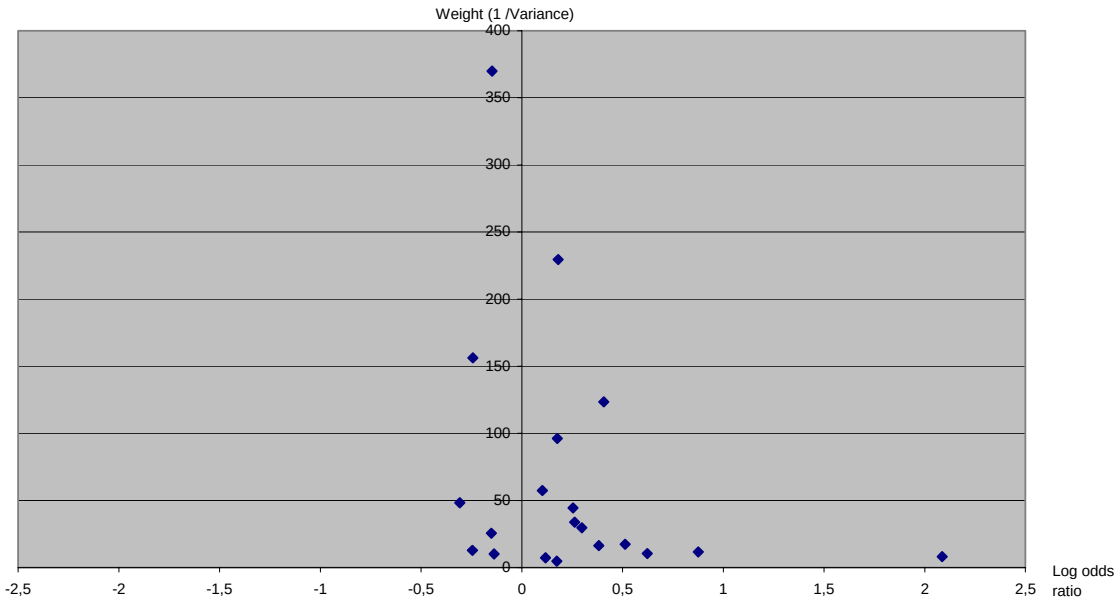



Figure 1. Funnel plot. Illustration based on arbitrary data. TØI report 692/2003.

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 $x_i = \frac{(x_i - \bar{x})}{\bar{v}_i^{\frac{1}{2}}}$,

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:

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

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.

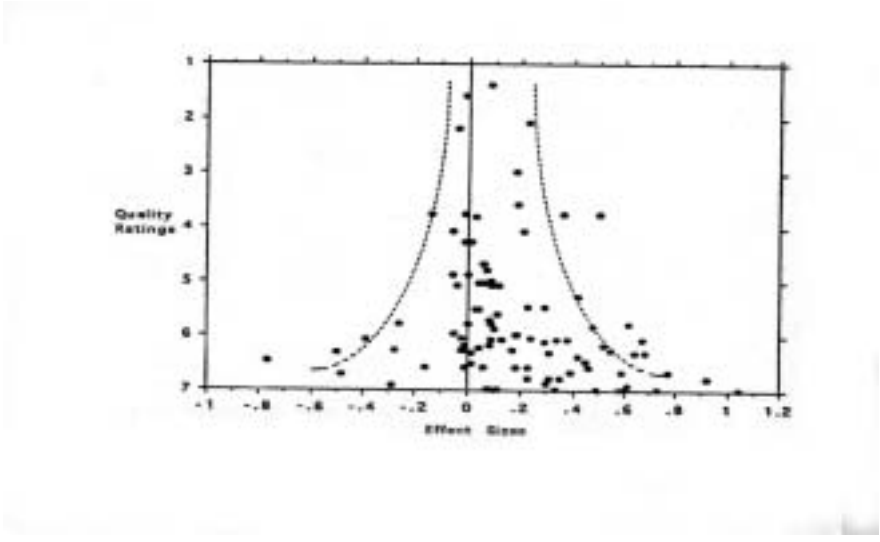



Figure 2. Scatter plot of the relation between recidivism effect sizes and ratings of methodological quality for studies of remedial programs for intoxicated drivers.
From Bangert-Drowns, Wells-Parker and Chevillard (1997).

### 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.

*TØI rapport 692/2003*
*Forfatter: Peter Christensen*
*Oslo 2003, 78 sider*

**Sammendrag:**

# Emner fra meta-analyse
## En litteraturstudie

Rapporten oppsummerer en litteraturstudie av metoder for meta-analyse. Formålet med litteraturstudien har vært å dekke felter innen meta-analyse hvor det fremdeles er uløste problemer eller hvor det er uenighet om hvordan meta-analyser bør gjennomføres. Feltene som har pekt seg ut er hvordan man skal behandle heterogene resultater, problemet med publikasjonsskjevhet og hvordan de enkelte undersøkelsers kvalitet skal vurderes og tas hensyn til.

## Hva er meta-analyse?

Ofte finner ulike undersøkelser forskjellige virkninger av samme tiltak. En meta-analyse er en beregning av gjennomsnittvirkningen av tiltaket med en systematisk metode for innhenting og bearbeiding av informasjon for sikre at data er så komplette som mulig og for å unngå feilkonklusjoner.

La $\theta$ være den sanne virkning av et tiltak. Estimatet for virkningen i den i'te undersøkelsen betegnes med $\hat{\theta}_i$ og variansen av estimatet med $\hat{\sigma}_i^2$. Hvis den sanne variansen $\sigma_i^2$ til estimatet er kjent er et optimalt estimat for den sanne virkning $\theta$ is gitt ved: $\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$, hvor k er antall undersøkelser og $w_i = \frac{1}{\sigma_i^2}$.

Siden $\sigma_i^2$ er ukjent brukes estimatet $\hat{\sigma}_i^2$ i vektene, $w_i = \frac{1}{\hat{\sigma}_i^2}$, og usikkerheten i $\hat{\sigma}_i^2$ blir normalt sett bort fra og vektene behandlet som om den sanne $\sigma_i^2$ er kjent. Dette er ikke noen alvorlig feil når undersøkelsene som er med i meta-analysen er forholdsvis store.

Fordi parameteren $\theta$ is antatt å være den samme (fixed) i alle undersøkelsene kalles dette "fixed effects" metoden.

## Heterogenitet

Fixed effects metoden beskrevet ovenfor antar at alle undersøkelser estimerer den samme sanne virkning, dvs at variasjonen i verdier mellom undersøkelser ikke er større enn at den kan forklares ved usikkerheten i de enkelte undersøkelsene. Variasjonen i resultatene kan imidlertid være for stor til at den kan forklares ved standardavviket til resultatene i undersøkelsene. Slik variasjon i resultatene kalles heterogenitet (Thompson and Sharp, 1999), eller mer presist statistisk heterogenitet.

*Rapporten kan bestilles fra:*
*Transportøkonomisk institutt, Postboks 6110 Etterstad, 0602 Oslo*
*Telefon: 22 57 38 00 Telefax: 22 57 02 90*

Om det er heterogenitet eller ikke er ikke alltid åpenbart. Det vil alltid være noe variasjon i resultatene og spørsmålet er om denne er for stor til å forklares med usikkerheten i de enkelte undersøkelsene. Det er derfor utviklet metoder for å påvise og måle heterogenitet.

Heterogenitet kan påvises både ved bruk av grafiske metoder og statistiske tester.

### Påvisning og måling av heterogenitet

Grafiske metoder for undersøke om det er heterogenitet er "skogdiagram" (forest plot), Galbraith diagram, en grafisk metode utviklet av Baujat et al og L´Abbé diagram . Disse er beskrevet i hovedteksten.

Standardtesten for heterogenitet er Cochrans Q-test. Test-observatoren er uttrykt ved:

$Q = \sum w_i (\hat{\theta}_i - \hat{\theta})^2$ hvor $\hat{\theta}$ er fixed effect gjennomsnittsestimatet av virkningen som er beskrevet ovenfor og $\hat{\theta}_i$ is den estimerte virkning i undersøkelse i. Q er tilnærmet kjikvadratfordelt.

Ifølge Higgins og Thompson (2002), er det velkjent at testen har liten styrke når det, som vanlig er, er få undersøkelser. Imidlertid er ikke andre tester noe bedre. Takkouche, Cadarso-Suarez, and Spiegelman (1999) undersøkte både type I feil og styrken til Cochrans Q test og fire andre tester for heterogenitet ved simulering. De konkluderte med at med hensyn på validitet, styrke og hvor enkelt det er å beregne testobservatoren, er Cochrans Q den beste. De dårlige nyhetene, som de uttrykker det, er at for de utvalgsstørrelser som er vanlig i epidemiologiske meta- analyser har ingen eksisterende tester akseptable styrke, hvis ikke heterogeniteten er betydelig.

Higgins and Thompson (2002) innfører tre ulike mål for å kvantifisere heterogenitet. Et av de to målene som de anbefaler er basert på Cohrans Q og er gitt ved:

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

### Meta-analyse når det er heterogenitet

Fixed effects metoden tar ikke hensyn til heterogenitet. Når det er heterogenitet vil derfor fixed effects metoden underestimere usikkerheten i det veiede gjennomsnittestimatet. En metode som tar hensyn til den økte usikkerheten er derfor nødvendig.

Dessuten vil ikke en beregning av det veide gjennomsnittestimatet når det er heterogenitet forklare heterogeniteten. Er den reell eller har den metodologiske forklaringer? Under hvilke forhold virker tiltaket? En forklaring på heterogeniteten er også av interesse. Det er derfor to ulike analyttiske angrepsmåter når det gjelder heterogenitet, å bare ta hensyn til den eller å forsøke å forklare den.

### Hvordan ta hensyn til heterogenitet

Random effects (RE) metoden tar hensyn til heterogenitet men forklarer ikke hvorfor den oppstår. RE metoden bygger på en antagelse om at den sanne virkning $\theta_i$ i den i'te undersøkelsen er tilfeldig valgt fra en normalfordeling av undersøkelser med gjennomsnittlig virkning $\theta$. Mer presist bygger RE metoden på følgende modell. Den observerte virkning $x_i$ i den i'te undersøkelsen er gitt ved:

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

$\sigma_i^2$ er variansen til resultatet i en undersøkelse og $\tau^2$ er variansen mellom undersøkelser. $Var(x_i)$ er nå gitt ved $Var(x_i)= \sigma_i^2 + \tau^2$ og vektene i fixed effects metoden er erstattet

med random effects vekter $w_i^* = \frac{1}{\sigma_i^2 + \tau^2}$. Bruk av disse vektene krever estimater for $\sigma_i^2$ og $\tau^2$.

Som nevnt tidligere for fixed effects metoden, så er $\sigma_i^2$ antatt å være kjent, dvs at estimatene $\hat{\sigma}_i^2$ for $\sigma_i^2$ er antatt å være uten usikkerhet. Denne antagelsen er rimelig hvis antall observasjoner eller tilfeller i undersøkelsene er stort. En tilsvarende antagelse for estimatet av $\tau^2$ er mindre rimelig siden antall undersøkelser i en meta-analyse vanligvis er ganske lite. Vanligvis vil usikkerheten i estimatet for $\tau^2$ være betydelig. Likevel tar ikke det mest vanlige estimatet for $\tau^2$ og den mest vanlige varianten av random effects analyse hensyn til usikkerheten til estimatet til $\tau^2$. I dette tilfellet er variansen til det veide gjennomsnittet gitt ved:

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

Det langt vanligste estimatet for $\tau^2$ er DerSimonian og Laird estimatet som bygger på Cochran's Q-test. Det er en moment-basert estimator som fås ved å sette den observerte verdien av Q lik forventningen og er gitt ved:

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

Når Q<k-1, settes $\tau_{DL}^2$ lik null. $\tau_{DL}^2$ er derfor et oppad avrundet estimat og følgelig forventningsskjevt.

Denne verdien av $\tau^2$ brukes i vektingsformelen over og det veide gjennomsnittet kan beregnes.

Alternativt kan beregnes en sannsynlighetsmaksimeringsestimator. I dette tilfellet beregnes $\tau^2$ og det veide gjennomsnittet $\hat{\theta}$ simultant. Ligningene for løsning er gitt i Hardy og Thompson (1996).

$$\hat{\theta} = \frac{\sum \frac{\hat{\theta}_i}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}}{\sum \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}} \quad \text{og} \quad \hat{\tau}^2 = \frac{\sum \frac{(\hat{\theta}_i - \theta)^2 - \sigma_i^2}{(\hat{\sigma}_i^2 + \hat{\tau}^2)^2}}{\sum \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)^2}}.$$ Det ses at ligningen for $\hat{\theta}$ er standard random effects Verdi for $\hat{\theta}$ gitt $\hat{\tau}^2$.

Ligningene løses ved iterasjon. Med en startverdi for $\tau^2$ kan finnes en løsning for $\hat{\theta}$. Brukes den verdien i den andre ligningen kan fås en ny verdi for $\tau^2$, osv. Hvis dette estimatet brukes uten å ta hensyn til usikkerheten i estimatet, vil det bli betegnet som enkel sannsynlighetsmaksimering.

Hardy og Thompson (1996) bruker profil sannsynlighetsmaksimering til å konstruere konfidensintervaller. Den følgende beskrivelsen er hentet fra deres artikkel.

Profil log-likelihood i to-parameter tilfellet er log-likelihood for én parameter gitt et estimat for den andre, dvs $l_1^*(\theta) = l(\theta, \hat{\tau}^2(\theta))$ og $l_2^*(\tau^2) = l(\hat{\theta}(\hat{\tau}^2), \tau^2))$. $\hat{\tau}^2(\theta)$ is sannsynlighetsmaksimeringsestimatoren for $\tau^2$ som en funksjon av

$\theta$ og $\hat{\theta}(\tau^2)$ er sannsynlighetsmaksimeringsestimatoren of $\theta$ som en funksjon av $\tau^2$. Et konfidensintervall for $\tau^2$ er gitt ved verdiene som tilfredstiller $l_2^*(\tau^2) > l_2^*(\hat{\tau}^2) - 3.84/2$. Konfidensintervallet er ikke nødvendigvis symmetrisk.

### Å forklare heterogenitet

Flere forfattere understreker betydningen av å forklare heterogenitet og ikke bare ta hensyn til den ved å bruke random effect-metoden. Den anbefalte måten å gjøre dette på er å bruke meta-regresjon. Betegnelsen meta-regresjon brukes når de uavhengige variablene beskriver egenskaper ved undersøkelsene, i motsetning til de regresjonsanalyser som er mulige når data på individnivå er tilgjengelige fra hver undersøkelse (Thompson and Higgins, 2002).

Meta-regresjon har to viktige egenskaper. For det første, siden undersøkelsene som er enhetene i analysen sjelden vil være av samme størrelse, og variansene til estimatene derfor er forskjellige, er det heteroskedastisitet, og vektet regresjon er nødvendig. For det andre, det er lite sannsynlig at regresjon vil forklare all heterogenitet og den gjenværende heterogenitet må tas hensyn til i den statistiske analysen. Den korrekte regresjonsmodell er derfor en random effects modell hvor vekten for hver undersøkelse er lik den inverse av summen av undersøkelsenes varians og den gjenværende varians mellom undersøkelser, analogt med random effects modellen beskrevet ovenfor.

Den gjenværende variansen mellom undersøkelser er bare kjent etter at regresjonsanalysen er gjennomført. En metode for å estimere regresjonsligningen og $\tau^2$ samtidig eller med iterasjon er derfor nødvendig. Thompson and Sharp (1999) beskriver fire metoder. De er beskrevet i hovedteksten.

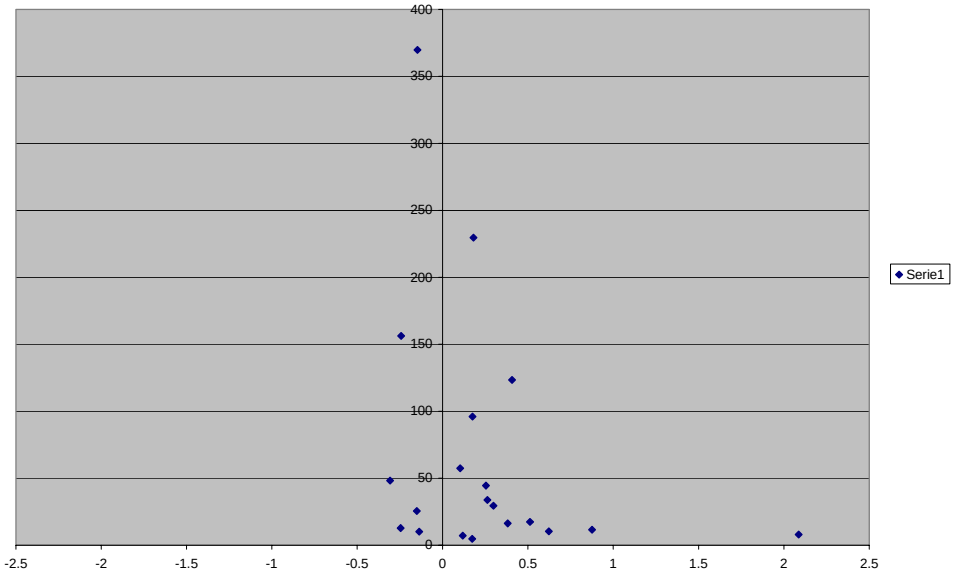
## Publikasjonskjevhet

Hvis undersøkelser som publiseres skiller seg fra undersøkelser som ikke er publisert med hensyn på hvilken virkning som er funnet, dvs at resultatene påvirker sannsynligheten for publisering, vil publiserte undersøkelser være et skjevt utvalg av alle undersøkelser. Dette er *publikasjonsskjevhet*.

Flere undersøkelser har vist at publikasjonsskjevhet er vanlig. Noen av disse er beskrevet i hovedteksten. Hovedtemaet i denne rapporten er imidlertid metoder for å undersøke om undersøkelsene som inngår i en meta-analyse er påvirket av publikasjonsskjevhet.

Et verktøy for å undersøke mulig publikasjonsskjevhet er traktdiagrammet. I et traktdiagram plottes virkningene som er funnet i et utvalg undersøkelser mot et mål for resultatenes presisjon.

Et eksempel på et traktdiagram er vist i figur 1.




Figur 1. Traktdiagram. Eksempel basert på vilkårlige data. TØI rapport 692/2003.

I figuren er virkningen uttrykt ved logaritmen til oddsforholdet og presisjonen er undersøkelsenes vekt, dvs den inverse til variansen til logaritmen til oddsforholdet.

Årsaken til betegnelsen traktdiagram er at når det ikke er publikasjonsskjevhet vil diagrammet se ut som en trakt snudd på hodet. Store undersøkelser vil være mer nøyaktige og vise mindre variasjon enn små undersøkelser. Når det ikke er publikasjonsskjevhet vil diagrammet være symmetrisk om gjennomsnittet.

Et forhold som skaper publikasjonsskjevhet er hvis resultater som ikke er signifikante ikke blir publisert. Siden små undersøkelser krever en større observert virkning for å være statistisk signifikant vil det være en tendens til at man finner større virkninger for små undersøkelser fordi små undersøkelser som finner liten eller endog en negativ virkning vil mangle. Det vil da mangle undersøkelser nede til venstre i traktdiagrammet. Dette ser ut til å være tilfellet i traktdiagrammet ovenfor og kan tolkes som et tegn på publikasjonsskjevhet.

Traktdiagrammet utnytter forskjellen i virkning mellom store og små undersøkelser. For at et traktdiagram skal være egnet er det derfor nødvendig at undersøkelsene varierer i størrelse.

For testing av publikasjonsskjevhet finnes statistiske metoder som er analoge til traktdiagrammet. Tre metoder beskrives her. To av metodene bygger på antagelsen om at publikasjonsskjevhet leder til en sammenheng mellom størrelsen på virkningen som er funnet og standardavviket til virkningen. En av metodene tester for en slik sammenheng ved bruk av en rangkorrelasjon test og den andre bruker regresjonsanalyse. Den tredje metoden tester for symmetri i traktdiagrammet.

Beggs test (Begg, 1994) er en test for uavhengigheten mellom størrelsen og variansen av virkningen og bygger på Kendalls tau. Testen er basert på en antagelse om at størrelsen av virkningene er statistisk uavhengige og identisk fordelte under nullhypotesen om at det ikke er publikasjonsskjevhet. Det er derfor nødvendig å standardisere størrelsen av virkningene for å kunne utføre testen. Betegnes størrelsen av virkningen og dens varians

med henholdsvis $x_i$ og $v_i$, brukes rangkorrelasjon til å teste sammenhengen mellom $x_i = \frac{(x_i - \bar{x})}{\bar{v}_i^{\frac{1}{2}}}$,

hvor $\bar{x}$ er fixed effects gjennomsnitt av virkningene og hvor $\bar{v} = v_i - \left( \sum_{j=1}^{n} v_i^{-1} \right)^{-1}$ er variansen til $x_i - \bar{x}$.

Testen går ut på å beregne P, antallet av alle mulige par hvor de tofaktorene er rangert i samme rekkefølge , og Q, antall par hvor de ikke er i rekkefølge. En normalisert test observator (z score) er da gitt ved:

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

Egger m fl (1997) bruker lineær regresjon til undersøke sammenhengen mellom virkningens størrelse og varians og på den måten teste for publikasjonsskjevhet. I regresjonen er den avhengige variabelen den standardiserte virkningen og den uavhengige variabelen er den inverse av standardavviket til virkningen. Betegnes virkningen med x og standardavviket med s er regresjonsligningen:

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

Testen for publikasjonsskjevhet er basert på koeffisienten *a*. En signifikant verdi tyder på publikasjonsskjevhet.

Begrunnelsen for testen er følgende. Den inverse av standardavviket er et mål for presisjon. Unøyaktige undersøkelser (vanligvis små undersøkelser) vil ligge nær origo på abscissen. Den standardiserte virkning vil da også være liten. Unøyaktige undersøkelser vil derfor ha små verdier på begge akser, dvs de vil ligge nær origo. Nøyaktige undersøkelser vil ligge langt fra origo på abscissen og hvis der er en virkning vil ordinatverdien også være stor. Regresjonslinjen gjennom de plottede undersøkelsene vil derfor gå tilnærmet gjennom origo med en vinkelkoeffisient lik den veide gjennomsnittsvirkningen. Dette er situasjonen når det ikke er publikasjonsskjevhet.

Når traktdiagrammet er asymmetrisk fordi det er publikasjonsskjevhet og mindre undersøkelser finner virkninger om avviker systematisk fra store undersøkelser vil ikke regresjonsligningen gå gjennom origo. Konstantleddet, koeffisienten *a*, vil derfor være et mål for asymmetrien. Fortegnet til *a* vil avhenge av hvordan virkningen måles. Når virkningen måles ved logaritmen til oddsforholdet og en negative verdi betyr en positiv virkning vil koeffisienten være negativ når det er publikasjonsskjevhet. En test for publikasjonsskjevhet fås følgelig ved å teste om koeffisienten *a* er forskjellig fra null. Da testen har lav styrke anbefaler Egger m fl å bruke er 10 % signifikansnivå.

"Trim and fill"-metoden til Duval and Tweedie (2000a, 2000b) bygger også på traktdiagrammet men i dette tilfellet er ikke utgangspunktet sammenhangen mellom virkningens størrelse og dens varians men symmetrien (eller mangel på symmetri) i traktdiagrammet.

Hvis det ikke er publikasjonsskjevhet, vil traktdiagrammet være symmetrisk. "Trim and fill"-metoden fjerner derfor et antall undersøkelser på en side av diagrammet slik at det blir symmetrisk (trimdelen), beregner et veid gjennomsnitt av de resterende undersøkelsene og genererer så et tilsvarende antall undersøkelser på den andre siden. De genererte undersøkelsene er symmetrisk til de fjernede undersøkelsene rundt det beregnete gjennomsnittet. Et eksempel med diagrammer er vist i hovedteksten.

## Kvalitetsvurdering og kvalitetscore

Bangert-Drowns, Wells-Parker and Chevillard (1997) påpeker at hvis kvaliteten på undersøkelsene ikke tas hensyn til, så betyr dette at undersøkelser med store utvalg er overlegne andre undersøkelser på grunnlag av bare denne egenskapen; utvalgsstørrelse.

Den eneste usikkerhet ved undersøkelsene som det tas hensyn til er den statistiske, som om det ikke er noen metodiske problemer. Dette er selvfølgelig meget urealistisk og er alene et argument for å vurdere kvaliteten på undersøkelsene.

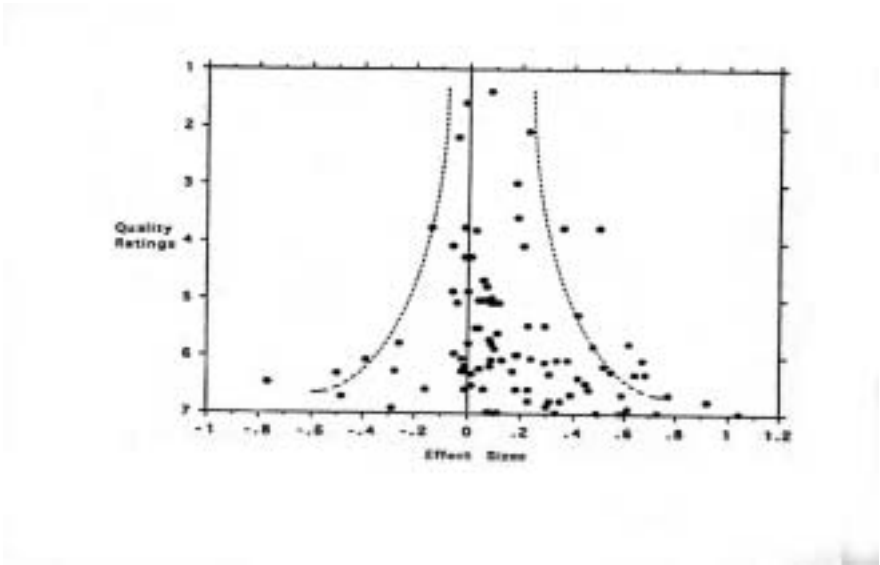
Bruk av kvalitetsvurderinger i meta-analyser krever svar på to spørsmål, hvordan kvalitet skal måles og hvordan kvaliteten på undersøkelsene skal tas hensyn til i meta-analysen. Svaret på siste spørsmål avhenger av de mulige virkninger av dårlig kvalitet av undersøkelser. Virkningen kan være:

1. En systematisk skjevhet
2. Økt varians (en større usikkerhet eller mindre presisjon)

Flere undersøkelser har funnet at undersøkelser av lav kvalitet har en tendens til å overvurdere virkningen. Imidlertid har andre undersøkelser ikke funnet at undersøkelser med lav kvalitet leder til systematiske skjevheter. Balk et al (2002) gjennomførte en empirisk undersøkelse av korrelasjonen mellom ulike kvalitetsmål og estimatet av virkningen av behandling. 24 kvalitetsmål ble analysert for 276 randomiserte kontrollerte forsøk fra 26 meta-analyser. Kvalitetsmålene ble omkodet til dikotome variable, høy kvalitet eller lav kvalitet. Betydningen av kvalitetsmålet ble estimert ved å beregne det relative oddsforholdet for hvert mål. Relative oddsforhold for undersøkelser med høy kvalitet mot undersøkelser med lav kvalitet varierte for de ulike kvalitetsmålene mellom 0.83 og 1.26. Ingen hadde signifikant sammenheng med størrelsen på virkningen.

Hvis resultatene fra undersøkelser av lav kvalitet varierer mer enn resultatene fra gode undersøkelser, fordi resultatene fra dårlige undersøkelser er mindre pålitelige, vil kvalitet ha samme virkning på variasjonen i resultatene som utvalgsstørrelse. Et plot av størrelsen på virkningen mot kvalitet vil oppføre seg på samme måte som et traktdiagram med størrelsen på virkningen mot statistisk vekt.

Dette er vist i figur 2 hentet fra Bangert-Drowns, Wells-Parker og Chevillard (1997). Lave tall betyr høy kvalitet.



Figur 2. Punktdiagram for sammenhengen mellom virkningen på residivisme og vurderinger av metodologiisk kvalitet av undersøkelser av rehabiliteringsprogrammer for promilleførere.
Fra Bangert-Drowns, Wells-Parker og Chevillard (1997).

### Hvordan ta hensyn til kvalitet

Den større usikkerheten til dårlige undersøkelser underbygger behovet for å ta hensyn til undersøkelsenes kvalitet på en eller annen måte. Det finnes (minst) fem måte å gjøre dette:

1. Utelate de dårligste undersøkelsene, dvs definere en terskel for kvalitet og bare ta med undersøkelser over denne terskelen i meta-analysen.
2. Stratifisere undersøkelsene med hensyn på kvalitet og gjøre en separat meta-analyse for hvert stratum.
3. Bruke kvalitetscore som vekter på same måte som man nå bruker statistiske vekter.
4. Bruke meta-regresjon til å uttrykke virkningen som en funksjon av enten kvalitet eller komponentene som inngår i kvalitetsmålet
5. Sekvensiell sammenveiing resultatene fra undersøkelsene basert på kvalitetscore (Detsky et al, 1992).

Ingen av de to første metodene er tilfredstillende. Enhver terskel vil nødvendigvis være vilkårlig. I tillegg vil ikke kvaliteten lenger være av betydning for de undersøkelser som ligger over terskelen. En undersøkelse vil enten ha ingen vekt eller full vekt avhengig av terskelen.

Å stratifisere etter kvalitet løser heller ikke problemet, det er bare utsettelse. Det gir ikke noe svar på spørsmålet om hva som skal gjøres med resultatene fra meta-analysene for hver stratum. Hvis man bare stoler på resultatet fra stratumet med høyest kvalitet er dette ekvivalent med bruk av en terskel. Hvis resultatene fra alle strata skal benyttes gjenstår fremdeles spørsmålet om hvordan resultatene fra de enkelte strata skal veies sammen.

Meta-regresjon bygger på den antagelse at det er en systematisk sammenheng mellom en kvalitetsskala eller komponentene til en kvalitetsskala og resultatene i en undersøkelse. Hvis det ikke er noen systematisk sammenheng vil meta-regresjonen tyde på at de metodologiske variable ikke påvirker forskjellen i resultater mellom undersøkelser. Likevel vil variasjoner som skyldes metodologiske mangler bidra til at størrelsen på feilleddet øker og følgelig blir konfidensintervallene videre. Dårlige undersøkelser vil imidlertid påvirke resultatet like meget som gode undersøkelser.

Sekvensiell sammenveiing av resultatene fra undersøkelsene basert på kvalitetscore kan betraktes som en spesiell form for bruk av en terskel for kvalitet og lider derfor av samme svakhet som bruk av en terskel.

Kvalitetsscoring har få tilhengere. Jüni, Altman and Egger (2001a) mener at "bruk av kvalitetsscore som vekter mangler statistisk og empirisk begrunnelse" og at det er ingen grunn til at kvalitetsscore skal modifisere estimatets presisjon. Problemene med kvalitestscore kan illustreres ytterligere av en undersøkelse av Jüni et al (1999).

De evaluerte bruken av 25 forskjellige vurderingsskalaer beskrevet av Moher m fl (1995). Disse skalaene ble brukt på 17 forsøk som sammenlignet bruk av ulike hepariner for å forebygge blodpropp etter operasjoner.

Selv om overenstemmelsen for standardiserte scorer for de 25 skalaene var betydelig (intraclass korrelasjonkoeffisient 0.72 (0.59,0.86)) varierte medianen av av kvaliteten til undersøkelsene fra 38.5% til 82.9% av maksimum score for de ulike skalaene. Ved bruk av kvalitetscore som vekter i en meta-analysis, ville konfidensintervaller basert på ska-

laen med den laveste median score være mer enn to ganger så vide som konfidensintervall basert på skalaen med den høyeste median score.

Resultatet ser ut til å være et argument mot å bruke kvalitetsscore i meta-analyser. Vårt syn er imidlertid at resultatet bare gjenspeiler vilkårligheten i eksisterende kvalitetsscore og kan unngås ved å utvikle bedre kvalitetsskalaer.

### Måling av kvalitet

Ulike kvalitetsvurderinger vil avvike dersom de underliggende kvalitetsbegrepene er forskjellige. Arbeidet med å vurdere kvalitet må derfor begynne med å avgrense hva som skal ligge i begrepet "kvaliteten av en undersøkelse".

I hvilken sammenheng begrepet skal brukes er viktig for definisjonen av kvalitet. En vurdering av kvaliteten av en undersøkelse innsendt til et tidskrift for publisering er noe annet en vurderingen av kvaliteten til an undersøkelse når den inngår i en meta-analyse. I det første tilfellet er kvalitetsbegrepet meget videre enn i det siste, det omfatter f eks interesse for leserne, resultatenes orginalitet osv. En undersøkelse som inngår i en meta-analyse kan betraktes som et instrument for å måle virkningen av noe. Kvaliteten til undersøkelsen er et mål for i hvilken grad man kan stole på resultatene, dvs undersøkelsens reliabilitet og validitet.

Dette leder til følgende definisjon av kvalitet: Den grad en undersøkelse har unngått metodologiske svakheter som kan påvirke resultatene. Dette er nesten (men ikke helt) det samme som begrepet "intern validitet" hos Shadish, Cook and Campbell (2002). Noen forfattere (Downs and Black 1998, Verhagen et al 1998) er av den oppfatning at kvalitetsbegrepet også bør omfatte ekstern validitet. Vårt syn er at ekstern validitet må tas hensyn til gjennom meta-analysen ved å sikre at de inkluderte undersøkelsene varierer med hensyn på enheter som inngår i undersøkelsene og forholdene de blir utført under. Meta-regresjon kan så benyttes til å analysere virkningen av variasjonen. Ekstern validitet bør ikke inngå i definisjonen av kvalitet.

Jadad et al (1996) nevner tre metoder for å vurdere kvaliteten til kliniske forsøk: individuelle markører, også kalt elementer eller komponenter, sjekklister og skalaer.

Individuelle markører er de mulige dimensjonene til kvalitet. Eksempler for randomiserte kontrollerte forsøk er randomiseringsmetoden og blindingsmetoden. For kvasi-eksperimentelle undersøkelser er det vanskeligere å gi eksempler på individuelle markører.

Sjekklister gir et kvalitativt anslag for kvaliteten til en undersøkelse med utgangspunkt i individuelle markører eller komponenter for å sammenligne undersøkelser (Moher, Jadad and Tugwell, 1996). For sjekklister beregnes ikke noe numerisk score.

En skala konstrueres ved å gi komponentene numeriske verdier og summere (hvis ønskelig veid) disse. For å vekte undersøkelser er kvalitetsskalaer å foretrekke.

Med den gitte definisjonen av kvalitet må kvalitetsskalaen være et mål for hvor godt faktorer som kan påvirke resultatet har blitt kontrollert for. En måte å kode undersøkelser er derfor å rangere ulike design for deres evne å kontrollere for bakgrunnsvariable og betrakte kvaliteten på undersøkelsen høyere jo lavere undersøkelsdesignet er rangert. Hvor gjennomført designet er implementert bør også tas hensyn til. Et alternativ er å lage en oversikt over mulige bakgrunnsvariable og å sjekke hvor godt undersøkelsen har kontrollert for dem. Jo flere bakgrunnsvariable som er kontrollert for jo høyere er kvaliteten på undersøkelsen.

Å evaluere validiteten til en kvalitetsskala er vanskelig. For å verifisere den teoretiske validiteten til en kvalitetsskala er det nødvendig å avlede konsekvenser av kvalitet som kan bekreftes empirisk. En slik konsekvens diskutert ovenfor er at virkningene funnet i undersøkelser av god kvalitet vil variere mindre enn virkningene funnet i undersøkelser

av dårlig kvalitet. Dette kan testes i et traktdiagram med virkningen langs abscissen og kvalitetsscoren langs ordinaten.

Det er to vesentlige forutsetninger for bruken av kvalitetsvurderinger i meta-analyser:

1. En undersøkelses kvalitet bør ha betydning for dens vekt i meta-analysen
2. Usikkerheten som skyldes metodologiske svakheter bør gjenspeiles i det veide estimatet for virkningen.

For å oppnå dette er en kvalitetskala nødvendig.

# Contents

# 1 Introduction

Until 1999 the use of meta-analysis for systematic reviews of earlier research was a solitary effort at the Institute of Transport Economics. This work led to a doctorate and several papers that reviewed the effects of various road safety measures (Elvik, 1999). The methods of meta-analysis used, however, were fairly simple and it was believed that the advancement of meta-analysis at the institute necessitated special funding. A so-called Strategic Institute Program was therefore initiated to develop the subject of meta-analysis at the institute.

The objective of the Strategic Institute Program is three-fold:

a) To enhance knowledge of meta-analytic methods at the Institute of Transport Economics and increase the skills of researchers in performing meta-analyses.

b) To improve on the simple methods that have been employed by studying more advanced methods and the pitfalls of meta-analysis.

c) To initiate the use of meta-analysis in other fields of transport research than road safety.

Of course, the first two objectives are not independent. If objective b is attained it will necessarily mean that objective a will be attained as well.

The work so far has mainly been pertinent to point b. Extensive literature surveys have been undertaken, both of various methods of meta-analysis and of the more contentious aspects of the principles of meta-analysis. This report describes the results of this work.

It starts with a short survey of meta-analytic methods. Then some problems like publication bias and quality scores are discussed briefly. With this basis the reasons are given for choosing some areas to go deeper into. The main chapters discuss the main areas chosen, ie heterogeneity, publication bias and quality scores. The final chapter discusses further work.

## 1.1 References

Rune Elvik. *Assessing the validity of evaluation research by means of meta-analysis. Case illustrations from road safety research*. TØI Report 430/1999. Oslo: Institute of Transport Economics.

# 2 An introduction to meta-analysis

## 2.1 What is meta-analysis?

The word meta-analysis was coined by Glass (1976). In the literal sense it means an analysis of analyses but in the sense introduced by Glass it indicates a study that summarises previous studies. However, summaries of earlier studies are also known as systematic reviews. Is there a difference between meta-analyses and systematic reviews?

Egger, Smith and Sterne (2001) define meta-analysis as the statistical combination of results from several studies to produce a single estimate of the effect of a treatment and systematic reviews as reviews that have been prepared using a systematic approach to minimise bias and explicitly address the issues of the completeness of the evidence identified, the quality of component studies and the combinability of studies.

Normand (1999) defines meta-analysis broadly as the quantitative review and synthesis of the results of related but independent studies while Petitti (2001) states that "in its most limited sense, meta-analysis is a statistical technique for combining quantitative data". She continues, however, with: "meta-analysis is, however, usually done in the context of a systematic review of the literature, although not all systematic reviews include a quantitative synthesis….. In this paper, the term 'meta-analysis' refers to quantitative synthesis after a systematic review of the literature".

In this report we will follow Petitti and define meta-analysis as the calculation of an overall mean of effect of a measure within the framework of a systematic review. In this view, there is no basis for a meta-analysis without a systematic review and since the purpose of a systematic review will normally be a meta-analysis, in practice the two terms can be used interchangeably.

A systematic review necessitates many tasks before the meta-analysis can be carried out, searching for and retrieving relevant studies, coding the results etc. These tasks will not be discussed in this report, which concentrates on the statistical methods of meta-analysis.

Some meta-analyses have employed individual data, that is data on units in the studies comprising the meta-analysis, but in most meta-analyses individual data has not been available. The basis for a meta-analysis is normally summary results like an effect size and its variance. The statistical methods described here are limited to the analysis of summary data from studies.

Bayesian methods are not discussed. This is partly due to lack of time for studying Bayesian methods, even for a Strategic Institute Program funds are limited, but also due to prejudice. The author is a non-Bayesian.

## 2.2 The basic meta-analytic method

Most studies evaluate something. Clinical studies evaluate drugs or treatments. Road safety studies evaluate measures aiming at reducing road accidents. The result of a study is normally an estimate of the effect of the treatment or measure. However, different studies of the same phenomenon often find different effects. The aim of a meta-analysis is to find the best estimate of the effect of a measure on the basis of all studies. This is done by taking the weighted mean of the effects found in the studies.

Let $\theta$ be the true effect of a treatment or measure (measurements of effect will be briefly discussed below). 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 $\sigma_i^2$ of the estimate had been known an optimal estimate for the true effect $\theta$ 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 $\sigma_i^2$ 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 disregarded and the weights treated as if the true $\sigma_i^2$ is known. When the studies included are fairly large this is not a serious error.

Various effect measures $\theta$ are possible. A general discussion of the possibilities will not be undertaken here. A very common effect measure in clinical trials is the odds ratio (OR). In road-safety research a quasi odds-ratio is often employed, ie an expression of the form $o = \frac{a_{12}a_{22}}{a_{11}a_{21}}$ where the *a*s represent a number of accidents.

It is assumed that the estimates $\hat{\theta}_i$ are (at least approximately) normally distributed. To fulfil this requirement the logarithm of the odds ratio or the quasi odds-ratio is used in the formula above. The approximation formula for the variance of the logarithm is the same for both:

$$Var[\ln[o]] = \frac{1}{a_{11}} + \frac{1}{a_{12}} + \frac{1}{a_{21}} + \frac{1}{a_{22}}$$

The weighting formula above assumes that the true effect is the same, or fixed, in every study. The differences in the results between studies are due to the uncertainties of the estimates in the studies. This is called the *fixed effects* model.

## 2.3 Pitfalls and problems

The description above left out some problems. Firstly, how can one retrieve all studies bearing on a certain problem? In practice this is well-nigh impossible. A random sample will of course do, but how can a random sample be obtained? In practice, published studies will often be the basis of a meta-analysis but published studies are very likely a biased sample of studies. This *publication bias* may be a serious problem for meta-analyses and a chapter discusses this problem. The main objective of our literature survey of publication bias has been to identify statistical methods to diagnose publication bias and if possible indicate its influence on the results. Chapter 4 on publication bias describes such methods.

Secondly, studies may be of varying quality. This is a particular problem in road safety research where most studies are observational and employing various methodological designs. Should quality be disregarded and all studies treated the same or should methodologically weak studies be left out? How to handle differences in quality will be discussed in chapter 5. The main questions treated in this chapter are the following: How

should quality be measured and how should quality be taken into account in meta-analyses? Before that, however, the case for somehow taking quality into account is argued.

The third problem is that the assumption that studies estimate the same true effect may not be tenable. The effects found in different studies may show too much variation. Chapter 3 is devoted to a discussion of heterogeneity of results, how to describe and measure heterogeneity and how to allow for it in the statistical analysis.

## 2.4 References

Petitti DB. Approaches to heterogeneity in meta-analyis. *Statist. Med*. 2001;20:3625-3633(2001).

Matthias Egger, George Dayey Smith and Jonathan AC Sterne. Uses and abuses of meta-analysis. *Clinical Medicine,* Vol 1 No 6 November/December 2001.

Sharon-Lise T. Normand. Tutorial in biostatistics -analysis formulating, evaluating, combining, and reporting. *Statist. Med*. 18, 321-359 (1999).

# 3 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 or the standard deviation of the estimated effect of the studies. 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. Related definitions of statistical heterogeneity are found in Thompson and Higgins (2002) where it is defined as the true effect of each study not being identical and in Thompson (2001) as "incompatibility in quantitative results".

Moses, Mosteller and Buehler (2002) explain heterogeneity in the following way: "If each 'small' trial were in fact very large we would expect each to arrive at somewhat different estimates of the clinical effect, since the various trials presumably differ in many respects".

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. Such methods are discussed in section 3.1. This is followed by a discussion of causes of heterogeneity. The main part of this chapter describes methods for analysing heterogeneous data.

## 3.1 Diagnosing and measuring heterogeneity

Both tests and graphical methods for investigating heterogeneity have been developed. The graphical methods can be used to explore whether there is heterogeneity and can also give an indication of which studies contribute to heterogeneity. The tests use quantitative indicators to determine whether there is heterogeneity. A problem for the application of these methods is that the number of studies in a meta-analysis is often small and the result often inconclusive.

### 3.1.1 Exploring heterogeneity

Heterogeneity can be explored by plotting the results of the individual studies to get a visual impression of their variation. Three different plots are commonly used, the forest plot, the Galbraith plot (Galbraith, 1988) and the L'Abbé plot (L'Abbe, Detsky, and O'Rourke, 1987). The plots give an indication of heterogeneity and can also show the studies with deviant results. Baujat et al (2002) have introduced a graphical method to identify trials or groups of trials that are sources of heterogeneity. The contribution of these trials to the overall result can also be evaluated with this method.

**The forest plot**

The forest plot is a plot of the treatment effect by trial. The estimated effect in the individual studies with their confidence intervals are plotted. Precise estimates are represented by short lines, and uncertain results by long lines. An example of a forest plot is given in figure 3.1. It is taken from Elvik (2002).

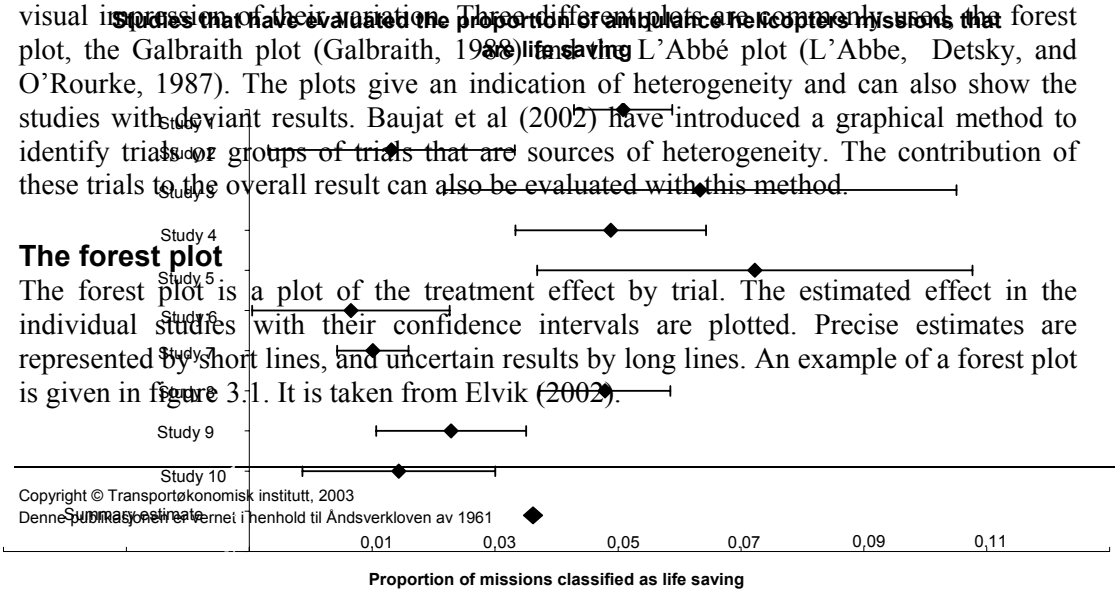


Figure 3.1. Forrest plot from Elvik (2002).

The figure indicates that there is considerable heterogeneity. The confidence interval of the summary value is not drawn, but since the standard deviation of the summary value must be smaller than the standard deviation of the most precise of the individual studies when the fixed effects method is applied, the confidence interval must be smaller than the one for study 7. Study 7 is therefore significantly different from the summary value and also significantly different from study 8.

### The Galbraith plot

In the Galbraith plot the z-statistic for each trial $\theta_i / \hat{\sigma}_i$ is plotted against the reciprocal standard error $1/\hat{\sigma}_i$. The regression line for $\theta_i / \hat{\sigma}_i$ on $1/\hat{\sigma}_i$ constrained through the origin can also be plotted. The slope of this line is the fixed effects estimate of the overall effect. Lines parallel with the regression line indicate 95% confidence intervals. An example is shown in figure 3.2 taken from Thompson (1993).

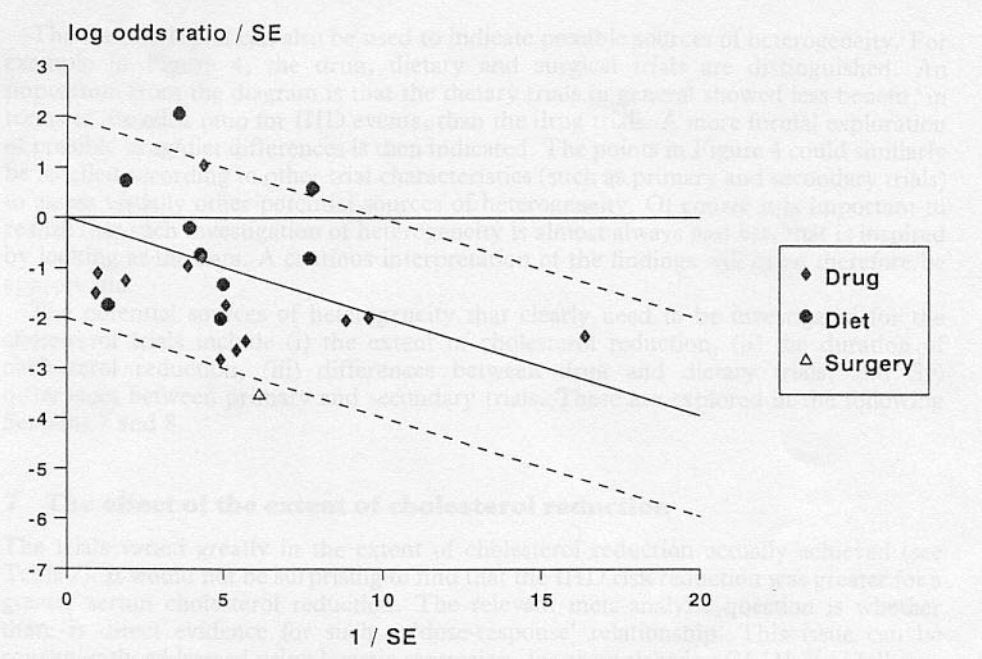



Figure 3.2. An example of a Galbraith plot taken from Thompson (1993).

### The graphical method of Baujat et al

The graphical method of Baujat et al (2002) is a plot of the contribution of a study to the overall Cochran Q-test for heterogeneity (see section 3.1.2) along the x-axis and the influence of the study, defined as the standardized squared difference between the treatment effects estimated with and without the study, along the y-axis. The contribution to the Cochran Q-test by study i is given by: $\hat{x}_i = \frac{(\hat{\theta}_i - \hat{\theta})^2}{\hat{\sigma}_i^2}$. $\hat{\theta}$ is the overall effect estimate calculated by the fixed effects model, The influence of the study is defined as the square of the difference between the treatment effect estimated with ($\hat{\theta}$) and without trial i ($\hat{\theta}_{-i}$), weighted by the inverse of the estimated variance of the treatment effect after

exclusion of the *i*th trial: $\hat{y}_i = \frac{(\hat{\theta}_{-i} - \hat{\theta})^2}{\hat{\sigma}^2_{-i}}$ where $\hat{\theta}_{-i} = \frac{\sum_{j \neq i} \frac{\hat{\theta}_j}{\hat{\sigma}_j^2}}{\sum_{j \neq i} \frac{1}{\hat{\sigma}_j^2}}$ and

$$\hat{\sigma}_{-i} = \frac{1}{\sum_{j \neq i} \frac{1}{\hat{\sigma}_j^2}}.$$

### The L´Abbé plot

The L´Abbé plot can be used for clinical studies with an experiment group and control group when the measure of effect is a relative risk. It is useful to identify not only the studies having different results from other studies, but also the study arms that are responsible for such differences. In a L´Abbé plot the percentage in the experiment group is plotted against the percentage in the control group. The overall estimated relative risk line can also be plotted and the plotted studies distance from the relative risk line used as an indication for outliers, ie studies that show a large deviation from the overall relative risk. An example is shown in figure 3.3 taken from Song (1999).

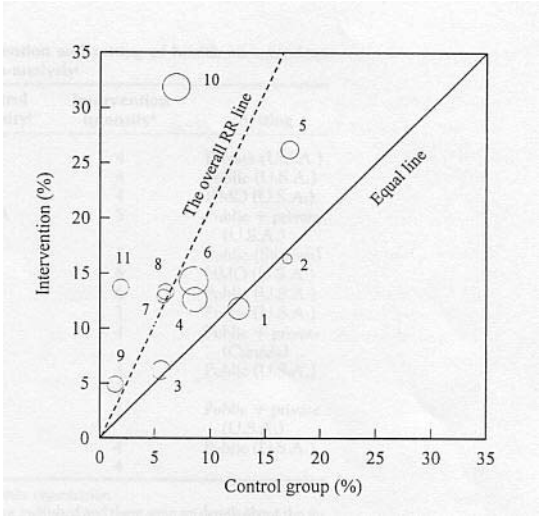



Figure 3.3. An example of a L'Abbé plot taken from Song (1999).

Song (1999) has shown by simulations that purely because of random variation, studies with event rates of around 50% are more likely to be identified as outliers in a L'Abbé plot. He gives a figure showing the simulated standard deviations and suggests that these standard deviations can be used to standardize the absolute distances between individual studies and the overall RR line. Song found that for eleven meta-analyses, the outlying trials identified by the Forrest plot were different from that identified by the adjusted and unadjusted L'Abbe method in 6 and 7 meta-analyses respectively.

Song sees it as a practical problem that different methods may identify different trials as outliers, and the conclusions of a meta-analysis may change by excluding different studies. Excluding studies is probably not the thing to do (this will be discussed later), and using plots for exploring heterogeneity will always be subjective. Analytical methods are preferable. Testing for heterogeneity will be discussed in the next section.

### 3.1.2 Testing for heterogeneity

The most common test for heterogeneity is the Q-test often ascribed to DerSimonian and Laird but previously proposed and discussed by Cochran (Takkouche, Cadarso-Suarez, and Spiegelman, 1999). For reasons made clear later this is the only test described here.

The Cochran Q-test 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. If the weights $w_i$ of the studies were based on the true standard deviations σ then under $H_0$ (homogeneity) Q would be chi square distributed with (k-1) degrees of freedom where k is the number of studies (Brockwell and Gordon, 2001). Since the weights are based on estimates for the standard deviations, Q is only approximately chi square distributed. The approximation is acceptable provided that each individual study has a sample size that is large compared with the number of studies and provided that var($\hat{\theta}_i$) is independent of $\hat{\theta}_i$ (Takkouche, Cadarso-Suarez, and Spiegelman, 1999).

According to Higgins and Thompson (2002) it is well known that the test has poor power in the common situation of few studies, and excessive power to detect clinically unimportant heterogeneity when there are many studies. Since few meta-analyses comprise many studies, excessive power is not much of a problem in practical work.

Lack of power, however, can certainly be a problem. Simulations to investigate the power of the test have therefore been carried out.

There are three characteristics of any meta-analysis which will have an impact on the expectation of Q and therefore on the power of the test for heterogeneity. They are: (a) the extent of heterogeneity present, that is, the value of $\tau^2$, the variance between studies (b) the number of studies k included in the meta-analysis; (c) the weight $w_i$ allocated to each study (Hardy and Thompson, 1998). Hardy and Thompson therefore varied these parameters in their simulation. They defined the total amount of information as $\sum w_i$ and studied how the variation between study weights $w_i$, given the total amount of information, influenced the power of the test. According to Hardy and Thompson the maximum value of E(Q) for a given total information, number of trials and between-study variance is achieved when the weights are the same.

When all studies have the same weight $w_i$ (and the variance $v_i$) and when the number of studies were 10 the power was 0.3 for $\tau^2/v_i$=0.5, 0.5 for $\tau^2/v_i$=1.0 and 0.8 for $\tau^2/v_i$=2.0. When weights are unequal, the total information required for a given k and $\tau^2$ in order to maintain the same expectation E(Q) as that in the equal weighting case increases. When one trial takes 90 per cent of the weight, the necessary total information increases by a factor of 4.76. Power is considerably reduced when the weight allocation is very uneven.

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, inter alia a likelihood ratio test, by simulation. Both asymptotic versions and bootstrap versions that they developed were compared.

In the studies of both type I error and of power, the results were virtually identical for every odds ratio considered. Only the case in which the odds ratio is equal to 2 was therefore presented and discussed.

Only the Q test gave the correct size under the null hypothesis. Although the Q statistic is strictly appropriate only when cov($\hat{\theta}_i$,var($\hat{\theta}_i$)) = 0, an assumption which is violated by binomial data as in these epidemiologic meta-analyses, violation of this assumption did

not have a detectable adverse impact on the validity of the test in the settings considered. Takkouche, Cadarso-Suarez, and Spiegelman 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 ($R_I$† ≥ 0.75). For small values of $R_I$, no test has a satisfactory power of, for example, 80 percent. It may thus be deceptive to use any homogeneity test when the proportion of between-study variance is lower than 0.4, as long as $k \leq 40$.

The components of Q can be used to investigate the possible sources of heterogeneity. To investigate whether study $i$ contributes to heterogeneity $q_i^2 = w_i(\hat{\theta}_i - \hat{\theta})^2$ can be compared to $\chi_1^2$ -distribution provided the number of trials $k$ is not too small (approximate because under the homogeneity assumption Q has a $\chi_{k-1}^2$ rather than a $\chi_k^2$ distribution) (Thompson, 1993). This investigation of heterogeneity is equivalent to using the Galbraith plot, $q_i$ is the vertical distance between each trial's point and the regression line.

As mentioned above, in the unlikely case that the number of studies is large, the test for heterogeneity may be significant even when heterogeneity is not very pronounced. It is therefore fruitful not just to test for heterogeneity but to quantify it. Measures of heterogeneity is discussed in the next section. $Q_i^2$ will also be recognized as the contribution to Cochran's Q employed in the graphical method of Baujat et al.

### 3.1.3 Measures of heterogeneity

Higgins and Thompson (2002) introduce three measures for quantifying heterogeneity, denoted by H, R and I respectively. H is the square root of the heterogeneity statistic Q divided by its degrees of freedom; R is the ratio of the standard error of the underlying mean from a random effects meta-analysis to the standard error of a fixed effects meta-analytic estimate, and $I^2$ is a transformation of H that describes the proportion of total variation study estimates that is due to heterogeneity. More precisely the formulas are:

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

$$R^2 = \frac{\sum w_i}{\sum \left( w_i^{-1} + \hat{\tau}^2 \right)^{-1}}$$

$$I^2 = \frac{H^2 - 1}{H^2}$$

Higgins and Thompson discuss the properties of the measures and outline ways of calculating the uncertainty of H. They propose H and $I^2$ as they favoured measures for quantifying heterogeneity in a meta-analysis.

Takkouche, Cadarso-Suarez, and Spiegelman (1999) suggest that heterogeneity can be quantified by means of $R_I$, the proportion of the total variance of the pooled effect measure due to between-study variance, and $CV_B$, the between-study coefficient of variation. Their measure $R_I$ is similar to the $I^2$ measure of Higgins and Thompson. Both measures are of the form: $\frac{\tau^2}{\tau^2 + \hat{\sigma}^2}$ where $\hat{\sigma}^2$ is a measure of the typical within-study

† R is defined in the next section.

variance, but different estimates for $\hat{\sigma}^2$ have been employed. Takkouche, Cadarso-Suarez, and Spiegelman use $\hat{\sigma}^2 = k\frac{1}{\sum w_i}$ while Higgins and Thompson use $\hat{\sigma}^2 = \frac{(k-1)\sum w_i}{\left(\sum w_i\right)^2 - \sum {w_i}^2}$. This choice gives the simple relation between $H^2$ and $I^2$.

The coefficient of variation suggested by Takkouche, Cadarso-Suarez, and Spiegelman is given by $CV_B = \frac{\sqrt{\tau^2}}{|\hat{\theta}|}$.

## 3.2 Causes of heterogeneity

### 3.2.1 Factual and methodological heterogeneity

Causes of heterogeneity are of two kinds, factual and methodological. The first is called clinical heterogeneity in clinical trials and might result from differences between patients, interventions compared or outcomes collected (Higgins et al, 2002). Factual heterogeneity is when there are real differences in effect between studies.

Methodological heterogeneity arises through the use of different study designs and different degrees of control over bias and does not represent real differences. This is a particular problem for observational studies where methodological problems abound and there are large variations in the quality of studies. Low study quality therefore contributes to heterogeneity.

While methodological heterogeneity should always be avoided, factual heterogeneity is not necessarily an evil. Factual heterogeneity may enhance the external validity, ie it may make the result apply to a wider population. Matt and Cook (1994) hold that: "The wider the range and the larger the number of substantively irrelevant aspects across which a finding is robust, and the better moderating influences are understood, the stronger is the belief that the finding will also hold under the influence of not yet examined contextual irrelevancies". Not everyone agrees. Feinstein (1995) believes that claims that a heterogeneous population gives the results "wider applicability" supports the idea that meta-analyses may be metaphysical or alchemistic.

Another possible advantage of heterogeneity is that it may permit analyses of the association between various factors and the effect. This is discussed in chapter 3.4.

### 3.2.2 Heterogeneity due to the choice of effect measure

One reason for heterogeneity might be an inappropriate choice for the measure of treatment effect (Whitehead and Whitehead, 1991) or heterogeneity may be an artefact of the summary measures used (Glasziou and Sanders, 2000). For clinical trials in particular, the choice of odds ratio, risk ratio or risk difference as effect measure may have a bearing on the heterogeneity found. This is basically due to the weights given to the individual studies. For example, compared with a trial where the probability of an outcome in the control group ($P_C$) and in the trial group ($P_T$) is 0.5, an equally sized trial in which $P_C$ = 0.05 and $P_T$= 0.01 would receive 8.71 times as much weight in an estimate of the fixed effects summary risk difference, but only one-fifth the weight when using the Mantel-Haenszel method to estimate the odds ratio (Engels et al, 2000). The risk difference metric gives large weight to trials with small proportions $P_T$ and $P_C$. The various summary

methods utilizing the odds ratio metric give large weight to trials with $P_T$ and $P_C$ near 0.50.

The consequences of the choice of effect measure have been investigated in several studies. Engels et al (2000) carried out an empirical study of 125 meta-analyses. They found no meta-analysis in which the summary risk difference and odds ratio were discrepant to the extent that one indicated significant benefit while the other indicated significant harm. However, they found that using risk differences led to more heterogeneity than using odds ratios. For 107 (86 per cent) of the meta-analyses, the Q-statistic P-value for the risk differences was less than that for the odds ratios. (sign test, $p < 0.0001$). For 18 meta-analyses (14 per cent) the risk differences were judged heterogeneous ((z-statistic $p < 0.10$) when the odds ratios were not, whereas for only three meta-analyses (2 per cent) were the odds ratios heterogeneous when the risk differences were not.

Deeks (2002) discusses the properties of the odds ratio, the risk ratio and the risk difference in detail. He points out that a choice of measure may give a lower Q-statistic not because its predictions are closer to the observed results for a set of trials, but because it gives outlying trials lower weight. Tang (2000) makes the same observation: "Outlier trials that are given very large weight and over-weighted in pooling the risk difference will thus be given very small weight and under-weighted in pooling the relative risk, and vice versa. The same outlier trials may thus affect differently the two combined measures of effect."

Deeks (2002) concludes that treatment effects tend to be more homogeneous across trials when expressed as relative rather than absolute effects. This seems to support Sterne and Egger's (2001) recommendation to employ relative measures like odds ratio or the risk ratio rather than the risk difference.

## 3.3 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. These two alternatives will be discussed in this chapter. A third alternative, to exclude outliers to remove heterogeneity will also be briefly discussed.

### 3.3.1 The random effects method

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 $\theta_i$ in the ith study is randomly selected from a normal distribution of studies with mean $\theta$. 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 $Var(u)=\tau^2$.

$\sigma_i^2$ is the within-study variance and $\tau^2$ 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 $\tau^2$.

As discussed for the fixed effects method, $\sigma_i^2$ is assumed to be known, ie the estimates $\hat{\sigma}_i^2$ 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 $\tau^2$ is less reasonable since the number of studies in a meta-analysis is usually fairly small. Normally the uncertainty of the estimate for $\tau^2$ will be considerable. All the same, the most common estimate for $\tau^2$ and the most common form for random effects analysis does not take the uncertainty of the estimate of $\tau^2$ 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 lack of precision in the estimate of does not necessarily affect the estimate of the overall treatment effect since the random effect weights in the formula above may be fairly independent of $\tau^2$, for example in the case where the fixed effects weights of the studies are nearly equal. Whether the value of $\tau^2$ used substantially affects the overall estimate of treatment effect can be investigated using a sensitivity plot of $\hat{\theta}$ against $\tau^2$, ie calculating the overall mean with different values of $\tau^2$ (Hardy and Thompson, 1996).

When calculating a confidence interval for $\theta$, the uncertainty in the estimate for $\tau^2$ can either be taken into account by employing a different distribution than the normal without explicitly estimating the uncertainty or estimating the uncertainty in $\tau^2$.

The next two sections describe estimates for $\tau^2$. Section 3.2.2 discusses estimates that does not include the uncertainty of $\tau^2$ while section 3.2.3 describes methods that calculate confidence intervals for $\tau^2$. Section 3.2.4 describes a method for constructing confidence intervals employing the t-distribution.

### 3.3.2 Estimates for $\tau^2$ without uncertainty calculation

The far most common estimate for $\tau^2$ 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 $\tau^2$ 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 $\tau^2$ and the weighted mean $\hat{\theta}$ are computed simultaneously. Equations for the solution are given in Hardy and Thompson (1996).

$\hat{\theta} = \dfrac{\sum \dfrac{\hat{\theta}_i}{(\hat{\sigma}_i{}^2 + \hat{\tau}^2)}}{\sum \dfrac{1}{(\hat{\sigma}_i{}^2 + \hat{\tau}^2)}}$ and $\hat{\tau}^2 = \dfrac{\sum \dfrac{(\hat{\theta}_i - \theta)^2 - \sigma_i{}^2}{(\hat{\sigma}_i{}^2 + \hat{\tau}^2)^2}}{\sum \dfrac{1}{(\hat{\sigma}_i{}^2 + \hat{\tau}^2)^2}}$. 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 $\tau^2$, $\hat{\theta}$ can be solved for. Using this value in the other equation a new value of $\tau^2$ is obtained, etc. When this estimate is used without taking the uncertainty of estimation into account, it will be referred to as simple likelihood.

A program to compute maximum likelihood estimates for $\theta$ and $\tau^2$ has been written in the programming language C.

### 3.3.3 Methods to calculate confidence intervals for $\tau^2$

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 $\tau^2$ as the value of $\hat{\theta}$ varies and $\hat{\theta}(\tau^2)$ is the MLE of $\theta$ as $\tau^2$ varies. A confidence interval for $\tau^2$ 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.

The C-program that computes maximum likelihood estimators also calculates confidence intervals by profile likelihood. It has been tested against data given in Hardy and Thompson's paper.

Biggerstaff and Tweedie (1997) derive an approximate distribution for $\hat{\tau}_M{}^2$, which is the non-truncated version of $\hat{\tau}_{DL}{}^2$, using an approximating gamma distribution for Q. They derive a formula for Var(Q) given by :

$$Var(Q) = 2(K-1) + 4(S_1 - \frac{S_2}{S_1})\tau^2 + 2\left(S_2 - 2\frac{S_3}{S_1} + \frac{S_2{}^2}{S_1{}^2}\right)\tau^4 \text{ where } S_r = \sum w_i^r$$

Given E(Q) and var(Q), Biggerstaff and Tweedie approximate the distribution of Q with a gamma distribution with shape parameter r and scale parameter λ. Based on this approximate distribution for Q, an approximate distribution for $\hat{\tau}_M{}^2$ is a location-shifted, scaled, gamma distribution. This density function is used to construct a CI for $\tau^2$.

To estimate the overall mean $\hat{\theta}$ the density for $\hat{\tau}_M^2$ is written as a density for $\hat{\tau}_{DL}^2$, $f_{DL}(t;\tau^2)$, consisting of a discrete and absolutely continuous part. When estimating the function $\hat{\tau}_{DL}^2$ is substituted for $\tau^2$.

The weights are then defined as $w_i^*(\tau^2) = E\left[w_i(\hat{\tau}_{DL}^2)\right]$.

Biggerstaff and Tweedie comment on the fact that the DerSimonian-Laird method gives too much weight to relatively small studies (with large variances $\sigma_i^2$) when $\hat{\tau}_{DL}^2$ is large. Their method gives more weight to large studies and less weight to small studies than the DerSimonian-Laird method. This implies that the *effective* value of $\tau^2$ used in the weights are considerably smaller than the estimate for $\tau^2$. The effective value also seems to be different in the weights of different studies.

Taking the uncertainty of the estimates of $\tau^2$ into account gives broader confidence intervals for the overall mean $\theta$ than when this uncertainty is ignored. The coverage probabilities‡ will therefore be higher and be a better approximation to the real confidence level. This has been confirmed by simulations described in section 3.3.5.

### 3.3.4 Using a t-distribution to calculate confidence intervals for $\theta$

Sidik and Jonkman (2002) suggest a new method of calculating a confidence interval for the random effects method using a t-distribution instead of a normal distribution as is common for the DerSimonian and Laird method.

They introduce the standard normally distributed statistic

$Z_w = \dfrac{\hat{\theta} - \theta}{\dfrac{1}{\sqrt{\sum w_i^*}}}$ and the statistic $Q_w = \sum w_i^*(y_i - \hat{\theta})^2$ which is $\chi^2$ distributed with (k-1) degrees of freedom. They show that $Z_w$ and $Q_w$ are independent and hence that $\dfrac{\sqrt{\left(\sum w_i^*\right)}(\hat{\theta} - \theta)}{\sqrt{\sum w_i^*(y_i - \hat{\theta})^2/(k-1)}}$ is t-distributed with (k-1) degrees of freedom. This t-distribution is used instead of the normal distribution for constructing a confidence interval for $\theta$.

In the derivation it is assumed that the true weights $w_i^* = \dfrac{1}{\sigma_i^2 + \tau^2}$ are known. For construction of confidence intervals, estimates for $\sigma_i^2$ and $\tau^2$ have to be used. It not made clear which estimate for $\tau^2$ is used but it seems to be the DerSimonian and Laird estimate.

Sidik and Jonkman (2002) show by simulation that confidence intervals based on the t-distribution have higher coverage probabilities than the DerSimonian and Laird method, particularly when the number of studies is small. At the same time the coverage

‡ The coverage probability is the probability that the true parameter value is included in the confidence interval. It should be 0.95 for a 95% confidence interval.

probabilities are also too small when using the t-distribution, again this is most pronounced for a small number of studies.

It will seen from the next section that the profile likelihood method is preferable.

### 3.3.5 Comparison of coverage probabilities for various meta-analytic methods

Brockwell and Gordon (2001) compared 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. The methods considered were the fixed effects method and the random effect method, used irrespectively of the value of Q, a Q-based method where the random effect method was employed if the value of Q was statistically significant using the $\chi^2$ test, the simple likelihood method and the profile log likelihood method.

The number of studies simulated varied between 3 and 35. The between studies variance, $\tau^2$ varied between 0.0 or 0.1. The variances of the studies, $\sigma_i^2$, were between 0.009 and 0.6.

As expected, the coverage probability of the fixed effects method was far too low when there is heterogeneity. For example, when the number of studies was 10 and $\tau^2$=0.1, the coverage probability was less than 0.65.

However, standard random effect methods also gave too low coverage probabilities. When the number of studies was 10 and $\tau^2$=0.1, the coverage probability was approximately 0.9 for the pure random effects method and approximately 0.85 for the Q-based method (values read from a figure in Brockwell and Gordon).

The simple likelihood method, somewhat unexpectedly, gave a smaller coverage probability than the pure random effect method. 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.

The profile likelihood and the simple random effects method gave slightly too large coverage probabilities when there was no heterogeneity, ie when $\tau^2$=0.0.

### 3.3.6 Fixed or random effects? Discussion

The simulation above shows that using a fixed effects model when there is heterogeneity leads to confidence intervals that are too narrow. Using random effects models will be more conservative and avoid too strong conclusions. However, Poole and Greenland (1999) show by examples that random-effects summaries can be farther from the null value and can have smaller p values, so they can appear more strongly supportive of causation or prevention than fixed effects summaries. One reason for this may be that random effect methods give larger weight to small studies and are therefore more susceptible to publication bias (see the next chapter). Thompson and Pocock (1991) believe that random-effects methods may give undue weight to small studies, emphasising poor evidence at the expense of good. Whether to use fixed effects methods or random effect methods has been contentious. In fact, Poole and Greenland refer to authors that find the inclusion of random effects "peculiar" or oppose it so strongly as to consider it 'wholly wrong'.

Another question is whether random effects should be used only after a test rejects homogeneity or without testing. One common recommendation is to test for heterogeneity and if the test rejects the null hypothesis of homogeneity, using a 10%

significance level due to the low power of the test, the random effects method should be employed. Other recommendations are also found. Higgins et al (2002) refer to The Cochrane Eyes and Vision Group, which recommends that for a p-value from the heterogeneity test greater than 0.10 a fixed effects model should be used; for a p-value between 0.05 and 0.10 both models should be used; and for a p-value less than 0.05 no meta-analysis should be performed. Others recommend that meta-analyses always use random effects models and thus include an estimate of the between-study variability regardless of the results of any test for heterogeneity (Takkouche, Cadarso-Suarez, and Spiegelman, 1999). After all, as pointed out by Normand (1999) it is almost always reasonable to believe that there is some between-study variation and few reasons to believe it is zero. Mosteller and Colditz (1996) also recommend using random-effects model because if there is no heterogeneity the estimate for $\tau^2$ will be zero, which is equivalent to a fixed effects model.

## 3.4 Explaining heterogeneity

Several authors stress the importance of explaining heterogeneity and not just to allow for it by random effect methods. Higgins et al (2002) holds that:"It is now generally accepted that meta-analysis should attempt to go beyond estimating a single average treatment effect. Reasons for heterogeneity should be explored in order to increase scientific understanding and clinical relevance". Greenland (1994a) goes further: "Estimates from random effects models are justifiable only after a thoughtful search for sources of the heterogeneity, and then only in a mixed model that includes fixed effects for important measured sources of heterogeneity". He also speaks of: "The mindless agglomeration of study results into a single summary estimate (with or without random effects) ". In another paper, Greenland (1994b) believes that: "Meta-analysis should be treated as a study of studies, rather than as a means for combining study results into a single effect".

Takkouche, Cadarso-Suarez, and Spiegelman (1999) also emphasize the importance of explaining heterogeneity: "It involves a decision about whether one should pool individual results into one summary measure or present separate results for subgroups only".

As mentioned earlier, heterogeneity can also be considered an opportunity. Berlin (1995) believes that because it involves comparisons across studies, meta-analysis can lead to insights when study design, exposure assessment or exposure levels, study populations, etc "are found to relate to study outcome". When treatments, therapies or measures vary between studies, this may provide estimates of the degree of benefit from a particular therapy and whether the benefit depends upon specific characteristics of the studies. This latter question capitalizes on the differences across studies; it may be of more interest to find a particularly effective treatment than in determining whether all studies, on average, involve effective treatments (Normand, 1999).

There is no doubt that the need for explanation of heterogeneity has broad support. This need is even more pronounced for observational studies. For observational, or quasi-experimental (Shadish, Cook and Campbell, 2002) studies, there is usually much larger methodological diversity and the effect of measures may vary with the design employed. This heterogeneity should also be explained, if possible.

The best and the most common method of explaining heterogeneity is using regression analysis with studies as the unit of analysis, so-called meta-regression. However, subgroup analysis is also employed. Subgroup analysis will not be separately discussed since this is conceptually very simple: classify the studies into subgroups by common characteristics and do a fixed effects or random-effects analysis within each subgroup.

Also, meta-regression can substitute for subgroup analysis and is preferable to subgroup-analysis for two reasons. A simple breakdown on groups, which can be handled in meta-

regression by dummies, may be improved on by introducing a variable describing the attributes of the groups.

Subgroup analysis and meta-regression have one thing in common, the danger of finding spurious associations or interaction through "data dredging". The next section discusses this problem.

### 3.4.1 False positive conclusions in subgroup analysis and meta-regression

Several authors warn against the problem of spurious findings. Thompson and Higgins (2002) regard false positive conclusions through 'data dredging' so important that one might label it as the principal pitfall in meta-regression. This is no less of a problem for subgroup analysis when groups are selected after looking at the data. Smith and Egger (2001) give an example where the treatment was said to be beneficial in patients under 65 but harmful in older patients. In subsequent studies these findings received no support. It can be shown that if an overall treatment effect is statistically significant at the 5% level ($P < 0.05$) and the patients are divided at random into two similarly sized groups, then there is a one in three chance that the treatment effect will be large and statistically highly significant in one group but irrelevant and non-significant in the other (Smith and Egger, 2001). Smith and Egger (2001) conclude that far from aiding clinicians, *post hoc* subgroup analyses may confuse and mislead.

Altman and Matthews (1996) also warn against post hoc subgroup analysis: "Exploratory examination of many such subgroups is almost certain to throw up some spurious significant interactions, and in practice we cannot tell if a specific interaction is real or spurious".

Berlin and Antman (1994), however, recommend to perform some stratified analyses for the purpose of understanding the data, suggesting what models may be appropriate, and illustrating the relations of interest. Colditz, Burdick and Mosteller (1995) believe that an initial stratification of results by study design is useful. A combined analysis should adjust for design features if there is heterogeneity across study designs or, alternatively, results should be reported separately for each design. This is, of course, acceptable as long as the subgroup analyses are considered as merely exploratory. Any statistical tests for testing the differences between subgroups will be invalid. If the significance of results is of interest, a Bonferroni adjustment to the significance level for each covariate can be carried out (Thompson and Higgins, 2002).

The general view, however, is that subgroups for a subgroup analysis or covariates for a meta-regression should be specified a priori. Higgins et al (2002) refer to three principal sources of guidance to people undertaking systematic reviews. All three sources clearly recommended that potential subgroup analyses be specified a priori. Higgins et al seem to believe that this is insufficient. They write: "Guidelines that address practical issues are required to reduce the risk of spurious findings from investigations of heterogeneity. This may involve discouraging statistical investigations such as subgroup analyses and meta-regression, rather than simply adopting a cautious approach to their interpretation, unless a large number of studies is available. The notion of a priori specification of potential effect modifiers for a retrospective review of studies is ill-defined…."

This is to go too far. As long as the danger of spurious associations is borne in mind, explaining heterogeneity by performing meta-regressions should be attempted. The next section discusses meta-regression.

### 3.4.2 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). Meta-regression can be used to investigate whether particular covariates (potential 'effect modifiers') explain any of the heterogeneity of treatment effects between studies (Thompson and Higgins, 2002) or to explain the study-to-study variation found in an empirical literature (Stanley, 2001). It is appropriate to use meta-regression to explore sources of heterogeneity even if an initial overall test for heterogeneity is non-significant since it is well known that this test often has low power and therefore a non-significant result does not reliably identify lack of heterogeneity (Thompson and Higgins, 2002).

The independent variables, covariates or factors, in a meta-regression are of two kinds, properties of the studies as such or average properties of the units studied. Examples of variables of the first kind are the country where the study has been carried out, study design, type of statistical analysis, etc. Examples of variables of the second kind are the average age of patients or the percentage males. A problem with this second kind is that relationship with patient averages across studies may not be the same as the relationship for patients within studies. This phenomenon is known as aggregation bias or ecologic bias (Greenland, 1987; Thompson and Higgins, 2002). Another problem is that aggregated values tend to exhibit little between-study variation, thus providing minimal information across the potential range of the factor (Schmid, 1999).

A problem with study characteristics is that studies may differ for other characteristics that have not been considered. These characteristics are confounding factors as in other observational studies. Thompson and Higgins (2002) view meta-regression as a study of the "epidemiology of trials". An association identified with one trial characteristic may in reality reflect a true association with other correlated characteristics, whether these are known or unknown. This problem is accentuated by the fact that the number of studies in a meta-regression is usually quite small.

For this reason, and because of the danger of "data-dredging" discussed above, a meta-regression should always be regarded as exploratory and the results tentative. That said, when there is an indication of heterogeneity, meta-regression should be carried out when possible.

#### Methods for meta-regression

Several papers describe methods for meta-regression or discuss their use, Berlin and Antman (1994), Berkey et al (1995), Thompson and Sharp (1999), Platt, Leroux and Breslow (1999) and Houwelingen et al (2002). Thompson and Sharp (1999), in particular, give an excellent introduction to meta-regression models.

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. If the weights are taken equal to the inverse of the within-study variance alone, the regression analysis is fixed effect.

The residual between-study variance $\tau^2$ is only known after a regression analysis has been done. A method for estimating the regression equation and $\tau^2$ simultaneously or

iteratively is therefore necessary. Thompson and Sharp (1999) describe four methods. The equations in their paper are given for the case of one covariate.

The simplest method does not require iteration, or more correct, just one iteration. A moment estimator of $\tau^2$ is derived from the heterogeneity statistic $Q = \sum w_i \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2$ where $\hat{\alpha}$ and $\hat{\beta}$ are estimated by fixed effects regression. $\tau^2$ is estimated by

$$\hat{\tau}^2 = \frac{Q - (k-2)}{F(w,x)} \text{ if Q>k-2, 0 if Q} \leq \text{k-2}$$

where $F(w,x) = \sum w_i - \frac{\sum w_i{}^2 \sum w_i x_i{}^2 - 2\sum w_i{}^2 x_i \sum w_i x_i + \sum w_i \sum w_i{}^2 x_i{}^2}{\sum w_i \sum w_i x_i{}^2 - \left(\sum w_i x_i\right)^2}$

Then a random effects regression is carried out with this $\hat{\tau}^2$ in the weights with no further iterations.

A maximum likelihood (ML) estimate is obtained by an iteration procedure which alternates between estimating the regression coefficients with random effects regression and estimating $\tau^2$ by $\tau^2 = \frac{\sum w_i^{*2}\left[\left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 - \frac{1}{\hat{\sigma}^2{}_i}\right]}{\sum w_i^{*2}}$. The weights are given by $w_i^* = \frac{1}{\hat{\sigma}_i{}^2 + \hat{\tau}^2}$. Starting with a value $\hat{\tau}^2 = 0$, ie for the first iteration the fixed effects model is employed, the iteration procedure continues until convergence.

The other two iterative methods described by Thompson and Sharpe (1999), restricted maximum likelihood (REML) and an empirical Bayes estimate, have slightly different formulas for the estimate of $\tau^2$, for the restricted maximum likelihood: $\tau^2 = \frac{\sum w_i^{*2}\left[\frac{k}{k-2}\left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 - \frac{1}{\hat{\sigma}^2{}_i}\right]}{\sum w_i^{*2}}$

and for the empirical Bayes method: $\tau^2 = \frac{\sum w_i^{*}\left[\left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 - \frac{1}{\hat{\sigma}^2{}_i}\right]}{\sum w_i^{*}}$.

The first differs from the maximum likelihood by the inclusion of the factor k/(k-2) and the second by using the weights instead of the square of the weights.

The four methods have all been implemented at the Institute of Transport Economics employing the econometric package LIMDEP. Which of the iterative methods is preferable in not clear. Thompson and Sharp (1999) states that the use of restricted maximum likelihood overcomes the tendency of ML methods to underestimate variances. However, Berkey et al (1995) found that restricted maximum likelihood consistently tended to underestimate $\tau^2$ in simulations they have carried out (see below).

All four methods take $\tau^2$ (and $\sigma_i{}^2$) as known and equal to its estimated value, ie the uncertainty of the estimate is not taken into account. The precision of the regression

parameters is therefore smaller than indicated by the regression results and using the normal distribution for confidence intervals or testing the parameters gives a false impression of precision. When the number of studies is limited, which is often the case in meta-analyses, the estimate of $\tau^2$ is imprecise and taking the uncertainty into account may lead to a noticeable increase in the confidence intervals. The uncertainty in the estimate of $\sigma_i^2$ is less of a problem since the number of subjects in a study is normally far larger than the number of studies in a meta-analysis.

Normally, uncertainty in the estimates of parameters in a normal model is handled by employing the t-distribution with the appropriate number of degrees of freedom. Berkey et al (1995) used simulation to determine this number empirically for meta-regression. They employed the empirical Bayes estimate for $\tau^2$ and found that the number of degrees of freedom should be reduced by 3 compared to the normal way of calculating the number of degrees of freedom. For a regression equation a constant and one coefficient they found that confidence interval with k-5 degrees of freedom gave the best coverage probability.

Berlin and Antman (1994) discuss the problem of collinearity for meta-regression. They discuss symptoms of collinearity and recommend that the correlation matrix showing the associations between pairs of predictors should always be calculated. Belsley (1990) describes a method for diagnosing collinearity and identifying the variables contributing, based on the eigenvalues of the normalized covariance matrix. This method has been implemented at the Institute of Transport Economic using the econometric package LIMDEP.

The methods described above are general, based on an assumption of the effect measure being (approximately) normally distributed. Random clinicial trials where the effect is derived from 2 X 2 tables can be handled by these methods since the logodds (the logarithm of the odds ratio) is approximately normal. However, there are methods to deal with the binomial structure directly. Thompson and Sharp (1999) discuss several varieties of logistic regression, both with and without residual heterogeneity (random effects). The random effects logistic regression has been implemented at the Institute of Transport Economic using the econometric package LIMDEP.

Normally, employing the approximately normally distributed logodds is acceptable. Thompson and Sharp (1999) conclude that logistic regression will in practice be likely to give similar results as the normal approximation. Only when all trials are small would it be necessary to take the binary nature of the data into account.

Platt, Leroux and Breslow (1999) use simulation to compare penalized quasi-likelihood (PQL) with the use of the logodds. Like Thompson and Sharp (1999), they conclude that it is reasonable to recommend the use of the transformation method (ie logodds) when the data are not sparse. For smaller numbers in the tables, they recommend PQL.

Higgins and Thompson (2003) evaluate, through Monte Carlo simulation under the null hypothesis of no true association, some properties of methods for meta-regression. In particular, they address the following questions: 1. How is the likelihood of a false-positive regression coefficient in meta-regression influenced by (i) true extent of heterogeneity, (ii) the number of studies, (iii) the relative weights awarded to studies, (iv) the number of covariates investigated, and (v) the extent of correlation between covariates? 2. How is the likelihood of a false-positive regression coefficient in meta-regression influenced by the method used?

The methods compared are the following:

1. Fixed effects meta-regression relating the test-statistic $t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$ to the standard normal distribution.
2. Random effects meta-regression with the REML estimate of heterogeneity variance $\tau^2$ described earlier relating the statistic t to the standard normal distribution.
3. Random effects meta-regression with the REML estimate of heterogeneity variance $\tau^2$ but relating a modified statistic due to Knapp and Hartung (2003) to the t-distribution with k-m-1 degrees of freedom where m is the number of covariates.

Data were simulated under the null hypothesis of no association between effect estimates and any covariate but with an unexplained component of heterogeneity according to the standard random effects meta-analysis model.

Parameters that were varied across simulation runs are the number of studies (k), the number of covariates (m), the correlation between covariates (ρ), the extent of genuine heterogeneity ($\tau^2$) and the set of weights ($w_i=1/\sigma_i^2$). Higgins and Thompson (2003) investigate k = 5, 10 and 100 studies; m = 1, 3 and 5 covariates; correlations ρ =0, 0.5 and 0.9 (when m = 3 or 5); and heterogeneity variances $\tau^2$ = 0,1 and 5.

For the within-study variances, $\{\sigma_i^2\}$, three patterns were selected, equal within-study variances, variable within-study variances and unbalanced within-study variances. These were assigned so that four-fifths of the trials have identical within-study variances and the remaining fifth are 'mega-trials' with 1/20 of the variance.

The numerical values for the variances were determined by making use of the 'typical' within-study variance described above (Higgins and Thompson, 2002):

$\hat{\sigma}^2 = \frac{(k-1)\sum w_i}{\left(\sum w_i\right)^2 - \sum w_i^2}$. $\hat{\sigma}^2$ is set to 1 in all cases so that the average size of a study is similar across all simulated meta-analyses. This also allows us to quantify the heterogeneity using a statistic $I^2 = \frac{\tau^2}{\tau^2 + \hat{\sigma}^2}$ that describes the proportion of total variation in effect estimates attributable to heterogeneity rather than within-study variability. For $\tau^2$ = 0, $I^2$= 0%, for $\tau^2$ = 1, $I^2$ = 50% and for $\tau^2$ = 5, $I^2$ =83%.

For a single covariate, fixed effects meta-regression gave the correct significance level in the absence of heterogeneity but was found to be to be unacceptable in the presence of heterogeneity. The observed false-positive rate is typically 20% for a moderate amount of heterogeneity ($I^2$=50%, se above for the definition of $I^2$) and can exceed 50% for extreme heterogeneity ($I^2$=83%).

For the random effects methods, in the absence of any heterogeneity only the standard normal test achieves the correct 5% level. In the presence of heterogeneity, the rate of false-positive findings, using the standard normal distribution increases substantially above 5% for small numbers of studies. The t-test of and Knapp and Hartung gives false positive rates below 5%, although it achieves the desired level approximately when there is substantial heterogeneity.

False-positive rates from multiple meta-regression analyses were found to be largely unaffected by correlation between covariates.

To achieve more correct significance levels Higgins and Thompson develop a permutation test. The permutation test is constructed by randomly shuffling the rows and

re-assigning them to the response vector. The linked pair of an effect estimate and its variance is taken as the 'response'. On each random re-allocation of covariates to responses, a test statistic is computed. The true significance level for the relationship between response and a particular covariate is determined by comparing an observed test statistic for the original data set with the distribution of test statistics across random re-allocations. For example, if in 12 out of 1000 re-allocations the original test statistic is equaled or exceeded, then the permutation test t-value is 0.012.

Higgins and Thompson (2003) describe algorithms for the permutation test for both a single and for several covariates. This algorithm has not yet been implemented at the Institute of Transport Economics but implementation has high priority.

Higgins and Thompson (2003) found that using a specific data set, among random effects analyses, test results using the approach of Knapp and Hartung are in good agreement with the permutation test p-values, both being considerably more conservative than the test based on a standard normal distribution for the statistic.In one example they found that for a single 'most significant' covariate among five, p=0.0006 for the random meta-regression based on the normal distribution, 0.03 based on the Knapp and Hartung method but increased to p = 0.11 in a permutation test in which distribution of 'most significant' findings from multiple regression was examined.

An alternative to the methods discussed are multi-level or hierarchical models. These are recommended both by NCR (1992) and Sutton et al (1999) referred in Higgins et al (2002). As yet, we have not been able to investigate hierarchical models.

### 3.4.3 Removing outliers

Analysts sometimes identify heterogeneity and deal with it by excluding studies until a satisfactory degree of homogeneity is achieved (Colditz, Burdick and Mosteller, 1995). Petitti (2001) (see the next page) also found meta-analyses where outlier studies where removed.

This practice should be avoided. In fact, Colditz, Burdick and Mosteller (1995) call it dangerous. Firstly, different methods and different measures may identify different studies as outliers (Song, 1999; Deeks 2002). Secondly, and worse, is that it may bring subjectivity into the meta-analysis. Studies that do not support the researcher's preconceived notions may be excluded as outliers.

### 3.4.4 Current practice of dealing with heterogeneity

A few studies have examined current practice in addressing heterogeneity in meta-analyses. Colditz, Burdick and Mosteller (1995) found that only eight of 26 studies in 1991 and seven of 29 studies in 1992 reported any assessment of heterogeneity.

Higgins et al (2002) carried out a study with the objectives: to collate recommendations on the subject of dealing with heterogeneity in systematic reviews of clinical trials; to investigate current practice in addressing heterogeneity in Cochrane reviews; and to compare current practice with recommendations. They found that fixed effects meta-analysis was preferred in practice. Among the 21 reviews that performed (or preferred) only fixed effects meta-analyses, four did this despite having outcomes with statistically significant heterogeneity ($p<0.05$).

No reviews discussed problems associated with random effects analyses of small numbers of studies (in which case it is difficult to estimate precisely the between-trial variation) and few reviews that looked at modifiers studied only those effect modifiers they had pre-specified, indicating the practical limitations of such a procedure. Meta-analyses may

therefore be expected to give too narrow confidence intervals for random effect models and to present spurious associations or regression coefficients.

A study of Petitti (2001) had similar objectives as Higgins et al (2002), viz "to assess whether tests of statistical heterogeneity were done, whether the results were reported, and how a finding of significance for a test of statistical heterogeneity was handled and the results interpreted". Petitti restricted the study to reviews of reproductive health topics. This field "is especially interesting because the application of the principles of meta-analysis to reproductive health pre-dates by several years the widespread application in other medical fields, and one might expect that the field would be more technically advanced" (Petitti, 2001). Tests of statistical heterogeneity were not done universally, for 23 of 32 meta-analyses (72 per cent), statistical tests of heterogeneity were reported to have been done (Two meta-analyses of special data are not included in this figure because there was no heterogeneity test for these types of meta-analysis). Of the 23 meta-analyses that did tests of statistical heterogeneity, four used exclusion of one or more outlier studies to eliminate statistical heterogeneity. In all four cases, the exclusion appeared to be post hoc.

Altogether, 16 of the 34 (47 per cent) meta-analyses of reproductive health topics explored heterogeneity. Two used meta-regression.

Although the consensus appears to be that heterogeneity tests are conservative for meta-analysis of studies and a probability value of 0.10 is preferred, many meta-analyses used the conventional value of 0.05 without providing a reason.

## 3.5 Heterogeneity. Recommendations

Many authors have given guidelines or recommendations for the handling of heterogeneity. For example, the US National Research Council panel (1992) believes that "CI (Combining Information) modeling would be improved by the increased use of random effect models in preference to the current default of fixed effects models".

Higgins et al (2002) sum up their collation of guidelines in the following way: "Generic guidance for addressing heterogeneity in systematic reviews of clinical trials appeared to be fairly consistent, with the key features being a priori specification of potential effect modifiers and planned subgroup analyses, recommendations to examine heterogeneity with regard to both a priori and post hoc specified effect modifiers (though with particularly cautious interpretation of the latter) and cautious use of fixed effects and random effects analyses (or a comparison of the two) when heterogeneity is present but cannot adequately be explained". Considering that one author has written papers stressing the importance of examining heterogeneity (Thompson, 2001), they are surprisingly lukewarm when it comes to recommending such an examination, "advocating a cautious examination of potential causes of heterogeneity" and suggesting "that in many cases it is not clear that sources of heterogeneity should be investigated unless a large number of studies is available". The risk of spurious findings due to the post hoc selection of covariates is probably the reason for this caution.

What if heterogeneity cannot be explained? Even a strong critic of random effects methods as Greenland accepts that in this case the random effects method may be the preferred choice: "If a large amount of unexplained heterogeneity (as measured by, say $\chi^2_{r,}$) remains after regression modelling and diagnostics, one may consider turning to random-effect models" (Greenland, 1987). Dickersin and Berlin (1992) are more dubious: "If the heterogeneity cannot be explained by study design features or populations, then the appropriateness of a pooled estimate of effect should be questioned".

Our recommendations are as follows:

1. If possible employ random effects meta-regression to explain heterogeneity. Use the permutation method of Higgins and Thompson to reduce spurious associations.
2. If meta-regression is not feasible, use the random effects method. The fixed effects will then be a special case of the random effects method when the estimate for $\tau^2$ is zero. The between studies variance $\tau^2$ should preferably be estimated by the profile maximum likelihood so that the uncertainty of the estimate is taken into account when calculating the uncertainty of the pooled effect estimate.

   The main purpose of employing the random effects method is to avoid confidence intervals that are too narrow, not to obtain a more correct estimate than with the fixed effects method. If estimates with the fixed effects and the random effects method differ significantly, no pooled estimate should be given.

## 3.6 References

Altman Douglas G and John N S Matthews. Statistics Notes: Interaction 1: Heterogeneity of effects. *BMJ* 1996; 313:486 (24 August).

Baujat, B C Mahe, J P Pignon, et al. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Statistics in medicine,* 21 (18): 2641-2652 Sep 30 2002.

Belsley, D A (1991). Conditioning Diagnostics, Collinearity and Weak Data in Regression. New York: *Wiley.*

Berkey, C S, D C Hoaglin, F Mosteller and G A Colditz. A random-effects regression modell for meta-analysis. *Statistics in medicine,* Vol. 14. 395-411 (1995).

Berlin J A. Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *American Journal of Epidemiology*, 142, 383-387 (1995).

Berlin, Jesse A and Elliot M Antman. Advantages and limitations of metaanalytic regressions of clinical trials data. The online journal of *Current Clinical Trails*, Vol 3, 1994.

Biggerstaff, B J and R L Tweedie. Incorporating variability in estimates of heterogeneity in the random effect modell in meta-analysis. *Statistics in medicine,* 1997, vol. 16, 753-768.

Brockwell, Sarah E and Ian R Gordon. A comparison of statistical methods for meta-analysis. *Statistics in medicine,* 2001; 20:825-840.

Colditz, Graham A, Elisabeth Burdick and Frederick Mosteller. Heterogeneity in Meta-analysis of Data from Epidemiologic Studies: A Commentary. *American Journal of Epidemiology*, Volume 142 Number 4 August l5,1995.

Deeks, Jonathan J. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in medicine* 2002; 21:1575-1600.

Dickersin, K and J Berlin. Meta-analysis: state-of-the-science. *Epidemiol Rev* 1992; 14: 154- 76.

Elvik, Rune. Cost-benefit analysis of ambulance and rescue helicopters in Norway: reflections on assigning a monetary value to saving a human life. *Applied Health Economics and Health Policy,* 2002:1(2) 55-63.

Engels, Eric A, Christopher H Schmid, Norma Terrin, Ingram Olkin and Joseph Lau. Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in medicine,* 2000; 19:1707-1728.

Feinstein, Alvan R. Meta-analysis: statistical alchemy for the 21 st century. *Journal of Clinical Epidemiology* 48, No 1, pp 71-79, 1995.

Galbraith, R. A note on the graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, 1988, 7:889-894.

Glasziou, P P and S L Sanders. Investigating causes of heterogeneity in systematic reviews. *Statistics in medicine,* 2002; 21:1503-1511.

Greenland, Sander. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews* 9, 1-30(1987).

Greenland, Sander. Can meta-analysis be salvaged? *Am J Epidmiol* 1994; 140: 753-7, 1994a.

Greenland, Sander. Invited commentary: A Critical Look at Some Popular Meta.Analytic Methods. *American Journal of Epidemiology,* Vol 140, No, 3 , 1994b.

Hardy, Rebecca J and Simon G Thompson. A likelihood approach to meta-analysis with random effects. *Statistics in medicine,* vol. 15, 619-629 (1996).

Hardy, Rebecca J and Simon G Thompson. Detecting and describing heterogeneity in meta-analysis. *Statistics in medicine,* 17, 841-856 (1998).

Higgins, Julian P T and Simon G Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine,* 2002; 21:1539-1558.

Higgins, Julian, Simon G Thompson, Jonathan Deeks and Douglas Altman. Statistical heterogeneity in systematic reviews of clinical trials : A critical appraisal of guidelines and practice. *J Health Serv Res Policy.* Vol 7 No 1 January 2002.

Higgins, Julian and Simon G Thompson. Controlling the risk of spurious findings from meta-regression. *Statistics in medicine,* in press, 2003.

Houwelingen, Hans C van, Lidia R Arends and Theo Stijnen. Advanced methods in meta-analysis : Multivariate approach and meta-regression. *Statistics in medicine,* 2002; 21: 589-624.

Knapp G and Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med* 2003; 22: 2693-2710.

L'Abbe K, A Detsky and K O'Rourke. Meta-analysis in clinical research. *Ann Intern Med* 1987; 107:224-33.

Lincoln E Moses, Frederick Mosteller and John H Buehler. Comparing results of large clinical trials to those of meta-analyses. *Statistics in medicine,* 2002; 21:793-800.

Matt, George and Thomas D Cook. Threats To The Validity Of Research Syntheses In: Cooper, H, and L V Hedges (Eds.) 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Mosteller Frederick and Graham A Colditz. Understanding research synthesis (meta-analysis). *Annu Rev. Public Health 1996*. 17:1-23.

National Research Council. Combining information. Statistical issues and opportunities for research. *National Academy Press.* Washington, D.C. 1992.

Normand, Sharon-Lise T. Tutorial in biostatistics - analysis formulating, evaluating, combining, and reporting. *Statistics in medicine,* 18,321-359 (1999).

Petitti D B. Approaches to heterogeneity in meta-analyis. *Statistics in medicine,* 2001;20:3625-3633

Platt, Robert W, Brian G Leroux and Norman Breslow. Generalized linear mixed models for meta-analysis. *Statistics in medicine,* 18, 643-654 (1999).

Poole C and S Greenland. Random-effects meta-analyses are not always conservative. *Am J Epideimol* 1999;150:469-75.

Schmid, Christopher H. Exploring heterogeneity in randomized trials via meta-analysis. *Drug Information Journal*, Vol. 33, pp. 211-224, 1999.

Shadish, William R, Thomas D Cook and Donald T Campbell. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company, Boston, New York, 2002.

Sidik, K and J N Jonkman. A simple confidence interval for meta-analysis. *Statistics in medicine* 21 (21): 3153-3159 NOV 15 2002.

Smith, George Davey and Matthias Egger. Going beyond the grand mean: subgroup analysis, in meta-analysis of randomized trials. In: Mathias Egger, George Davey Smith and Douglas G Altman: *Systematic reviews in Health care. Meta-analysis in context*. BMJ books, Second edition 2001.

Song, Fujian. Exploring heterogeneity in meta-analysis: Is the L'Abbe Plot useful? *J Clin Epidemiol* 1999;52:725-30.

Stanley, T D. Wheat From Chaff. Meta-Analysis As quantitative literature Review. *Journal of Economic Perspectives*, volume 15, number 3, Summer 2001, Pages 131-150

Sterne JAC and M Egger. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal Of Clinical Epidemiology* 54 (10): 1046-1055 Oct 2001.

Sutton A J, D R.Jones, K R Abrams, T A Sheldon and F Song. Systematic reviews and meta-analysis, a structured review of the methodological literature. *Journal of Health Services Research & Policy* 1999; 4: 49-55.

Takkouche, Bahi, Carmen Cadarso-Suarez, and Donna Spiegelman. Evaluation of Old and New Tests of Heterogeneity in Epidemiologic Meta-Analysis. *Am J Epidemiol* 1999:150:206-15.

Tang, Jin-Ling. Weighting bias in meta-analysis of binary outcomes. *Journal of Clinical Epidemiology* 53. (2000) 1130-1136.

Thompson, S G. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research*, 2, 173-192 (1993).

Thompson, Simon G. Why and how sources of heterogeneity should be Investigated. In: Mathias Egger, George Davey Smith and Douglas G Altman: *Systematic reviews in Health care. Meta-analysis in context.* BMJ books, Second edition 2001.

Thompson, Simon G and Stuart J. Pocock. Can meta-analyses be trusted? *The Lancet,* Vol 338: Nov 2, 1991.

Thompson, Simon G and Julian P T Higgins. How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine,* 2002; 21:1559-1573.

Thompson, Simon G and Stephen J Sharp. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in medicine*, 18, 2693-2708 (1999).

Whitehead, Anne and John Whitehead. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in medicine*, vol. 10, l665-1677 (1991).

# 4 Publication bias

Statistical estimation is based on the assumption of an unbiased sample. A biased sample leads to a biased estimate. Meta-analysis is a form of statistical estimation of an effect based on a sample of studies of the effect. For unbiased estimation the sample of studies must be unbiased.

Not all studies are published. 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*.

Published studies are easier to retrieve than unpublished studies. For this reason, and because published studies are believed to be of better quality due to the peer review process, meta-analyses tend to be based on published studies. If there is publication bias, the results from meta-analyses will themselves be biased. In addition, publication bias may have substantial effects on estimation of the between-study variance even when the estimate of the mean is not strongly affected (Champney, 1983, referred in Sutton et al, 2000).

Given the risk of publication bias, meta-analyses may lead to spurious precision. The aggregated sample size may be very large, providing estimates that are apparently accurate and precise but are none-the-less biased (Begg and Berlin, 1994).

Systematic reviews should therefore investigate the possibility of publication bias. This is not always done. Sutton et al (2000) found that methods for testing publication bias were used in 33 of 132 meta-analyses and in none of 61 narrative systematic reviews. Of reviews with significant or positive conclusions, 48% of the meta-analyses and 12% of the narrative systematic reviews discussed or tested for publication bias.

Steps should also be taken to minimize the occurrence of publication bias. However, for the practising meta-analyst the detection of publication bias is more important. The emphasis in this chapter is therefore on methods to detect publication bias. What is known about the existence of publication bias and other biases in published material is presented first. For a discussion of ways of reducing publication bias it is referred to Song et al (2000).

## 4.1 Bias in published studies

In addition to publication bias, several other related form of bias exist. Sutton et al (2000) list the following:

1. Pipeline bias – studies with null or non-significant results take longer to publish than those with significant results.
2. Subjective reporting bias – studies are published but only the most significant results included in the report
3. Duplicate reporting bias – the same results are reported in multiple sources
4. Language bias - potentially a problem when meta-analyses restrict the languages of published reports included in a review, because there is evidence that non-native

English-speaking researchers are more likely to publish non-significant findings in their native tongue, compared to significant results

Song et al (2000) uses the name "outcome reporting bias" for subjective reporting bias above and "multiple reporting bias" for duplicate reporting bias. They also introduce the name "citation bias" if the chance of a study being cited by others is dependent on its results. They refer to a study by Ravnskov (1992) who investigated the citation frequency of cholesterol-lowering trials. It was found that supportive trials were cited almost six times more often than others (the mean annual number of citations was 40 versus 7.4).

Trying to include studies in other languages than English does not necessarily mean that bias is reduced. According to Thornton and Lee (2000), certain countries only publish positive results, so that including papers in all languages may actually introduce more bias into a meta-analysis.

The biases listed by Sutton et al (2000) can all be considered sub-groups of publication bias. In the following these forms of bias will not be discussed separately.

The most common reason for publication bias is that studies that do not find a significant effect are not published (empirical indication of this is described in section 4.2). An indication of publication bias is that small studies show a larger effect than larger studies. The explanation for this is as follows: The smaller a study, the larger the treatment effect necessary for the results to be statistically significant and therefore published. In addition, the greater investment of money and time in larger studies means that they are more likely to be of high methodological quality and published even if their results are negative (Sterne, Gavaghan and Egger, 2000). Therefore, the small studies that are published will show larger effects. If the true effect is zero the small studies will show a negative effect as often as a positive effect and there will be no bias. But if there is a small positive effect or negative effects are less likely to be published, small studies will be predominantly positive and there will be publication bias. Sterne, Gavaghan and Egger (2000) found that smaller studies tended to show greater treatment effects than larger studies in 69% of a set of meta-analyses studied.

Sterne, Gavaghan and Egger (2000) suggest the term "small-study effects" to describe a trend for the smaller studies in a meta-analysis to show larger treatment effects. Small-study effects may, however, have other explanations than publication bias. One possibility is that the effect really is larger in small studies than in larger studies. Sterne, Gavaghan and Egger (2000) suggest possible mechanisms for this. Effects may be larger when high-risk patients are studied. Trials conducted in high-risk patients will also tend to be smaller, because of the difficulty in recruiting such patients. Interventions may have been implemented less thoroughly in larger trials, thus explaining the more positive results in smaller trials. Also, a larger study is necessary if a small effect is expected.

In the last case, when the size of a study is dependent on the expected effect, for example when the necessary study size is based on a power calculation, a small study effect will of course not be an indication of publication bias. An assumption that power calculations are not a major or at least not decisive factor in the determination of sample size must therefore be considered to be an underlying assumption for the statistical methods for investigating publication bias described below.

If power calculations are a decisive factor in the determination of sample size and there is a basis for investigators prior expectation as to effect so that there is a negative correlation between sample size and effect, statistical methods for investigating publication bias will not work. But the consequences are more serious than that. Meta-analyses will generally be biased. Large effects will have too little weight in the analysis and small effect too large. This will underestimate the effect.

If a "small study effect" is due to publication bias a meta-analysis will overestimate the effect, if it is due to power calculations a meta-analysis will underestimate the effect. A number of studies have given indications of publication bias. It is time to investigate empirically how important power calculations are.

Sterne, Egger and Davey Smith (2001) note that small trials are generally conducted before larger trials are established. In the intervening years standard (control) treatments may have improved, thus reducing the relative efficacy of the experimental treatment.

Sutton at al (2000) point out that other sources of bias, such as those introduced by including studies of varying quality, can produce similar symptoms to publication bias. However, for varying quality to introduce a small-study effect the quality of a study will have to be correlated with its size.

These possible mechanisms show that a small-study effect does not necessarily entail publication bias. But the assumed association between publication bias and sample size is the cornerstone of many methods for detecting publication bias (Song et al, 2000). Also, as the next section will show, publication bias seems to be ubiquitous. Therefore, small study-effects will be taken as a likely sign of publication bias and the tests described in a later chapter, which really tests for small-study effects, will be considered tests for publication bias.

## 4.2 Empirical studies of publication bias

### 4.2.1 The significance level in published studies

Sterling (1959), referred in Song et al (2000), found that the results of 97% of studies published in four major psychology journals were statistically significant, concluding that non-significant results are under-represented.

Song et al refer 12 other studies of the same kind. One found that the percentage of significant results was 35%, but apart from this the lowest percentage was 71.

Sterling, Rosenbaum and Weinkam (1995) repeated Sterling's 1959 study with eight psychological journals and three medical journals. They found that of articles that used statistical tests, 96% rejected $H_0$ in psychological journals and 85% in medical journals.

Sterling, Rosenbaum and Weinkam (1995) also make assumptions on the power of the statistical tests employed and on the proportion of the scientific hypothesis tested for which the null hypothesis is really false, and do an approximate calculation of the percentage of studies that should reject the null hypothesis. They conclude that the percentage of tests rejected should be considerably less than 80. The large percentage of null hypotheses rejected is therefore an indication of publication bias.

Murtaugh (2002) puts forward the hypothesis that if the magnitude and statistical significance of an estimated effect influences the likelihood that a study's results will be published, then one may expect to find the strongest results in the journals of highest quality. He tests this hypothesis by using the impact factor of a journal, measured by the average number of times a paper is cited, as a measure of quality. He finds that out of ten analysed relations between quality and strength of results, nine were positive, whereof two significantly so. Murtaugh then extrapolates to unpublished studies. These will have even weaker results and therefore the association between journal quality and strength of results indicate publication bias.

This result is supported by other studies. Simes (referred in Sutton et al, 2000) found that positive trials (indicating a significant survival difference) appeared in prominent journals such as the NEJM and *Cancer*, while less widely circulated journals published only negative trials. In a study by Misakian and Bero (1998) one investigator stated explicitly

that they chose a less prestigious journal to publish their statistically non-significant results.

These results can be considered indirect evidence of publication bias. Another example of indirect evidence is the small-study effect described above, ie the fact that small studies show larger effects than larger studies. The use of the funnel plot and statistical methods based on the small-study effect therefore entails only indirect evidence of publication bias.

The small-study effect is in practice the most important indicator of publication bias since this indicator will be the only method available when carrying out a meta-analysis and the only information is the effect sizes and their variance. Methods for investigating small-study effects will be described below.

The direct evidence for significant results being published more often are of three kinds (Song et al, 2000):

1. Surveys of investigators
2. Comparisons between published and unpublished studies
3. Follow-up of cohorts of registered studies

A fourth kind of evidence comes from studies of the behaviour of referees and editors.

In the following a few studies showing publication bias are described. As we are more interested in methods for detecting publication bias in a particular sample of studies than to show that publication bias exists, these studies have not been selected in any systematic way. The evidence of publication bias is therefore more extensive than these few studies. On the other hand, publication bias may also be a problem for studies on publication bias. Studies finding no publication bias may not be published.

### 4.2.2 Surveys of investigators

Coursol and Wagner (1986) sent a questionnaire to a random sample of members of the American Psychological Association. The decision to submit a paper for publication was significantly related to outcome. It was decided to publish for 82% of the positive results and 43% of the negative.

Greenwald (1975) (cited in Sohn, 1996) found that researchers were eight times more likely to submit a manuscript for publication if the results were positive than they were if the results were negative.

Misakian & Bero (1998) asked investigators to indicate reasons for unpublished results. The reasons stated most frequently were ongoing data collection or analysis (n=33 times), lack of time (n = 26), and competing priorities (n = 11). Statistically non-significant results were cited as a reason for failure to publish for only 2 studies. In spite of this, the median time to publication was 5 years (95% confidence interval 4-7 years) for statistically non-significant studies and 3 years (3-5 years) for statistically significant studies. This finding cannot be attributed to differences in the sample sizes, funding sources, or health outcomes measured.

### 4.2.3 Follow-up of cohorts of registered studies

Dickersin, Min and Meinert (1992) referred in (Cook et al, 1993) conducted a survey of 737 studies approved by institutional review board serving The John Hopkins Institutions before or during 1980. Significant studies were considerably more likely to be published (odds ratio, 2.54). The bias was due to non-significant trials not being submitted for publication rather than their being rejected once submitted. Similar findings come from a

survey of 487 research projects approved by the Central Oxford Research Ethics Committee between 1984 and 1987 (Easterbrook et al, 1991, referred in Cook et al, 1993). Studies with statistically significant results were more likely to be published than those finding no difference between study groups (odds ratio, 2.32). Neither survey found methodological differences between published and unpublished work.

Dickersen and Min (1993) investigate studies approved by institutional review board serving The Johns Hopkins Institutions during 1980. This is a subset of the studies investigated in Dickersin, Min and Meinert (1992). Clinical trials funded by the National Institute of Health and the studies included in Easterbrook et al (1991) are also analysed. Employing all studies, the odds ratio for the association between significant results and publication was 2.33 (2.13-3.90). Restricting the analysis to clinical trials the odds ratio increased to 5.96 (2.33-15.22). For randomised trials the odds ratio was 8.92 (1.96-40.65). The more rigorous the study, the more important it seems to be that the results are statistically significant for the study to be published. The confidence intervals for the odds ratios are large, however, so the trend is not significant.

In Finland, 274 notifications of the commencement of a clinical drug trial were received in 1983 by the National Agency for Medicine. By the end of 1993, 68 of these 274 trials reported their results; It was found that the rate of reporting was 38% for trials with positive results, 18% for those with inconclusive results, and 20% for those with a negative outcome ($p = 0.023$) (Bardy, 1998, cited in Song et al 2000).

### 4.2.4 Comparisons between published and unpublished studies

Sohn (1996) refer two studies showing a difference in the effect size for published and unpublished studies. Lipsey and Wilson (1993), in a meta-analysis of 92 meta-analyses of outcome research in the areas of psychotherapy and education, found that the average effect size was .53 and .39 for published and unpublished research, respectively. In their meta-analysis of psychotherapy outcome research, Smith, Glass, and Miller (1980) compared effect sizes of published and unpublished findings. They found that the effect sizes for journals and dissertations were .87 and .66, respectively.

### 4.2.5 Behaviour of referees and editors

Coursol and Wagner (1986), referred above, also found that of the papers that were submitted, positive papers were more likely to be accepted. The acceptance rate was 80% for positive studies and 50% for negative studies. The overall publication rate was then 66% for positive studies and 22% for negative.

Chalmers (1991) gives an anecdotal example of journals' preference for significant results, also for meta-analyses. A meta-analysis of RCTs of type I anti-arrhythmic drugs had revealed a detrimental effect on mortality that was not statistically significant. When a large trial reported in a preliminary fashion documented a statistically significant detrimental effect the meta-analysis was redone and resubmitted to the journal that had previously rejected it, it was promptly accepted - an illustration of potential publication bias.

Song et al (2000) refer a study (DeBeelfeuile et al, 1992) that found that submitted oncology abstracts were more likely to be presented. However, no statistically significant association between study outcome and full publication was observed in seven other studies.

## 4.3 Publication bias in observational studies

The studies above indicate that publication bias definitely exists. How common it is, is more difficult to determine. Studies of publication bias may themselves not be exempt from publication bias. However, when publication bias is found for experimental studies, there is a danger that publication bias is even more common for observational studies. Randomised studies must be prospectively organized. Consequently, the time commitment and expense required will generally be much greater than a historically controlled study, which can be constructed on a more ad hoc basis (Begg and Berlin, 1994). In the extreme case, existing data may be used to test a hypothesis and if the hypothesis is confirmed (ie the null hypothesis is rejected) the result is published. If the hypothesis is not confirmed it does not seem worthwhile to publish (or a submitted paper is rejected).

The more data there is that is easily available, the more likely it is to find ad hoc studies. Since accident data are collected routinely in most countries, it is fairly easy to carry out road safety studies with existing data. There is therefore reason to believe that publication bias is at least as great a problem for observational road safety studies as for randomised controlled trials.

## 4.4 The funnel plot

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. The choice of a precision measure will be discussed further below.

An example of a funnel plot is shown in the figure 4.1.

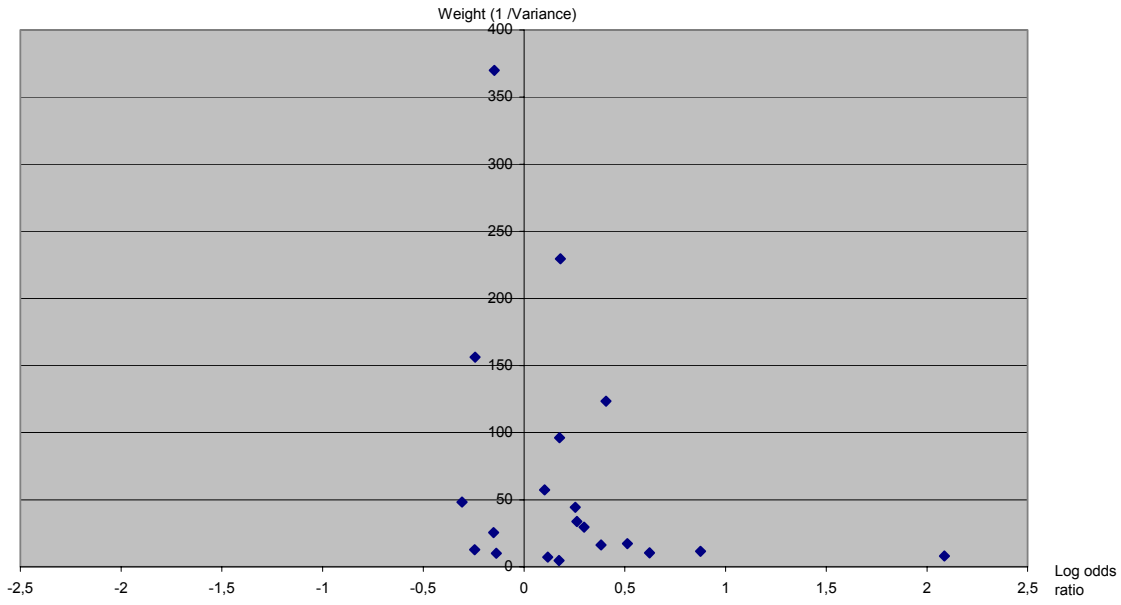



Figure 4.1. Funnel plot. Illustration based on arbitrary data. TØI report 692/2003.

In figure 4.1. 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.

As discussed above, 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.

Many articles introducing the funnel plot explain this as plotting the effect size against the sample size. Others use other precision measures for the dependent variable. The choice of precision measure may however affect the shape of the plot and the conclusion as to whether publication bias is indicated or not. Tang and Liu (2000) used two different measures of precision, the inverse of the standard error and the trial size. They also used two different effect measures, the relative risk and the risk difference. The relative risk is considered the preferable measure (see below) and the results based on the risk difference will not be discussed here. The results presented are calculated from table 2 in Tang and Liu (2000).

Tang and Liu examined 198 meta-analyses. When using the relative risk as effect measure, publication bias was indicated in 31 of these for either trial size or the inverse variance used for precision. Of these 31, in 11 cases publication bias was indicated for trial size only, in 18 cases for the inverse variance only and in only 2 cases was publication bias indicated for both precision measures. If the two alternatives for a measure of precision are regarded as of equal value, results of funnel plots will be fairly arbitrary. However, there are reasons for asserting that the inverse variance is superior to the sample size.

Sterne and Egger (2001) discuss the choices of axes for funnel plots and suggest guidelines. For the vertical axis the possible choices discussed are the variance, the inverse variance, the standard deviation, the inverse standard deviation (which they call precision), the sample size and log sample size.

An argument for using the inverse variance is that this is the weight employed in fixed effects meta-analysis. Sterne and Egger believe that the standard deviation is the preferable choice of the vertical axis. There are two reasons for this. Interpretation of funnel plots is facilitated by inclusion of lines representing the 95% confidence limits around the summary treatment effect. This indicates the expected distribution of studies in the absence of heterogeneity or selection biases. The lines will be straight only when standard error is used for the vertical axis. Also, funnel plots with standard error along the vertical axis correspond to Egger's regression test to detect publication bias and therefore allow the identification of trials that are influential in the regression analysis. When standard error is used the vertical axis must be inverted (standard error 0 at the top).

More important than this recommendation, however, is the rejection of sample sizes and functions of sample size as the dependent variable. The fundamental assumption underlying the use of the funnel plot is that in the absence of bias a plot from trials estimating the same treatment effect will be symmetrical and bear some resemblance to a funnel. This is not necessarily the case for sample size. For example, when the log odds ratio is used as a measure of effect, the precision of the estimate of the effect depends both on the total sample size and the number of participants developing the event of interest. Studies with very different sample sizes may therefore have the same standard

error. Even without publication bias the funnel plot does not have to be symmetrical when sample size is used for the vertical axis. Also, Dickersin (1997) found that the sample size of a study was only weakly associated with the probability of publication whereas a consistent association was found with statistically significant results.

The conclusion is therefore that funnel plots should employ some function of the standard deviation of studies for the vertical axis.

Sterne and Egger (2001) also discuss the choice of effect measure or horizontal axis. They consider the log odds ratio, log risk ratio and the risk difference. They conclude that a measure based on a ratio should be employed, preferably the log odds ratio. One advantage of the odds ratio is that funnel plots have the same shape whether an outcome is defined as the occurrence or non-occurrence of the event of interest. The shape of funnel plot based on risk ratios will differ depending on the definition of outcome as occurrence or non-occurrence.

## 4.5 Statistical methods analogous to the funnel plot for detecting publication bias

The funnel plot can be employed to obtain a visual indication of publication bias but visual inspection is by necessity subjective. For example, Vandenbroucke (1988, referred in Egger et al 1998) suggested that publication bias might explain the association found between passive smoking and lung cancer but a more objective method described below found no evidence of asymmetry in the funnel plot (p=0.80) (Egger et al, 1997).

This section describes three statistical methods analogous to the funnel plot to test for publication bias. 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.

A fourth method (Sugita et al, 1992) will also be briefly discussed but this method is of doubtful value.

### 4.5.1 Begg's rank correlation test

Begg's (Begg, 1994) test is a test for the independence of effect size and the variance (or the standard deviation since there is a one to one correspondence between the standard deviation 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 $x_i = \frac{(x_i - \bar{x})}{\bar{v}_i^{\frac{1}{2}}}$,

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:

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

Begg and Mazumdar (1994) have investigated the properties of the rank correlation test by simulation and conclude that the power is small for the number of studies normally included in meta-analyses. Some results from their simulations are presented in the section 4.6.1.

Begg's rank correlation test has been implemented at the Institute of Transport Economics as part of a program for meta-analysis written in C**.**

### 4.5.2 The regression method of Egger et al

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. This corresponds to a regression analysis of the Galbraith (1988) plot where the inverse of the standard deviation is plotted along the abscissa and the standardized effect is plotted along the ordinate. 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%.

According to Macaskill, Walter and Irwig (2001), the Egger approach is intrinsically biased because: (i) the independent variable is subject to sampling variability; (ii) the standardized treatment effect is correlated with its estimated precision; and (iii) for binary data, the independent regression variable is a biased estimate of the true precision, with larger bias for smaller sample sizes. Whether these flaws are serious enough to invalidate the regression test has to be decided on the basis of practical experience. One possibility

is to study the properties of the regression test studied by simulation. Some results from simulations that have been carried out are presented in section 4.6.

This regression analysis can of course be performed in any statistical package. A small program has nevertheless been written in the programming language C. An advantage is that this program can be combined with programs for Begg', method and the trim and fill method described next to make a package to perform all tests.

Thompson and Sharp (1999) point out that allowing for heterogeneity in the regression (see the chapter on heterogeneity) may make a difference. They give an example where a regression without allowing for heterogeneity indicated publication bias but when heterogeneity was allowed for there was no longer any indication of publication bias. This has to be further looked into.

### 4.5.3 The trim and fill method of Duval and Tweedie

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 may make this clearer.

Start with the funnel plot from the previous chapter where there seems to be missing studies at the left. Without the studies removed by the trim part of the trim and fill method the funnel plot looks as follows:

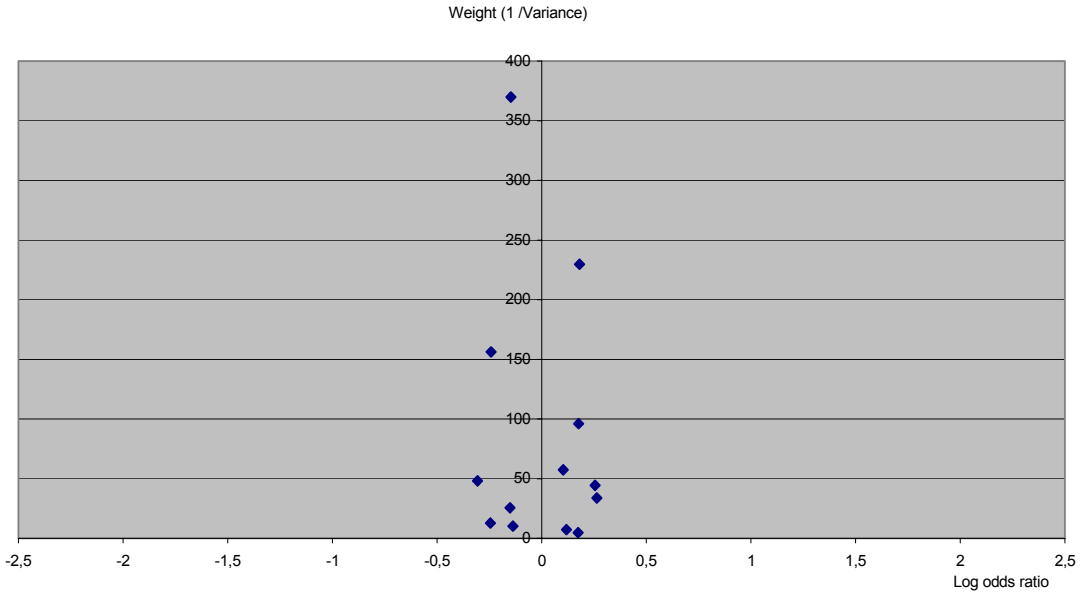



Figure 4.2. Funnel plot for the data in figure 4.1. after trimming the plot with the trim-and fill method. TØI report 692/2003.

These studies are used to calculate the weighted mean and then missing studies are imputed by reflecting the trimmed studies around the mean. All studies, including the imputed studies, are shown in the following funnel plot. The original studies also found in figure 4.1. are marked as diamonds and the imputed studies as triangles.

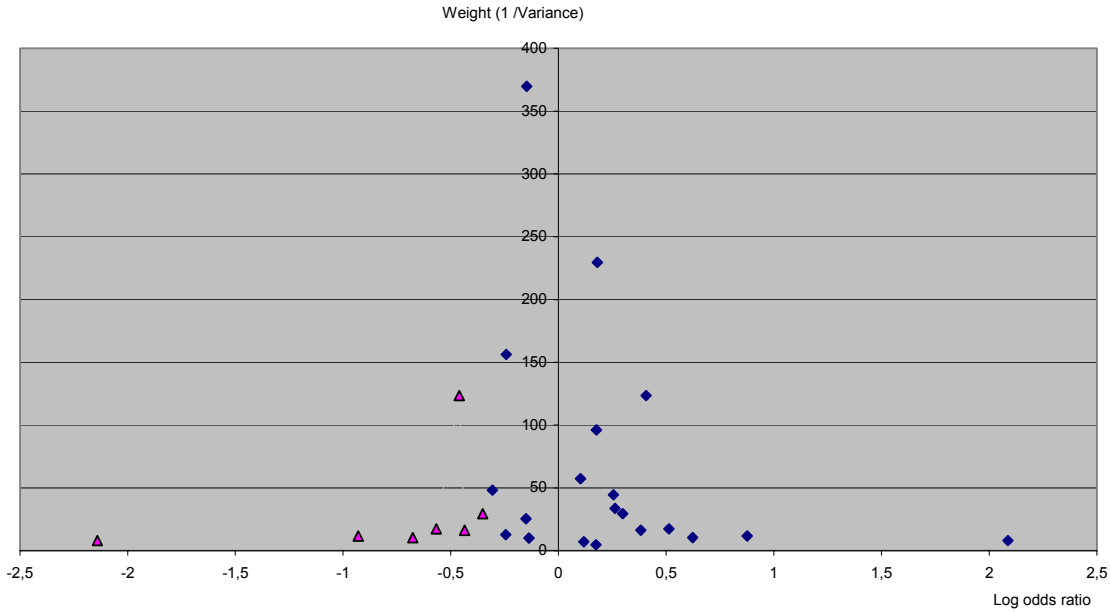



Figure 4.3. Funnel plot for the data in figure 4.1. after employing the trim-and fill method. TØI report 692/2003.

Duval and Tweedie (2000a) develop the trim and fill method in two steps. First they consider a situation where the median of each effect is assumed to be zero, and then they consider a nonzero and unknown effect. The key assumption is that the $k_0$ most negative studies (a positive value indicates effect) have been suppressed. This seems to be the case in the first of the funnel plots above. The number of remaining studies is denoted by n. To estimate the number of suppressed studies, the effects found in the studies are ranked by size and estimators for the number of studies are based on the ranks of the positive effects.

Three different estimators for the number of missing studies are defined. Two of these will be described here. Simulation studies (Duval and Tweedie, 2000b) indicate that these perform better than the third.

One estimator is based on the length of the rightmost number of ranks associated with positive effects. Denoting this length with $\gamma$, the estimator is defined by $R_0=\gamma-1$. The second estimator is based on the sum of ranks for the positive effects. Denoting the ranks by $r_i$, and the effect sizes by $x_i$ the sum of ranks is defined by $T_n = \sum_{x_i>0} r_i$ and an estimator of the number of missing studies is defined by $L_0 = \frac{4T_n - n(n+1)}{2n-1}$.

Duval and Tweedie (2000a) derive the properties of these estimators. They also show by simulations that rounding the estimators to integers hardly affects their properties.

However, the assumption of zero median effect does not necessarily hold in practice and the true median effect is unknown. To handle this, an iterative procedure is developed.

The mean $\Delta$ is estimated by the fixed effects or the RE estimator and centred values are defined by $y_i=x_i-\Delta$. The number of missing studies is then estimated by one of the estimators above and the trimmed observations are used to estimate a new mean. The new mean is used to define new centred values and the number of missing studies is estimated anew. This procedure is repeated until convergence, ie until the number of missing studies does not change.

The properties of these estimators are analytically intractable. Duval and Tweedie (2000b) therefore investigate their properties by simulation. For n+k0 equal to 25, 50, and 75 and values of k0 of 0,5 and 10, where n is the number of included and $k_0$ is the number off missing studies, 1000 sets of funnel plots are generated and the mean estimated number of missing studies calculated. The estimators perform reasonably well. For example, when n=20 and there are 5 missing studies ($k_0$=5) the mean number of missing studies found were 4.6 and 3.7 for the R estimator and L estimator respectively. To be significantly (5% level) different from zero the R estimator must be 4 or more. This is valid for any n. The critical value of the L estimator depends on n. For n=19, L must be 5 or higher, for n=50, L must be 7 or higher.

Sterne (2000) finds by simulation that the trim and fill method has a high false positive rate. Depending on the number of trials the percentage of simulated meta-analyses with "missing" trials varied from 35% to 45%.

The trim and fill method has been implemented at the Institute of Transport Economics as a program written in C. It estimates both the R and L estimators using both fixed effects and RE (when applicable) models. It also calculates the overall mean and variance, both without and with the imputed missing studies.

### 4.5.4 The method of Sugita et al

Sugita et at (1992) assume that the effect sizes will be normally distributed without publication bias. Publication bias leads to a deviation from the normal distribution.

Sugita et al assume that there is one missing study and develop a method to estimate its effect size and weight. This study therefore represents a weighted average of all missing studies. By including the missing study a corrected weighted mean is calculated. When there is no publication bias the missing study will have little effect on the weighted mean.

The effect size and weight of the missing study are estimated by to equations expressing conditions on the third and fourth central moments of the normal distribution. However, these equations, being non-linear, have multiple solutions. Sugita et al do not discuss how to select the appropriate one. The fact that the equations have multiple solutions is not even mentioned.

Experiments with the method lead to seemingly reasonable corrections to the weighted mean but it turned out that the estimated weight of the missing study was negative. This method does not seem reliable and will not be discussed further.

## 4.6 Simulations to investigate the power of methods to detect publication bias

In a letter commenting on the article in BMJ where the Egger regression method was introduced, Irwig et al (1998) asserts that the graphical test, as they call the regression method, is itself biased. Their reasons for considering the test biased have been referred above. They also give an example of the test incorrectly asserting that there is publication bias in simulated data. Egger et al (1998) respond by giving results from 10 000 simulations based on the same assumptions as Irwig et al where it was found that on

average 4.99% of test were significant at the 5% level and 9.63% at the 10% level when there was no bias.

Both before and after this exchange, the properties of tests for publication bias have been investigated by simulation. Begg and Mazumdar (1994) investigated Begg's rank correlation test. Sterne, Gavaghan and Egger (2000), Macaskill, Walter and Irwig (2001) and Schwarzer, Antes and Schumacher (2002) all investigated both the rank correlation and the regression test. The following sections describe these studies in some detail so that the methods and results may be compared. This is followed by a section that discusses and compares the results.

The trim-and-fill methods has also been studied by simulation (Terrin et al , 2003). This study is described in a separate section.

### 4.6.1 Simulations of Begg's test and the Egger regression method

#### Begg and Mazumdar

Begg and Mazumdar (1994) used simulation to investigate the power of Begg's rank correlation test. The effect sizes t in the studies was assumed to be normal, t~N(δ,v), where v is the sampling variance of the study.

##### Scenarios

A literature survey found that for meta-analyses of medical and epidemiological studies, the number of number of studies ranged from 6 to 79 with a mean of 23. For meta-analyses in psychology, social science or education, the mean number of component studies was 73. Two values for the number of studies per meta-analysis were therefore employed, 25 and 75. Only the simulations with the smaller number of studies will be presented here.

The parameter reflecting the effect under study was varied from zero (the null value) through 3.0 standard deviations from the null, where the scale is in standard deviation units for the effect estimator for a study in the "middle" group, i.e., with v = 1. In each simulation, studies were generated in such a way that after selection for publication there were three (approximately) equal-sized groups of studies with different variances, in each case the middle group having a standardized variance of 1. The sizes of the groups were 8, 9 and 8. Two ranges of standardized variances were used: large (v = 0.1, 1.0, 10.0) and small (v = 0.5, 1.0, 2.0).

##### Study selection

The selection model (or censoring mechanism) employed depended on the significance level p given by $p = \Phi(-\frac{t}{v^{\frac{1}{2}}})$ and is given by $s(p) = \exp(-bp^a)$. Two alternatives of this model was used, strong selection bias with a=1.5 and b=4 and moderate selection bias with a=2 and b=4.

An alternative selection model where the chance of publication depended on the effect size was also simulated. The results were similar to the results described below for the selection based on the significance level and will not be described here.

Each simulated meta-analysis was generated in the following way. First, an effect size was randomly generated from a normal distribution N(δ,v) where v is the variance under study. The probability of selection for publication was calculated depending on the configuration under study, i.e., depending on the calculated p-value or the effect size. The decision to include or exclude this study in the meta-analysis was made based on a biased-coin randomisation using the calculated probability of publication. The process

was repeated until the requisite number of studies with variance v was selected. The whole process was repeated 5000 times for each configuration. The estimates of power, ie the percentage of cases that the test was significant at the 5% level (two-sided) have a maximum standard error of 0.7%.

### Results

Tables 4.1 to 4.3 show respectively the power, percentage of studies selected and the bias due to publication bias for different treatment effects. The treatment effect δ is expressed in standard deviation units relative to the variance of the effect size estimate in the average study (viz v = 1).

Table 4.1. The power of Begg's rank correlation test to detect publication bias for various treatment effects, selection strengths and ranges of variances. One-sided selection. Two-sided test. From table 1 in Begg and Mazumdar (1994). Per cent. TØI report 692/2003.

| | Strong selection strength | | Moderate selection strength | |
|---|---|---|---|---|
| **Range of variances** | **Large** | **Small** | **Large** | **Small** |
| **Treatment effect δ** | | | | |
| 0.0 | 60 | 23 | 35 | 13 |
| 0.5 | 54 | 23 | 25 | 12 |
| 1.0 | 40 | 19 | 15 | 9 |
| 1.5 | 29 | 14 | 9 | 7 |
| 2.0 | 21 | 10 | 6 | 5 |
| 2.5 | 13 | 6 | 5 | 4 |
| 3.0 | 9 | 5 | 3 | 4 |

Table 4.2. The percentage of studies selected for various treatment effects, selection strengths and ranges of variances. One-sided selection. From table 1 in Begg and Mazumdar (1994). Per cent. TØI report 692/2003.

| | Strong selection strength | | Moderate selection strength | |
|---|---|---|---|---|
| **Range of variances** | **Large** | **Small** | **Large** | **Small** |
| **Treatment effect δ** | | | | |
| 0.0 | 36 | 37 | 57 | 57 |
| 0.5 | 54 | 52 | 74 | 73 |
| 1.0 | 65 | 67 | 82 | 85 |
| 1.5 | 72 | 79 | 87 | 92 |
| 2.0 | 78 | 88 | 90 | 96 |
| 2.5 | 82 | 93 | 92 | 98 |
| 3.0 | 86 | 96 | 94 | 99 |

Table 4.3. The bias (difference between the true treatment effect and the estimated) for various treatment effects, selection strengths and ranges of variances. One-sided selection. From table 1 in Begg and Mazumdar (1994). TØI report 692/2003.

| | Strong selection strength | | Moderate selection strength | |
|---|---|---|---|---|
| **Range of variances** | **Large** | **Small** | **Large** | **Small** |
| **Treatment effect δ** | | | | |
| 0.0 | 0.35 | 0.75 | 0.25 | 0.54 |
| 0.5 | 0.16 | 0.54 | 0.09 | 0.35 |
| 1.0 | 0.07 | 0.37 | 0.04 | 0.20 |
| 1.5 | 0.05 | 0.23 | 0.03 | 0.11 |
| 2.0 | 0.03 | 0.13 | 0.02 | 0.05 |
| 2.5 | 0.02 | 0.07 | 0.01 | 0.03 |
| 3.0 | 0.02 | 0.04 | 0.01 | 0.02 |

A discussion of the results will be found later in this chapter.

## Sterne, Gavaghan and Egger

Sterne, Gavaghan and Egger (2000) study the properties of the Egger regression method and Begg's rank correlation method for detecting publication bias. Their method for generating bias is very different from the way bias has been generated in the other studies. The extent of bias was defined by assuming a linear relationship between treatment effect (log odds ratio) and its standard error (SE):

log(OR) = true treatment effect + (bias coefficient x S.E of log OR).

The Egger regression equation can be written t=as+b. This has the same form as the equation for generating bias, with the bias coefficient a and true treatment effect b. It may seem, therefore, that the method of generating bias favours the Egger method.

The values of the bias coefficient employed were 0, -0,5 and -1.0, defined as no bias, moderate bias and severe bias. For simulations with non-zero bias, an iterative procedure was used to calculate the actual treatment odds ratio, and hence the treatment group event rate, in each trial. An iterative procedure was necessary because the odds ratio depends on the standard error, which in turn depends on the odds ratio.

The choice of control group event rates was based on event rates found in 78 published meta-analyses, where the median event rate ranged from 8.9% to 18.1%. Control group event rates of 5%, 10% and 20% were used to include this range.

The number of trials in the simulated meta-analyses and the sample sizes of the trials were also based on the 78 published meta-analyses. Four hypothetical meta-analyses containing 5, 10, 20 and 30 trials were defined. The sample sizes of the trials included in these are shown in table 4.4.

Table 4.4. Sample size of trials included in four hypothetical meta-analyses which were used in simulations. From table 2 in Sterne, Gavaghan and Egger (2000). TØI report 692/2003.

| Sample size | Meta-analyses based on 5 trials (No. of trials) | Meta-analyses based on 10 trials (No. of trials) | Meta-analyses based on 20 trials (No. of trials) | Meta-analyses based on 30 trials (No. of trials) |
|---|---|---|---|---|
| **24-49** | 1 | 1 | 2 | 3 |
| **50-74** | 1 | 1 | 3 | 3 |
| **75-99** | 0 | 1 | 3 | 3 |
| **100-149** | 1 | 2 | 3 | 5 |
| **150-249** | 1 | 1 | 3 | 4 |
| **250-499** | 0 | 2 | 3 | 6 |
| **500-999** | 1 | 1 | 2 | 3 |
| **1000-4999** | 0 | 1 | 1 | 3 |

The power of the methods to detect bias was examined by simulating 10,000 meta-analyses for each "typical" meta-analysis, degree of bias and control group event rate (1,080,000 simulations). These simulations assumed odds ratios of 1 (no true treatment effect), 0.5 and 0.25 (protective treatment effects). In each simulation the number of events in each group in each trial was generated randomly using the binomial distribution. The power (the proportion of simulations which gave evidence for bias) was calculated for each method, in each set of simulations. The results are given in tables 4.5 to 4.7.

In most cases the results were fairly similar for the different values of the control group event rate. Only the results for the event rate of 10% are therefore shown here. A few cases where the results differ are shown in separate tables.

Table 4.5. Power (per cent of simulations with P<0.1) of two tests for small study effects when the true treatment OR=1, for different amounts of simulated bias and control group event rate of 10%. From table 3 in Sterne, Gavaghan and Egger (2000). TØI report 692/2003.

| Bias | | 5 Trials | 10 Trials | 20 Trials | 30 Trials |
|---|---|---|---|---|---|
| **None** | Weighted regression | 10.9 | 10.3 | 10.8 | 10.7 |

| | | | | | |
|---|---|---|---|---|---|
| | Rank correlation | 7.2 | 8.5 | 7.5 | 7.0 |
| **Moderate** | Weighted regression | 13.8 | 23.6 | 38.5 | 49.7 |
| | Rank correlation | 7.4 | 14.3 | 17.5 | 24.5 |
| **Severe** | Weighted regression | 22.8 | 57.5 | 90.1 | 96.4 |
| | Rank correlation | 8.9 | 23.7 | 36.5 | 57.0 |

Table 4.6. Power (per cent of simulations with P<0.1) of two tests for small study effects when the true treatment OR=0.5, for different amounts of simulated bias and control group event rate of 10%. From table 4 in Sterne, Gavaghan and Egger (2000). TØI report 692/2003.

| Bias | | 5 Trials | 10 Trials | 20 Trials | 30 Trials |
|---|---|---|---|---|---|
| **None** | Weighted regression | 10.7 | 10.4 | 10.0 | 10.1 |
| | Rank correlation | 5.9 | 7.1 | 6.5 | 7.2 |
| **Moderate** | Weighted regression | 14.3 | 23.9 | 41.5 | 55.0 |
| | Rank correlation | 4.9 | 9.9 | 14.6 | 7.8 |
| **Severe** | Weighted regression | 25.4 | 68.0 | 93.7 | 98.5 |
| | Rank correlation | 3.9 | 16.1 | 13.7 | 30.6 |

Table 4.7. Power (per cent of simulations with P<0.1) of two tests for small study effects when the true treatment OR=0.25, for different amounts of simulated bias and control group event rate of 10%. From table 5 in Sterne, Gavaghan and Egger (2000). TØI report 692/2003.

| Bias | | 5 Trials | 10 Trials | 20 Trials | 30 Trials |
|---|---|---|---|---|---|
| **None** | Weighted regression | 9.3 | 8.7 | 8.5 | 9.2 |
| | Rank correlation | 4.2 | 3.9 | 3.3 | 3.6 |
| **Moderate** | Weighted regression | 11.7 | 24.2 | 38.6 | 51.5 |
| | Rank correlation | 3.5 | 3.3 | 2.2 | 2.6 |
| **Severe** | Weighted regression | 12.7 | 57.7 | 84.0 | 96.6 |
| | Rank correlation | 6.3 | 2.6 | 13.9 | 5.4 |

The power when there is no bias is the true significance level for the test. It is seen that for the regression method the level is fairly near the correct level in all cases. The highest value found by Sterne, Gavaghan and Egger (2000) was 12.9 % for 5 trials and a control group event rate of 5% (not shown in the tables above).

The true significance level of the rank correlation method is seen to be consistently lower than the nominal level of 10%, in some cases less than half.

When the true treatment effect is extremely large and the event rate is low the true significance level of the tests increases drastically. Results of simulations with an odds ratio of 0.1 and no bias are shown in table 4.8.

The real significance level is substantially larger than the nominal for both tests, but only for the lowest event rate.

Table 4.8. True significance level (per cent of simulations with P<0.1 when there is no bias) of two tests for small study effects when the true treatment OR=0.1, for different control group event rates. From table 6 in Sterne, Gavaghan and Egger (2000). TØI report 692/2003.

| Event rate | | 5 Trials | 10 Trials | 20 Trials | 30 Trials |
|---|---|---|---|---|---|
| **5%** | Weighted regression | 14.4 | 14.8 | 33.5 | 27.1 |

| | | | | | |
|---|---|---|---|---|---|
| | Rank correlation | 14.2 | 24.3 | 38.4 | 48.2 |
| **10%.** | Weighted regression | 6.6 | 6.6 | 12.0 | 8.8 |
| | Rank correlation | 7.8 | 6.3 | 8.6 | 6.9 |
| **20%** | Weighted regression | 7.1 | 7.4 | 7.0 | 6.9 |
| | Rank correlation | 3.7 | 2.7 | 1.6 | 1.3 |

One would expect the power of the tests to increase with the number of trials and with the degree of severity of bias. This is found to be the case for the regression method, and consistently so. For the rank correlation method there is tendency, but the results are not as consistent as for the regression test. In table 4.6. power diminishes with the degree of severity for the case of 5 trials. In table 4.7. power is reduced from 13.9% to 5.4% for severe bias. It is also strange that the power is 30.6 for severe bias and true OR=0.5 (table 4.6.) but the power is reduced to 5.4% when OR=0.25 (table 4.8.). With 10 000 simulation behind each value the standard deviation in the percentages is less than 0.5%, so that the differences cannot be explained by random variation.

As indicated previously, in some case the results for different event rates in the control group differed. This was most common in the case of severe bias. These results are shown in tables 4.9 and 4.10.

Table 4.9. Power (per cent of simulations with P<0.1) of two tests for small study effects when the true treatment OR=0.5, for severe simulated bias and different control group event rates. From table 4 in Sterne, Gavaghan and Egger (2000). TØI report 692/2003.

| Event rate | | 5 Trials | 10 Trials | 20 Trials | 30 Trials |
|---|---|---|---|---|---|
| **5%** | Weighted regression | 20. 0 | 58.4 | 87.5 | 96.6 |
| | Rank correlation | 6.3 | 2.7 | 10.8 | 3.9 |
| **10%.** | Weighted regression | 25.4 | 68.0 | 93.7 | 98.5 |
| | Rank correlation | 3.9 | 16.1 | 13.7 | 30.6 |
| **20%** | Weighted regression | 27.3 | 64.7 | 93.4 | 98.3 |
| | Rank correlation | 9.9 | 32.9 | 50.9 | 71.8 |

Table 4.10. Power (per cent of simulations with P<0.1) of two tests for small study effects when the true treatment OR=0.25, for severe simulated bias and different control group event rates. From table 5 in Sterne, Gavaghan and Egger (2000). TØI report 692/2003.

| Event rate | | 5 Trials | 10 Trials | 20 Trials | 30 Trials |
|---|---|---|---|---|---|
| **5%** | Weighted regression | 22.9 | 23.4 | 41.2 | 66.4 |
| | Rank correlation | 35.1 | 28.2 | 82.7 | 89.3 |
| **10%.** | Weighted regression | 12.7 | 57.7 | 84.0 | 96.6 |
| | Rank correlation | 6.3 | 2.6 | 13.9 | 5.4 |
| **20%** | Weighted regression | 24.7 | 70.6 | 95.6 | 99.1 |
| | Rank correlation | 4.4 | 20.7 | 22.9 | 45.1 |

It looks as if there may have been an error in transcribing the results. For the event rate of 5% in table 4.10. (OR=0.25) the power of the rank correlation method is larger than the power of the regression method. For no other combination of true treatment, bias or control group event rate is the same tendency found. In particular, in table 4.9. (OR=0.5) there is a large difference in power in favour of the regression method for the same event rate and bias.

On the whole, the results are very consistent for the regression method. The findings indicate that the level of significance of the test is approximately correct when the treatment effect is not too extreme, and that power is reasonable satisfactory, at least when the number of trials is at least as large as twenty. The large significance level, ie the

large percentage of false positives, is not likely to be a problem in practical work since such a large effect is unlikely.

The results for the rank correlation method are less consistent. The power is substantially lower than the regression method. The significance level is lower than the nominal level and this contributes to the low power. However, it is difficult to explain why the power in same cases is lower when there is bias than when there is no bias.

These simulations indicate that the regression method is preferable to the rank correlation method. On the face of it, the method of introducing bias favours the regression method but whether this explains the difference in power is doubtful.

### Macaskill, Walter and Irwig

Macaskill, Walter and Irwig (2001) used simulation to investigate the power of Begg's rank correlation test (BV) and the Egger regression method to detect publication bias. For the Egger method both an unweighted model (EU) and a model weighted by the inverse of the variance of each study (EW) were used.

An alternative of Begg's rank correlation test where the sample size was used instead of the within-study variance was also simulated. From the earlier discussion of the choice of axis in a funnel diagram, variance is preferable to sample size. The results for this alternative are therefore not described here.

The authors introduced another method as an alternative to Egger's method, which was also investigated. This method entails fitting a regression directly to the data using the treatment effect ($t_i$) as the dependent variable, and study size ($n_i$) as the independent variable (funnel plot regression). The observations are weighted by the inverse variance of the estimate to allow for possible heteroscedasticity (FIV). Since the weights are based on the observed data and therefore stochastic, there may be bias in the slope. An alternative weighted regression with the weights being the reciprocal of the pooled variance for each study was therefore simulated as well (FPV). According to the authors this form of weighing should reduce the correlation between the weight and the dependent variable.

The situation considered is a treated group and a control group with a binary outcome. The measure of effect is the log odds ratio, with a value of less than 0 to indicate a favourable effect.

#### Scenarios

The scenarios were chosen to reflect the number of studies and their sample sizes commonly included in meta-analyses and based on characteristics of 70 meta-analyses published in seven leading medical journals between 1990 and 1995. Each simulated meta-analysis comprised a total of 21 studies. Two configurations of study size were used: (i) configuration A, 11 studies of 100 per treatment group, 6 of 200 per group and 4 of 300 per group (7000 total subjects ) (ii) configuration B, 10 of 100 per group, 5 of 200 per group, 3 of 300 per group, 2 of 500 per group, 1 of 1000 per group (9800 subjects). The underlying treatment effects considered were log-odds ratios of 0.0, -0.405, -0.693 and –1.386, corresponding to odds ratios of 1, 2/3, 1/2 and 1/4. The underlying outcome proportion (event rate) in the control group was taken to have a uniform distribution between 0.1 and 0.5 and chosen by a random number generator. The corresponding proportion in the treated group was calculated from the assumed value of the true odds ratio and the simulated results for each study were then generated using a binomial random number generator.

The type I error rate was defined as the percentage of "significant" test for the scenario of no publication bias and the power as the percentage for the scenarios with publication

bias. Based on 10000 replication the maximum standard error of type I error rate was 0.4 per cent and for the power 0.5 per cent.

### Study selection

Macaskill, Walter and Irwig (2001) used both one- and two-sided censoring and one and two-sided test. Only the one-sided cases, ie one-sided censoring and one-sided test, will be described here.

The censoring mechanism used by Begg and Mazumdar (1994) and described above was employed. The values for a and b was as in the strong bias case of Begg and Mazumdar.

After computing the censoring probability for each study the decision to select a study was based on a simulated biased coin toss. The number of selected studies included in each meta-analysis was fixed at 21 and the specified distribution of sample size was maintained by repeating the above procedure for each chosen sample size until the required number of studies was selected. This means that the number of missing studies will vary.

### Results

Table 4.11. shows the percentage of simulated meta-analyses that incorrectly shows significant (5 per cent level) publication bias when there is no bias , ie the real level of significance for the different tests.

Table 4.11. The percentage of simulated meta-analyses that incorrectly shows significant (5 per cent level) publication bias. The percentages are approximate, as they have been read off figures in the original work. Based on figures 5 and 6 in Macaskill, Walter and Irwig (2001).
TØI report 692/2003.

| True odds ratio | Con-figuration | Per cent censored | Egger method | | Funnel plot regression | | Begg's non-parametric test |
|---|---|---|---|---|---|---|---|
| | | | EU | EW | FIV | FPV | BV |
| **1.0** | A | 0 | 5 | 5 | 5 | 5 | 5 |
| | B | 0 | 6 | 12 | 5 | 5 | 5 |
| **0.67** | A | 0 | 7 | 7 | 5 | 5 | 6 |
| | B | 0 | 7 | 12 | 5 | 5 | 6 |
| **0.5** | A | 0 | 8 | 7 | 4 | 4 | 7 |
| | B | 0 | 6 | 8 | 4 | 4 | 6 |
| **0.25** | A | 0 | 14 | 8 | 3 | 4 | 12 |
| | B | 0 | 8 | 8 | 3 | 4 | 8 |

When there is no effect (the true odds ratio equals one), the significance level is near the nominal for all tests for configuration A, while for configuration B the Egger method with weighting is too sensitive. When the effect increase, ie the log odds ratio becomes more negative the real significance level increases for both Egger method and for Begg's test. The real significance level is slightly decreased for the funnel plot regressions. Both the unweighted Egger method and Begg's test seems to have a sensitivity considerably higher than the nominal significance level when the odds ratio is 0.25, ie when the log odds ratio is -1.386.

Table 4.12. shows the power of the tests for different log odds ratios and configurations.

Table 4.12. The percentage of simulated meta-analyses that shows significant (5 per cent level) publication bias (power of the test). From table II in Macaskill, Walter and Irwig (2001).
TØI report 692/2003.

| True odds ratio | Con-figuration | Per cent censored | Egger method | | Funnel plot regression | | Begg's non-parametric test |
|---|---|---|---|---|---|---|---|
| | | | EU | EW | FIV | FPV | BV |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **1.0** | A | 64.5 | 37.0 | 31.8 | 25.6 | 26.3 | 33.4 |
| | B | 64.5 | 58.2 | 56.8 | 40.2 | 40.8 | 43.3 |
| **0.67** | A | 21.5 | 29.8 | 23.9 | 16.4 | 17.5 | 26.1 |
| | B | 19.5 | 31.8 | 29.7 | 18.3 | 19.5 | 26.4 |
| **0.5** | A | 7.7 | 18.6 | 13.6 | 8.3 | 9.2 | 16.3 |
| | B | 7.0 | 15.0 | 15.4 | 7.8 | 8.6 | 14.0 |
| **0.25** | A | 0.8 | 15.4 | 9.6 | 3.4 | 4.7 | 13.5 |
| | B | 0.7 | 9.2 | 9.2 | 3.1 | 4.7 | 9.7 |

As noted by Begg and Mazumdar (1994), intuitively one should expect that a large variation in sample sizes (variances) will have increased power over a small variation. Their comment pertains to Begg's method, but since the other methods also exploit the difference in effect between small and large studies, the comment should be valid for them as well. Since there is more variation in sample size in configuration B, this configuration should show the largest power. This turned out to be the case when the true log odds ratio is small in absolute value, but not when it is large.

The percentage of meta-analyses that show publication bias diminishes when the odds ratio becomes smaller, ie when the true log odds ratio becomes more negative. However, it is doubtful whether this should be interpreted as a reduction of power. The way the simulation has been carried out, the publication bias nearly disappears when the log odds ratio takes it most negative value (less than one per cent censored).

The Egger method without weighting has the highest power. Macaskill, Walter and Irwig (2001) comment that this method is too liberal when there is no bias, but as seen from table 4.11. this is predominantly a problem when the true log odds ratio has a large negative value. When there is no effect, and this is probably when publication bias is most insidious, the Egger method has approximately the correct significance level.

Macaskill, Walter and Irwig (2001) conclude that based on the results for the type I errors, their funnel plot regression weighted by the inverse of the pooled variance is the preferred approach. However, the power of the Egger method is higher, and as discussed above, the significance level is approximately correct when the log odds ratio does not differ much from zero. In our view, the Egger method is therefore the preferred method.

### Schwarzer, Antes and Schumacher

The aim of the simulation study of Schwarzer, Antes and Schumacher (2002) was to examine the performance of the linear regression and rank correlation test in meta-analyses with binary outcome measures under the null hypothesis of no (publication) bias. Both the odds ratio and relative risk were considered as measures of treatment effect. Simulation design and results were only reported for the odds ratio because the results for the relative risk were similar.

10000 meta-analyses were conducted for each combination of the following factors resulting in a total of 90 different configurations, an odds ratio $\theta$ of 0.50, 0.67, 1.00. 1.50 and 2.00, an average event rate $p_A$ of 0.1 and 0.3 and a number of trials in the meta-analysis of 10, 20 and 50. The average event rate $p_A$ was defined as $\text{logit}(p_A) = (\text{logit}(p_E) - \text{logit}(p_C))/2$ with event rate $p_E$ in the experimental group and $p_C$ in the control group. The event probabilities in both trial groups $p_E$ and $p_C$ were calculated as $\text{logit}(p_E) = \text{logit}(p_A)+\text{logit}(\theta)/2$ and $\text{logit}(p_C) = \text{logit}(p_A)-\text{logit}(\theta)/2$.

This study also looked at the effect of heterogeneity by introducing a between-trial variance $\tau^2$ determined as a percentage of 0, 25 and 50 per cent of the within-trial variance of a trial with sample size 100. When $\tau^2$ was positive the logarithm of $\theta$ was generated according to a normal distribution with mean $\log(\theta)$ and variance $\tau^2$.

The sample size of the individual studies was based on a survey conducted at the German Cochrane Centre. All issues from 1948 to 1998 of eight German medical journals were examined and information from all published primary randomized clinical trials was extracted. A normal distribution was fitted to the log sample sizes of these 1555 trials. Thus a normal distribution with mean 3.798 and variance 1.104 was used to generate log sample sizes. The sample size n was rounded to the next even number to get treatment groups each of size n/2. Of the generated trials only trials with a total sample size of at least 30 patients were considered. Accordingly, the probability of generating a trial with total sample size less than 100 is about 66 per cent and to generate a trial with total sample size larger than 500 is about 1.7 per cent. Finally, the number of events were generated according to a binomial distribution with probability $p_E$ ($p_C$ ) and sample size n/2.

The results of the simulations when there is no heterogeneity are shown in table 4.13. for the average group event rate of 0.1. The odds ratios of 1.5 and 2 are the inverse of 0.67 and 0.5 and therefore do no contribute any novel information. Results are therefore only given for the odds ratios 0.5, 0.67 and 1.0.

Table 4.13. True significance level (per cent of simulations with P<0.1 when there is no bias) of two tests for publication bias for different true treatments. Average group event rate 0.1. Based on figure 6 in Schwarzer, Antes and Schumacher (2002). The figures are read off the graphs in figure 6 and are approximate. TØI report 692/2003.

| Effect (OR) | | 10 Trials | 20 Trials | 50 Trials |
|---|---|---|---|---|
| **1.0** | Weighted regression | 13 | 12 | 12 |
| | Rank correlation | 13 | 12 | 12 |
| **0.67** | Weighted regression | 12 | 14 | 15 |
| | Rank correlation | 12 | 14 | 17 |
| **0.5** | Weighted regression | 15 | 17 | 23 |
| | Rank correlation | 13 | 15 | 30 |

The real significance level is found to be larger than the nominal for both tests. The difference is fairly small when the odds ratio is 1 but increases with a reduction in the odd ratio and is quite large when the odds ratio is 0.5. The increase is most pronounced when the number of trials is large.

The detailed results for an average group event rate of 0.3 are not given in the paper. It is however pointed out that the tests perform better in this case. The results of the simulations when there is heterogeneity are shown in table 4.14. and table 4.15. The standard form of regression test has been employed, not the version that allows for heterogeneity.

Table 4.14. True significance level (per cent of simulations with P<0.1 when there is no bias) of two tests for publication bias for different true treatment effects. Average group event rate 0.1. Heterogeneity 25% (see text above for explanation). Based on figure 6 in Schwarzer, Antes and Schumacher (2002). The figures are read off the graphs in figure 6 and are approximate.
TØI report 692/2003.

| Effect (OR) | | 10 Trials | 20 Trials | 50 Trials |
|---|---|---|---|---|
| **1.0** | Weighted regression | 13 | 12 | 14 |
| | Rank correlation | 11 | 10 | 10 |
| **0.67** | Weighted regression | 13 | 16 | 18 |
| | Rank correlation | 12 | 13 | 17 |
| **0.5** | Weighted regression | 17 | 19 | 28 |
| | Rank correlation | 12 | 14 | 28 |

Table 4.15. True significance level (per cent of simulations with P<0.1 when there is no bias) of two tests for publication bias for different true treatment effects. Average group event rate 0.1. Heterogeneity 50% (see text above for explanation). Based on figure 6 in Schwarzer, Antes and Schumacher (2002). The figures are read off the graphs in figure 6 and are approximate.
TØI report 692/2003.

| Effect (OR) | | 10 Trials | 20 Trials | 50 Trials |
|---|---|---|---|---|
| **1.0** | Weighted regression | 13 | 15 | 16 |
| | Rank correlation | 10 | 9 | 10 |
| **0.67** | Weighted regression | 14 | 16 | 19 |
| | Rank correlation | 11 | 12 | 15 |
| **0.5** | Weighted regression | 16 | 19 | 28 |
| | Rank correlation | 11 | 13 | 26 |

For the regression method, there is a slight tendency for the significance level to increase with heterogeneity. For the rank correlation method there is no such increase, on the contrary, there seem to be a very slight decrease.

### Discussion and comparison of results

The methods of the simulation studies vary. Begg and Mazumdar (1994) assume normally distributed data while the other three studies all simulate a control group and a treatment group with binary data. The effect sizes and trial sizes are therefore not comparable.

Begg and Mazumdar (1994) only consider the power for different level of bias (selection strength) and do not estimate the true significance level without bias. On the other hand, Schwarzer, Antes and Schumacher (2002) only investigate the true significance level without bias and do not consider the power when there is bias.

Both the studies of Sterne, Gavaghan and Egger (2000) and of Macaskill, Walter and Irwig (2001) investigate both the significance level and power. However, the way bias is generated differs. The way bias is generated by Macaskill, Walter and Irwig (2001) leads to few studies being left out when the effect is large, ie when the odds ratio is small. The empirical power is then small, but this is not because the methods do not detect publication bias, but because there is very little publication bias to detect. The way bias is generated by Sterne, Gavaghan and Egger (2000) avoid this effect. There is still large bias even when the odds ratio is small. The power, particularly of the regression method, is then large. However, it is not obvious that this is a realistic and fruitful way of generating bias. Bias generation should reflect the way bias is possibly generated when papers are published (or not published).

The only results that are fruitfully compared between these studies are therefore the true (empirical) significance levels. All studies that consider the significance level include simulations with approximately 20 trials (Macaskill, Walter and Irwig use 21). Both the

studies of Sterne, Gavaghan and Egger (2000) and of Schwarzer, Antes and Schumacher (2002) employ the event rate of 10%. However, the former study defines the event rate in the control group, while the latter defines the average event rate for control and test group. When the odds ratio is one, the definitions are equivalent. Macaskill, Walter and Irwig use a uniformly distributed random event rate (in the control group) between 10% and 50%, ie the mean event rate is 30%. The simulations of Sterne, Gavaghan and Egger (2000) are not very sensitive to the event rate but the difference should still be borne in mind when interpreting the comparison.

Another difference between the studies is the size distribution of trials. The trials tend to be larger in Macaskill, Walter and Irwig (2001) and smaller in Schwarzer, Antes and Schumacher (2002) than in Sterne, Gavaghan and Egger (2000). The distribution of sample sizes is shown in table 4.16.

Table 4.16. Distribution of trial sample sizes in the simulations carried out by Sterne, Gavaghan and Egger (2000), Macaskill, Walter and Irwig (2001) and Schwarzer, Antes and Schumacher (2002). Schwarzer, Antes and Schumacher (2002) use random trial sizes and percentages are indicated. Number of trials. TØI report 692/2003.

| Sample size | Sterne, Gavaghan and Egger (2000) | Macaskill, Walter and Irwig (2001) Configuration A | Macaskill, Walter and Irwig (2001) Configuration B | Schwarzer, Antes and Schumacher (2002) Percentage |
|---|---|---|---|---|
| **24-49** | 2 | 0 | 0 | **66** |
| **50-74** | 3 | 0 | 0 | |
| **75-99** | 3 | 0 | 0 | |
| **100-149** | 3 | 0 | 0 | **32** |
| **150-249** | 3 | 11 | 10 | |
| **250-499** | 3 | 6 | 5 | |
| **500-999** | 2 | 4 | 3 | **2** |
| **1000-4999** | 1 | 0 | 3 | |

Macaskill, Walter and Irwig (2001) employ a 5% significance level while the other two studies employ 10%. Table 4.17. therefore shows the ratio of the real significance level to the nominal significance level.

Table 4.17. The ratio of the real significance level, as determined by simulation, and the nominal level. From Sterne, Gavaghan and Egger (2000), Macaskill, Walter and Irwig (2001) and Schwarzer, Antes and Schumacher (2002). TØI report 692/2003.

| | Egger's regression method | | | | Begg's rank correlation method | | | |
|---|---|---|---|---|---|---|---|---|
| **Odds ratio** | **Sterne et al** | **Macaskil et al Conf A** | **Macaskil et al Conf B** | **Schwarzer et al** | **Sterne et al** | **Macaskil et al Conf A** | **Macaskil et al Conf B** | **Schwarzer et al** |
| | **Reg** | **EU A** | **EU B** | **Reg** | **Rank** | **BV A** | **BV B** | **Rank** |
| **1** | 1.08 | 1 | 1.2 | 1.2 | 0.75 | 1 | 1 | 1.2 |
| **0.67** | na | 1.4 | 1.4 | 1.4 | Na | 1.2 | 1.2 | 1.4 |
| **0.5** | 1 | 1.6 | 1.2 | 1.7 | 0.65 | 1.4 | 1.2 | 1.5 |
| **0.25** | 0.85 | 2.8 | 1.6 | Na | 0.33 | 2.4 | 1.6 | na |

Sterne, Gavaghan and Egger (2000) find lower true levels of significance for both tests than the other studies. They find that the Egger regression test has approximately the correct level while for the Begg's rank correlation test the real significance level is lower than the nominal. The other two studies find that the real significance level tends to be too high for both tests, and is increasing with a reduction in the odds ratio. The real significance level is therefore particularly high when the odds ratio is low. Sterne,

Gavaghan and Egger (2000) find that the significance level *decreases* with a reduced odds ratio.

Given the small standard deviation in the percentages the differences cannot be explained by random variations and it is difficult to come up with any other explanation. It cannot be due to the different ways of introducing bias since this should not affect the simulations without bias. However, while the simulations compared have been selected for being as similar as possible, some differences remains. When the odds ratio differs from one, the event rate as defined in Sterne, Gavaghan and Egger (2000) is different from the event rate defined by Schwarzer, Antes and Schumacher (2002). The event rate in Macaskill, Walter and Irwig (2001) is random and the average is larger than the event rate in the other two studies. The size distribution of trials also differs between the studies. These differences may explain the difference in results.

An explanation of the differences may be found by a new simulation study. This study should be much bigger than the studies referred. It should use a fine grid of values of odds ratio, number of studies, distributions of study size and event rates. It should also employ different ways to generate bias. The simulations should cover the methods and values used in the referred studies. A multivariate analysis can then be carried out on the simulated data to determine how the significance level and the power depend on the variables that are investigated.

### 4.6.2 Simulation of the trim-and-fill method

#### Terrin et al

Terrin et al (2003) tested the performance of the trim-and-fill method by simulation. They used fixed effects models to simulate homogeneous meta-analyses and random effects models to simulate heterogeneous meta-analyses. The random effects models incorporated the effect of baseline risk on between-study variation.

They also did simulations where sample size was determined by power calculations. Sample size was determined by requiring 80 per cent power for the two-sided $\alpha$= 0.05 test of the null hypothesis that the true odds ratio is 1 against the alternative that the true odds ratio is $\theta_i$. As discussed earlier, all the methods for testing for publication bias are based on the small sample effect or the symmetry of the funnel plot in the absence of publication bias. If the sample size is based on power calculations and if the expected effects are realistic, these assumptions are not fulfilled and the methods for showing publication bias are invalid. No simulations are needed to prove that. The results of these simulations are therefore not described here.

However, as discussed earlier, power calculations may not be a problem only for demonstrating publication bias but for meta-analyses in general. Whether it is more than a potential problem should be studied empirically by investigating to what extent sample sizes are determined by power calculations and not by other factors, economic considerations, for example.

In the simulations of homogenous meta-analyses the values used for the true odds ratio $\theta$ were 0.5, 0.8 and 1.0. The values used for the true rate in the control group $\pi_C$ were 0.15 and 0.30.

In the simulations of heterogeneous meta-analyses the model is hierarchical and incorporates the influence of the control rate on the outcome. It assumes the true control rate for the rth study, $\pi_{C_i}$, and the true log-odds ratio $\theta_i$,conditioned on $\pi_{C_i}$, are normally distributed.

$$\pi_{C_i} \sim N(\mu_C, \tau_C^2) \text{ and } \log(\theta_i \mid \pi_{C_i}) \sim N(\mu + \beta(\pi_{C_i} - \mu_C), \tau^2)$$

The parameter $\mu$ is the mean log odds ratio at the average control rate $\mu_C$, $\tau_C^2$ is the variance of the control rates across studies, ie the variance of $\pi_{C_i}$, $\beta$ is the slope of the regression of log $\theta_i$ on $\pi_{C_i}$ and $\tau^2$ is the residual variance from the regression. When $\beta$ is zero, there is no control rate effect.

The values employed for the simulations were for $\mu_C$ and $\tau_C^2$ 0.15 and 0.005 or 0.30 and 0.02, for $\beta$ -2 or 0 and for $\tau^2$ 0.01 or 0.15.

Several combinations of number of studies and sample size distributions of studies were employed. Results were however only shown for two combinations, typical size, 10 studies and 50 to 500 subjects per study, large size, 25 studies and 100 to 10000 subjects per study.

For the trim-and-fill estimates the L estimator with the random effects method was used.

For each configuration of the simulation parameters, the coverage probability was calculated. This is the fraction of meta-analyses for which the 95 per cent confidence interval for the log-odds ratio contains the true mean log-odds ratio.

For homogenous meta-analyses the coverage probability for a 95% confidence interval varied between 0.92 and 0.95, ie not to far from the correct value.

For heterogeneous meta-analyses the coverage probabilities were too small. When residual variance was large, the drop in coverage probability from unadjusted pooling to trim and fill was between 0.05 and 0.09 for typical size meta-analyses and between 0.04 and 0.06 for large meta-analyses. When residual variance was small, the difference between the two methods was smaller (less than 0.04 for typical size meta-analyses, and less than 0.02 for large meta-analyses).

A positive control rate effect seemed to reduce the coverage probability for small studies and increase it for large studies but the effect is small, the largest difference was 0.03.

The smallest coverage probability was 0.86, ie approximately 10% to small.

Terrin et al (2003) conclude that the trim-and-fill methods spuriously adjust the global effect estimate when the studies were heterogeneous if either the variability among studies caused some precisely estimated studies to have effects far from the middle, simply due to chance or if sample size depended on power calculations. Larger random effect variances increased the problem. They infer from this that the funnel plot itself is inappropriate for heterogeneous meta-analyses.

As discussed above, the result for power calculations is no surprise but (hopefully) not relevant in practice. In the other case, underestimating the size of the confidence interval by approximately 10% may be a small price to pay to avoid overestimating the effect because of publication bias. Until it has been shown empirically that the small study effect is due to power calculations, the trim-and-fill method seems to do reasonably well.

## 4.7 Other methods for detecting publication bias

Several methods for detecting publication bias not related to the funnel plot have been developed. The simplest, and probably best known method, is the fail-safe N (Rosenthal, 1979). This method estimates the number of studies with zero effect necessary to make an effect found in a meta-analysis to be no longer significant. This method is very conservative. The necessary number of missing studies is usually fairly large, for example

compared to the number of missing studies indicated by the trim and fill method. The methods described above are more sensitive and the fail-safe method will not be discussed further.

The other methods discussed in this chapter have one thing in common, they employ fairly advanced statistical methods. This makes them difficult to learn and more difficult and time-consuming to implement. The methods have not been implemented and have therefore not been tried out at the Institute of Transport Economics. Lack of software has in general limited their use.

The methods described are just a selection of methods, in particular no Bayesian methods are described. Sutton et al (2000) discuss a number of methods not discussed here.

The methods described are all based on modelling the probability of a study being published.

Iyengar and Greenhouse (1988) assume that publication bias (or other kinds of bias) can be described by a weight function w(x) where x is the observed effect. The probability density of the obtained effect x is given by f(x;θ) where θ is a parameter, for example the expectation. The probability density of *published* effects is then f(x;θ)w(x).

Iyengar and Greenhouse assume that the probability of publication depends on the significance level. It is assumed that the test is two-sided and that the observed effects follow a student distribution with q degrees of freedom. All significant results are published. Two alternative weight functions are employed for non-significant results. The weight function are given by:

$$\mathrm{w}_1(x;\beta,q) = \frac{|x|^{\beta}}{t(q,0.05)^{\beta}} \text{ if } |x| \le t(q,0.05) \text{ and 1 otherwise}$$

$$\text{and } \mathrm{w}_2(x;\gamma,q) = e^{-\gamma} \text{ if } |x| \le t(q,0.05) \text{ and 1 otherwise}$$

respectively.

With the first weight function, the probability of publication increases from zero when x equals zero to 1 when x is reach significance. With the second weight function, the probability of publication is constant as long as x is non-significant.

Iyengar and Greenhouse estimate maximum likelihood estimators for θ and β or θ and γ based on data from Hedges and Olkin (1985). Without correcting for publication bias the estimate for the average effect (the parameter θ) was 0.057. The maximum likelihood estimator for θ was 0.026 and 0.022 respectively when $w_1$ and $w_2$ was employed.

In this case, the corrected value became less than half of the value without bias correction and the result is robust with respect to the choice of weight function.

Hedges (1992) generalises the weight functions of Iyengar and Greenhouse. He refers to empirical studies carried out by psychologists indicating that the level of significance is considered important. A result that is significant at the 4.5% level is regarded as much more interesting than a result significant at the 5.5% level. However, results significant at the 3.5% level and 4.5% level are regarded as equivalent. The weight function employed is therefore a step function with points of discontinuity at the standard levels of significance, 5%, 1%, 0.5% and 0.1%.

The weight function can accordingly be expressed:
$w(p) = \omega_1 \quad if\ 0 < p_i \le a_1$

$w(p) = \omega_j \quad if\ a_{j-1} < p_i \le a_j$

$$w(p) = \omega_k \quad if\ a_{k-1} < p_i \le a_k$$

Only the relative weights can be estimated. One can therefore be chosen. A natural choice is to set $\omega_1$ to 1. This entails that all studies significant at the lowest level are published.

For estimation, it is necessary to express the weight function by the observed effects instead of the significance level p. Hedges assumes that the observations $x_i$ are normally distributed with an expected value $\delta$ (corresponds to Iyengar's and Greenhouse's $\theta$) and standard deviation $\sigma_i^2$. The level of significance $p_i$ depends on both $x_i$ and $\sigma_i$. The weight function may therefore be expressed by $x_i$ and $\sigma_i$. $\sigma_i$ is estimated from the individual studies and is assumed known.

In contrast to Iyengar and Greenhouse, who employed a fixed effects model, Hedges uses a random effects model, ie he assumes that $\delta$ is normally distributed with expectation $\Delta$ and standard deviation $\tau$. This makes $x_i$ normally distributed with expectation $\Delta$ and standard deviation $\sqrt{\tau^2 + \sigma_i^{\ 2}}$.

The probability of the observations $x_i$ can be expressed by the unknown parameters $\Delta$, $\tau$ and $\omega_1$ to $\omega_k$. Maximum likelihood estimation is used for estimating the parameters. The equations derived are solved with various numerical methods.

Hedges also describes a test for publication bias based on a comparison of the observed levels of significance with the expected distribution. To calculate the expected distribution, estimates for $\tau^2$ and $\Delta$ are necessary. These are obtained in a simpler way than using maximum likelihood. The estimates are given by the formulas:

$$\hat{\tau}^2 = \sum_{i=1}^{n} \frac{(X_i - X)^2}{n-1} - \sum_{i=1}^{n} \frac{\sigma_i^{\ 2}}{n} \text{ and}$$

$$\Delta = \frac{\sum_{i=1}^{n} X_i / (\hat{\sigma}^2 + \tau^2)}{\sum_{i=1}^{n} 1/(\hat{\sigma}^2 + \tau^2)}$$

A standard chi-square test is employed to test the difference between the observed and expected distribution.

Hedges employs his method on a meta-analysis with as many as 755 studies. With such a large number of studies, detecting publication bias ought to be substantially more feasible than for the typical meta-analysis. However, Hedges does not find any significant difference between the expected and actual distribution of levels of significance. The estimate for the mean effect is also nearly the same, whether it is estimated with or without an assumption of publication bias. If no bias is assumed, the estimate of the mean is 0.26. When bias is assumed, the estimate is 0.25.

Probably, there is no publication bias in this case. It is unlikely that a lack of power of the test explains this result.

Terrin et al (2003) use this method with only one cutpoint at p=0.05 on the simulated data used for the trim-and-fill simulation described above. They find that the coverage probabilities are slightly larger than when the trim-and-fill was employed.

Like Hedges (1992), Dear and Begg (1992) employ a step function that is constant between the significance levels for a weight function. Unlike Hedges, they do not work with predetermined significance levels but let the significance levels be determined by the significance levels of the studies included.

Vevea and Hedges (1995) generalise Hedges (1992) to a general linear model. The result from a study is assumed to be normally distributed with expectation $\delta_i$ and variance $\sigma_i^2$. The expectations $\delta_i$ are themselves assumed to be normally distributed with expectation $\Delta_I$ and variance $\sigma^2$. $\Delta_i$ is a linear function of a number of explanatory variables:

$\Delta_i = \sum \beta_j X_{ij}$ . The weight function is the same as in Hedges (1992) but unlike Hedges where the two-sided test were employed, one-sided test are employed. The values of the regression coefficients $\beta$ and the coefficients of weight functions $\omega$ are estimated simultaneously.

With this method it is feasible to test for publication bias by setting the coefficients of weight functions $\omega$ to one and employ a likelihood ratio test to see whether the fit is significantly reduced. It is also possible for publication bias to depend on confounding variables. It is, for example, possible that there are more stringent requirements for publishing an observational study than for randomised controlled trials.

Copas (1998) models publication bias by introducing an extra variable z which depends on the sample size n by the formula:

$$z = \gamma_0 + \gamma_1 n^{\frac{1}{2}} + \delta$$

The sample size is sum of all elements, both in the experimental and control group. $\gamma_0$ og $\gamma_1$ are coefficients and $\delta$ is an error term having the standard normal distribution.

The effect based on a 2x2 table is measured by the $\phi$-coefficient and its value is assumed to be expressed by:

$$y = \mu + (\tau^2 + n^{-1})^{\frac{1}{2}} \varepsilon$$

$\tau$ is the standard deviation of the variation between studies (random effect) and $\varepsilon$ is a normally distributed error term. $n^{-1}$ is the variance of the $\phi$-coefficient. $\mu$ is the expectation, ie the value in need of estimation.

Publication bias is introduced by letting the probability of an estimated value being published depend on z at the same time as y and z are correlated by the error terms $\delta$ og $\varepsilon$ being correlated with a correlation coefficient $\rho$. An observed value is published if $z > 0$.

It is not possible to estimate the values of the parameters $\gamma_0$ and $\gamma_1$ that affects publication bias. By stipulating values for these, the remaining parameters $\mu$, $\tau$ and $\rho$ can be estimated by maximum likelihood estimation. A sensitivity analysis may be carried out by investigating how different choices for $\gamma_0$ and $\gamma_1$ affect the estimates of the other parameters.

Since this method does not estimate publication bias but has to assume parameter values that entail a certain publication bias, it is of little interest for diagnosing or testing for publication bias. It may, however, be used for evaluating the consequences of different degrees of publication bias.

The choice of the $\phi$-coefficient as a measure of association is unusual. Much more common is to use the odds ratio or the risk ratio.

Gleser and Olkin (1996) describe two methods to estimate the number of unpublished studies. Both models assume that the null hypothesis tested in the studies is true.

One model is based on the assumption that out of a total of k+N studies the k studies with the lowest level of significance are published. A variation of this model assumes that the m studies with the lowest level of significance are published and the k-m other published studies are randomly selected from the N+k-m remaining studies. On the basis of these

assumptions and additional assumptions that the studies are independent and that the test statistic is continuous, an estimate can be derived for N, the number of unpublished studies.

In an alternative model the probability of publication is assumed to be a function of the level of significance, P{The study is published | The level of significance is p}=g(p). To be able to estimate the number of unpublished studies, some knowledge of the function g(p) is necessary. It is, however, not necessary to have complete information on g(p). It is sufficient is to know the integral of g(p) over an interval of p. By counting the number of levels of significance in this interval N can be estimated. For example, the integral is known if all studies with a level of significance below a certain value are assumed to be published.

Practical experience with the method has found that the estimated number of missing (unpublished) studies is often quite large, much larger than what is found by the trim and fill method. Since simulations have found that the trim and fill method give reasonable estimates for the number of missing studies, Gleser's and Olkin's method tends to overestimate the number of missing studies.

## 4.8 Publication bias. Recommendations

The possibility of publication bias should always be considered in a meta-analysis. Tests for publication bias should therefore be carried out.

While some of the simulation results described earlier seem contradictory, a general tendency is that the Egger regression test has larger power than Begg's test. Provisionally the Egger test is therefore recommended. The level of significance should be 0.1 or even higher.

Even if there is publication bias, this will not necessarily affect the overall effect estimate. If the small study assumption is true, the missing studies will be small and have low weight, though the effect will be larger in a random effects model. The trim-and-fill method can be used to estimate the bias due to missing studies. It is therefore recommended to perform a trim and fill and estimate the overall effect anew, including the imputed studies.

## 4.9 References

Bardy, A. Bias in reporting clinical trials. *Br J Clin Pharmachol* 1998;46:147-150.

Begg, C B and M Mazumdar. M.Operating characteristics of a rank correlation test for publication bias. *Biometrics 50, 1088-1101, 1994.*

Begg, C B. Publication bias. In: Cooper, H. and L V Hedges. *Handbook of research synthesis.* Russel Sage Foundation, New York 1994.

Begg, Colin B and Jesse A Berlin. Publication bias: A problem in interpreting medical data. *J. R. Statist Soc. A* (1988).

Campney, T F. *Adjustment for selection : publication bias in quantitative research synthesis.* Doctoral thesis, University of Chicago, 1983.

Chalmers T C. Problems induced by meta-analyses. *Stat Med* 1991; 10: 971-80.

Cook D J, G H Guyatt, G Ryan, J Clifton, L Buckingham, A Willan, W McIlroy, AD Oxman AD. Should unpublished data be included in meta-analyses? *JAMA* 1993;269:2749-53.

Copas, J. What works?: Selectivity models and meta-analysis. *J. R. Statist. Soc A* (1999) 162, Part 1, pp 95-109.

Coursol, Allan and Edwin E Wagner. Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias. *Professional Psychology Research and Practice* 1986, Vol 17, No 2 136-137.

Dear, K B G and C B Begg. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical science,* 1992, Vol 7, No 2, 237-245.

DeBellefeuille, C, C A Morrison, and I F Tannock. The fate of abstracts submitted to a cancer meeting: Factors which influence presentation and subsequent publication. *Ann Oncol* 1992;3:187-191.

Dickersin K. How imporant is publication bias? A synthesis of available data. *Aids Educ Prev* 1997;9:15-21.

Dickersin, K, Y Min and C L Meinert. Factors influencing publication of research results. *JAMA* 1992;267:374-378.

Dickersin, Kay and Yuan-I Min. Publication Bias: The Problem That Won't Go Away. In: Kenneth S. Warren and Frederick Mosteller. Doing more good than harm: The evaluation of health care interventions. *Annals of the New York academy of sciences*, volume 703.

Duval, S and R Tweedie. A non-parametric "Trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association,* 2000a.

Duval, S and R Tweedie. Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* (2000b) 56, 276-284.

Easterbrook, P J, J A Berlin, R Gopalan and D R Mathews. Publication bias in clinical research. *Lancet* 1991;337:867-872.

Egger, M, G D Smith, M Scheider and C Minder. *Bias in meta-analysis detected by a simple graphical test*. BMJ 1997;315:629-634.

Egger, M, G D Smith, M Scheider and C Minder. Reply to letters. *British Medical Journal,* 1998; 469.

Galbraith, R (1988). A note on the graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine,* 7:889-894.

Gleser, J and I Olkin. Models for estimating the number of unpublished studies. *Statistics in Medicine,* Vol. 15, 2493-2507 (1996).

Greenwald, A G. Consequences of predjudice against the null hypothesis. *Psychological Bulletin*, 82, 1-20.

Hedges, L V. Modelling publication selection effects in meta-analysis. *Statistical Science,* 1992, Vol 7. No 2, 246-255.

Hedges, L and I Olkin. Statistical Methods for Meta-analysis. *Academic,* Orlando, Fla, 1985.

Iyengar, S and J B Greenhouse. Selection models and the file drawer problem. *Statistical science.*1988, Vol 3, No 1, 109-135.

Irwig, L, P Macaskill, G Berry and P Glasziou. Graphical test is itself bias. *Letter, British Medical Journal,* 1998; 316.

Lipsey, M W and D B Wilson. The efficacy of psychological, educational and behavioral treatment. *American psychologist,* 48, 1992, 1181-1209.

Macaskill, P, S D Walter and L Irwig. A comparison of methods to detect publication bias in meta-analysis. *Statistics in medicine*, 2001; 20:641-654.

Misakian A L and L A Bero. Publication bias and research on passive smoking. Comparison of published and unpublished studies. *JAMA* 1998;280:250-3.

Murtaugh P A. Journal quality, effect size, and publication bias in meta-analysis. *Ecology* 83 (4): 1162-1166 APR 2002.

Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *BMJ* 1992; 305:15-19.

Rosenthal, R. The ”file-drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638-641, 1979.

Schwarzer G, G Antes, M Schumacher. Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Statistics in medicine* , 21(17):2465-2477 Sep l5 2002.

Smith, M L, G V Glass and T I Miller. *The benefits of psychotherapy*. Baltimore, John Hopkins, 1980.

Sohn D. Publications bias and the evaluation of psychotherapy efficacy in reviews of the research literature. *Clin psychol rew* 1996;16:147-56.

Song F, A Easterwood, S Gilbody, L Duley and A J Sutton. Publication and other selection biases in systemtic reviews. *Health Technology Assessmen*t 2000 4(10): 1-115.

Sterling, T D. Publication decisions and their possible effect on inferences drawn from tests of significance - or vica versa. *Am Stat Assic* J 1959;54:30-34.

Sterling TD, W L Rosenbaum and J J Weinkam. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *Am Stat* 1995:49: 108-12.

Sterne, Jonathan A C, David Gavaghan and Matthias Egger. Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* 53 (2000) 1119-1129.

Sterne J A C and M Egger. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal Of Clinical Epidemiology* 54 (2001): 1046-1055.

Sterne, Jonathan A C, Matthias Egger and George Davey Smith. Investigating and dealing wih publication bias and other biases. In: Matthias Egger, George Davey Smith and Douglas G Altman: Systematic reviews in health care. Meta-analysis in context. *BMJ books* 2001.

Sugita, M, M Kanamori, T Izuno and M Miyakawa. Estimating a summarized odds ratio whilst eliminating publication bias in meta-analysis. *Jpn J Clin Oncol* 22(5) 1992.

Sutton, A J, F Song, S M Gilbody and K R Abrams. Modelling publication bias in meta-analysis: A review. Statistical Methods in Medical Research 2000: 9: 421-445.

Tang, J L and J L Y Liu. Misleading funnel plot for detection of bias in meta-analysis. J *Clin Epidemiol* 53 (2000) 477-484.

Terrin, Norma, Christopher H Schmid, Joseph Lau and Ingram Olkin. Adjusting for publication bias in the presence of heterogeneity. *Statist. Med*. 22 (13): 2113-2126 Jul 15 2003.

Thornton, A and P Lee. Publication bias in meta-analysis: its causes and consequences. *J Clin Epidemiol* 53 (2000) 207-216.

Thompson Simon G and Stephen J Sharp. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statist. Med.* 18, 2693-2708 (1999).

Vandenbroucke J P. Passive smoking and lung cancer: a publication bias *British Medical Journal,* 1988;296:391-2.

Vevea, J L and L V Hedges. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, Vol 60, No 3, 419-435.*

# 5 Assessment of quality and quality scores

Whether the quality of studies should somehow be taken into account in meta-analyses, for example in the form of quality scores, has been contentious. Eysenck (1978) considers it an "exercise in mega-silliness" to lump together studies of various quality. Feinstein (1995) also believes quality should be assessed and says that criteria for the scientific quality of the individual studies should be developed. Spector and Thompson (1991) similarly believe that in principle it would seem desirable to down weight studies of "doubtful" quality relative to "good" quality studies, because of their greater likelihood of bias.

On the other hand Greenland (1994) asserts that "quality scoring adds the analyst's subjective bias to the results, wastes information, and can prevent the recognition of key sources of heterogeneity: it should be completely replaced by meta-regression on quality items (the score components)". Greenland uses even stronger words: "Perhaps the most insidious form of subjectivity masquerading as objectivity in meta-analysis is 'quality scoring'"

This chapter argues the case for quality scores. The quality of studies need to be assessed and studies of low quality should count less than better studies in a meta-analysis. Whether Greenland's recommendation of using meta-regression solves the quality problem will be also be discussed.

The case for quality assessment is based on the possible consequences of including low-quality studies in meta-analyses. These consequences will be discussed in section 5.1. Section 5.2 will discuss various methods of incorporating/employing quality assessments in meta-analyses. These two sections leave the definition and measurement of quality open. However, developing principles for quality assessment is no simple matter. This will be discussed in section 5.3.

## 5.1 Why quality assessment should be carried out

The main argument for including all studies regardless of quality, rather than leaving out the methodologically weak ones, is that the reviewer's own biases may influence the assessment of studies as "good" or "poor". This argument can only be countered by showing that it is feasible to develop principles for quality scoring that are reasonably objective. This is left to a later section. In this section it is argued that the consequences of not assessing studies for quality are serious and quality assessment is therefore the lesser evil, even if a perfectly objective scoring system for quality should not be feasible.

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 effects of the low quality of a study may be:

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

Both effects will lead to heterogeneity, ie the differences between the results found in different studies that are larger than what can be explained by the statistical errors in the studies. However, to distinguish between the two effects is important. Glass (1980) (referred in Barley 1988) holds the view that if there is no difference in effect size as a result of quality, studies can be pooled resulting in a larger data base from which to answer questions about substantive characteristics. Glass view seems to be that as long as there is no systematic bias in low quality studies all studies can be considered to be of equal value. Here, this view is considered erroneous. This will be discussed below.

A number of studies have explored the possible biases found in studies of poor quality by studying the relation between a quality indicator of a study and the effect found in the study. Colditz, Miller and Mosteller (1989) found that the likelihood of success of standard psychological therapy was substantially grater for non-random studies than for randomised studies. They mention that similar results is often found in other investigations but that not all investigations had shown this result.

Miller, Colditz and Mosteller (1989) observed that non-randomized studies tended to report larger gains than did the randomized studies but the tendency was not statistically significant at a 5% significance level.

Schultz et al (1995) found that trials with certain indications of inadequate methodological quality tended to report more extreme beneficial effects. They point out that the results potentially could have materialized, in part, because of publication bias if trials of lower methodological quality tended to be published more often if their results were more extreme.

In an empirical study of the relation of quality scores to treatment differences in published meta-analyses of controlled randomised clinical trials (RCT), Emerson et al (1990) found no relation between treatment difference and quality score. Nor did they find any relation between quality score and variation in treatment difference. They conclude that the findings support the view of Glass that meta-analysis should include all studies of a particular question. However, they warn that RCTs generally incorporate the more important features of good research design, whose presence or absence in other designs can be associated with the outcome.

Several authors conclude that there is not necessarily any clear relationship between quality and the results of the study, ie low quality does not necessarily lead to any systematic bias. Dickersin and Berlin (1992) state that: “The intuition that poorer quality studies, or those thought to be most susceptible to “bias” (e.g., case-control studies) tend to show larger effects than better studies, is not always supported by the data.” Khan et al (2001) believes that: “The general consensus seems to be that in specific reviews, biases due to lack of randomisation can distort effects in either direction and that it is impossible to predict whether bias has been avoided in any particular non-randomised study.”

Kunz and Oxman (1993) conclude that failure to use adequately concealed random allocation can distort the apparent effects in either direction, causing the effects to seem either larger or smaller than they really are. The size of these distortions can be as large as or larger than the size of the effects that are to be detected. Verhagen et al (2001) also conclude that components of quality can influence the effect estimates but the direction of this influence is not consistent. Low study quality can both underestimate and overestimate the true effect.

Balk et al (2002) confirm this in an empirical study of the correlation of quality measures with estimates of treatment effects. Twenty-four quality measures were analyzed for 276 RCTs from 26 meta-analyses. The quality measures were dichotomized into high quality

vs low quality. The effect of quality measures was estimated by calculating relative ORs (ROR)s of treatment effect for each measure. Relative ORs 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.

However, the fact that the quality of studies does not always entail a systematic bias does of course not imply that the quality of studies are of no importance and that all studies should be included in a meta-analysis on an equal basis as Glass recommends. Firstly, some studies do find that low-quality studies give biased results compared to better studies. Secondly, even if there is no systematic relation between quality score and the outcome, studies of lower quality may lead to larger variances. Studies of low quality are more uncertain and therefore provide less information. The estimated standard error of the outcome underestimates the real uncertainty of the study. Combining studies of different methodological quality may lead to confidence intervals that are spuriously narrow because the methodological uncertainty of low-quality studies has not been taken into account. On the other hand, including low-quality studies may introduce spurious heterogeneity and lead to wider confidence intervals when random effects models are used.

Emerson et al (1990) did not find any relation between quality score and variation in treatment difference. On the other hand, Kunz and Oxman (1993) conclude that failure to use adequately concealed random allocation can distort the apparent effects in either direction, causing the effects to seem either larger or smaller than they really are. If inadequately concealed random allocation can distort the effects in any direction so can other methodological flaws. In fact, it may be assumed à priori that flawed, or just less than perfect, studies are bound to give results that are possibly distorted. If the distortion can be in either direction, the statistical error of the study underestimates the real variation in possible results. It is also reasonable to assume that the variation will be larger the more flawed a study is.

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 5.1 taken from Bangert-Drowns, Wells-Parker and Chevillard (1997). Low numbers indicate a high quality.

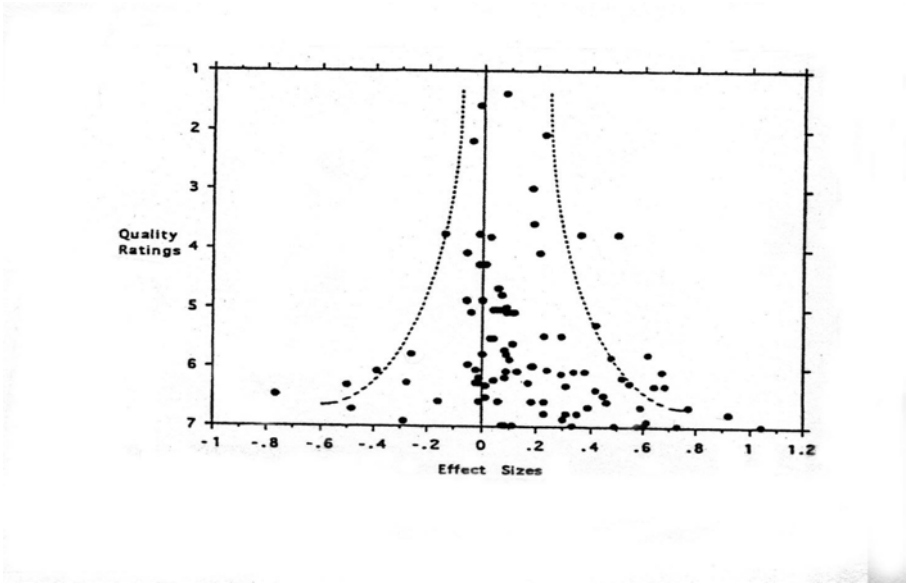


Figure 5.1. Scatter plot of the relation between recidivism effect sizes and ratings of methodological quality for studies of remedial programs for intoxicated drivers.
From Bangert-Drowns, Wells-Parker and Chevillard (1997). TØI report 692/2003.

The funnel shape of the plot can clearly be seen, even if there had been no dotted lines to indicate the funnel.

The larger uncertainty of studies of lower quality supports the case for somehow incorporating the quality of studies in meta-analyses. How this should be done, will be discussed in the next section.

## 5.2 How to incorporate/employ quality

In an interview with Barley (1988) Fredrick Mosteller, an expert on meta-analysis, said that even if he had a good measure of quality, he did not see how to adjust for it in the meta-analysis. Several methods have, however, been suggested. Five will be discussed her, 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). This method is described by the authors in the following way: "First, we construct a set of overall pooled confidence intervals for the estimate of effect……, starting with just the highest quality study and then sequentially adding the next highest quality study that was not included in the previous group".

### 5.2.1 Quality thresholds

This method suffers from two important weaknesses, not to say fundamental flaws. Firstly, any threshold will by necessity be arbitrary. Secondly, 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. To use quality thresholds is therefore not a fruitful method to handle quality differences. This method will not be discussed further.

A variant of quality thresholds is "best evidence synthesis" (Slavin, 1995). Slavin's proposal is that if there are good studies bearing on a problem, poor studies should be disregarded, ie only the best evidence should be taken into account.

### 5.2.2 Stratifying on quality

This method 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.

### 5.2.3 Quality scores as weights

The weights used in meta-analysis for weighting the individual studies are normally based on the variance of the estimated effect found in the studies. As discussed above, this implicitly assumes that there is no other source of uncertainty than the statistical. However, since there is always a methodological uncertainty, because no studies are perfect, it is tempting to treat the methodological uncertainty in the same way as the statistical, by using weights reflecting the methodological uncertainty.

Mathematically this can be done multiplying the statistical weight by a number between 0 and 1 that reflects the quality of the study. This number is the study's quality score. A perfect study, ie one with no methodological weaknesses, would have the value 1. For this study the statistical weight will be the weight applied. For all studies with a quality score less than 1 the weight will be reduced relative to the statistical weight. This way, high quality studies will have a higher weight than low quality studies of the same size.

Many authors are critical to this weighting. 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.

A perceived problem with this is that confidence intervals will be affected by weighting, if the weights are smaller than one, confidence intervals will be larger. To avoid this problem, it is recommended to divide each score by the mean score, to leave the confidence intervals unchanged (Sutton et al, 2000). This view is supported by Detsky et al (1992) who state: "The amount of the widening that automatically results by weighting in logistic regression is completely without empirical support. The amount of widening can easily be modified by multiplying the quality scores by a constant."

This recommendation is misplaced. The fact that studies are of less than perfect quality should be reflected in wider confidence intervals. For a methodologically perfect study there is no other contribution to the uncertainty than the statistical error. The inverse variance is then the correct weight. The real uncertainty of a less than perfect study is larger than the random error and such a study should accordingly have a smaller weight than the inverse variance. How much smaller, should depend on the methodological deficiencies. The extent of methodological deficiencies should be measured by some quality score.

Another problem with by leaving the confidence intervals unchanged after weighting is that this is to give the higher quality studies a larger weight and therefore better precision than their precision (variance) entails.

Using quality scores as weights entails stringent requirements on the scale used. A score of 0.5 means that the effective sample size of the study is halved. A score of 0.25 that is reduced to a fourth. Values of 0.5 and 0.25 should therefore reflect that these studies are only worth a half and a quarter respectively of a perfect study. This requires that scores be measured on the level of a ratio scale. To construct scales to comply with this requirement will require substantial work on scale development.

If different quality scales give different values, the confidence intervals will be different. A comparison of two scales used to score the same studies found the rank order of studies was similar but the value of the score differed (Cho and Bero, 1994). Both scales had a possible range between 0 and 1. Cho's and Bero's scale gave values between 0.36 and 0.74 while the scores obtained by Detsky et al (1987) gave values between 0.04 and 0.54. The means were 0.60 and 0.23 respectively. Confidence intervals based on the scores of Detsky et al would be more than twice as large as confidence intervals based on the Cho's and Bero's scale.

This problem will be further discussed below after the last two methods of using quality scores has been discussed.

### 5.2.4 Meta-regression with either quality or the components of quality as independent variables

Instead of defining a summary quality score it is also possible to define independent components or dimensions of quality. This was Mosteller's suggestion in the interview referred to above (Barley, 1988). He thought that a possibility was to think of a study as having a number of design dimensions that take place in a design space.

This method has the advantage of being able to take into account that different methodological flaws may have different effects. While a composite score may show no systematic bias, the separate components may do so. In that case the effect of the methodological flaws on the results can be estimated by meta-regression with the scores for the methodological dimensions included among the independent variables.

Khan et al (2001) are supporters of using individual quality components. They write: "Scales, in particular, have been criticised for ignoring the direction of bias in their schema. Therefore, in exploring the impact of quality on estimation of effect, it is increasingly considered preferable to use individual components of methodological quality rather than summary scores." And they say further: "The prevalent view is that for exploring the impact of quality, individual quality components are preferable to quality scores."

However, 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. 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.

### 5.2.5 Sequential combination of trial results based on quality score

This method can be regarded as a special way of using quality thresholds and therefore suffers from the same weaknesses as quality threshold. The method generates a sequence of results based on successively less demanding thresholds. Detsky et al (1992), however, do not say what to do with this sequence. If one of the results is chosen, this is equivalent to using a threshold.

### 5.2.6 Discussion of methods to incorporate quality scores

Of the five methods of using quality scores discussed, only two is worthy of further discussion, quality weighting and the use of quality scores or quality components in meta-regression.

Meta-regression would be the preferable method if the effects of methodological flaws were systematic biases. But in so far as the deviations incurred by flaws in the studies can be in any direction, meta-regression will fail to establish any association between methodological flaws and the results of the study. Poor studies will be given the same weights as better studies.

Quality weighting, on the other hand, can handle non-systematic variation caused by methodological flaws by treating this variation in a way analogous to statistical variation. Poor studies will be given less weight than better studies. If the variation is in fact systematic, this cannot be handled by quality weighting.

While some studies have found a systematic effect of methodological flaws in studies, many studies show no systematic effect. It is therefore more important to handle unsystematic variations than systematic. Quality weighting is therefore preferable to meta-regression. Quality weighting is the only method that heeds the dictum that *poor studies should count less than studies of high quality*. Not taking quality into account is implicitly to judge a study by the sample size only, which is plainly silly.

Meta-regression should, however, be used to complement the quality weighting. To show that there is systematic bias is of interest in its own right.

As referred earlier, Glass recommends that when there are no systematic differences between poor and good studies, all studies should be included in a meta-analysis. No weighting is recommended and this means that he believes that all studies should be treated equally, independent of quality. The discussion above indicates that Glass' recommendation does not take into account the larger uncertainty of poor studies and the recommendation should therefore be rejected.

Although weighting of studies is regarded here as the preferred use of quality scores, a weighting exercise will be futile until a proper scale is developed. The problems with quality scales 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. However, the alternative use of quality scores, to reject or select studies for meta-analysis, fares no better. Using the thresholds for definition of high-quality that were defined for the scales, or the median if no threshold was provided by the authors, the effect of one kind of heparin (LMWH) relative to standard heparin was determined separately for high-quality and low-quality studies. For six of the scales there was no significant difference for high-quality studies (as defined by those scales) but low-quality studies found that LMWH was significantly better than the alternative. Seven scales showed the opposite. High-quality studies found a significant effect while low-quality studies found none.

Meta-regression analysis confirmed the differences between scales. Depending on the scale used, the effect size either increased or decreased with increasing trial quality.

None of the 25 scales yielded a statistical significantly association between scores and effect sizes.

Of course, what this boils down to is that *any* use of quality scores requires that the quality scale is valid and reliable. Jüni et al comment that many scales included items that are more closely related to reporting quality, ethical issues or to the interpretation of the results than to the internal validity of trials.

An example of weighting by quality scores is found in Berard and Bravo (1999). Unfortunately their method of weighting seems to be flawed.

Berard and Bravo use quality scores to adjust the weights in a meta-analysis of prevention of bone loss in postmenopausal woman. The quality scores were based on a modified

version of Chalmers et al's (1981) quality assessment scale and measured quality on a scale from 0 to 100.

Berard and Bravo compared both the fixed effects method and random effects method both with and without quality weighting. They found that heterogeneity still remained when quality scores were incorporated in the weights but the overall effect size estimates of the random effects model was associated with a narrower confidence interval.

In our view quality weighting should not lead to narrower confidence intervals. The narrower intervals in their study may be due to their method of incorporating the quality weights into the random effects weights, which seems dubious. Denoting the quality scores q their modified weights are given by $w_i^{'} = qw_i$, where $w_i$ is the fixed effects weight, and similarly $w_i^{'} = qw_i^{*}$ where $w_i^{*}$ is the random effects weight. However, the quality score for a study should modify the contribution to the weight from the study, not the total weight. More correct would therefore be to define the new weights in the random effects case by: $w_i^{'} = \frac{1}{\frac{1}{qw_i} + \tau^2} = \frac{1}{\frac{\sigma_i^2}{q} + \tau^2}$. This assumes that the quality scores have been scaled to a value between 0 and 1.

The next section discusses the construction of quality scales.

## 5.3 The measurement of quality

The previous section discussed the use of quality assessments as if the quality of studies was known. However, the assessment of quality is not without problems.

Some of the results described above indicated that different assessment of quality give different results. In fact, the impossibility of objective quality assessment has been the reason for some people rejecting the use of quality in meta-analyses.

The fact that people disagree about the quality of studies is also shown by the assessment of papers submitted for publication. Slavin (1995) refers studies finding that evaluations of journal articles show considerable variation from reviewer to reviewer. However, peer reviews are bound to be influenced by views on the subject matter, more so than quality evaluations for meta-analysis. For quality assessment for meta-analysis it should be feasible to concentrate on methodological issues.

In this section the view is taken that assessing quality is very demanding but not impossible. It discusses what the concept of quality should encompass and suggest principles for assessment. No system for assessment is presented. Preliminary work with a scoring system has been done (Elvik, 2002) but the system is unsatisfactory and will have to be developed further.

Before discussing the concept of quality and the assessment thereof, it should be pointed out that the goal is to assess studies within the field of road safety, ie quality assessment for social science studies. In social science in general, and also in road safety research, many (probably most) studies are non-experimental or observational. Since the goal of these studies is to approximate experimental studies, following Shadish,Cook and Campbell (2002) they will be referred to as quasi-experimental studies. The number of possible confounders and sources of error is much larger for quasi-experimental studies than for experimental and constructing a quality scale or checklist will be correspondingly more difficult than for experimental studies. Nearly every quality scale constructed pertains to random controlled trials. A medline search from 1990 to January

1997 failed to identify any for the assessment of non-randomised studies (Downs and Black, 1998).

On the other hand, the variation in quality will presumably be much larger for quasi-experimental than for experimental studies for the very same reason that they are difficult to score for quality. Quasi-experimental studies span the range from simple before-and-after studies without any control to advanced designs with several control groups and measurements or data collection at several different times. The best studies therefore control for a much larger number of potential confounders than the worst. Considerations of quality are accordingly more important for meta-analysis of quasi-experimental studies.

### 5.3.1 Definition 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, however, other approaches have also been tried.

Verhagen et al (1998) carried out a Delphi study of items for quality assessment. As an alternative to reach consensus about a definition of quality they considered the possibility of trying to achieve consensus on items that, according to the participants, measure quality of a trial and infer from those a definition, or a description of the concept of quality. The initial list of 209 items categorized by 17 headings generated intense disagreement: on 25% of the items (n = 52) five or more participants scored "strongly agree" to include this item, whereas five or more other participants scored "strongly disagree" to include that item. The disagreement was in part due to different formulations of the items but also to the different priorities of the statisticians and the epidemiologists regarding the inclusion of statistical items.

After three Delphi round a fair consensus was reached (and the large number of items reduced to nine) but the initial disagreement probably reflects that there was no consensus on the definition of quality and it is more fruitful to reach consensus on this before any items are considered.

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.

If there are no methodological weaknesses in the study, the conclusions of the study can be trusted, ie the study is valid. Methodological weaknesses may be of various kinds and therefore it is fruitful to distinguish various kinds of validity. Shadish, Cook and Campbell (2002) discuss four kinds.

External validity reflects the possibility of generalising a study to other situations and other populations. Construct validity indicates the possibility of generalising from operations to constructs, or whether the measurements reflect the theoretical concept intended. Statistical conclusion validity indicates whether there really is a statistical association between variables and internal validity reflects whether the association found is causal.

Of these four validities external validity is not relevant to quality as defined above. Downs and Black (1998) believe that external validity should be included and consider the exclusion of any consideration of external validity in existing instruments to be a limitation. In the Delphi study of Verhagen et al (1998) a consensus was achieved for including three dimensions in the definition of the concept of quality, "internal validity," "external validity," and "statistical considerations."

As pointed out by Jüni, Altman and Egger (2001b): "There is no external validity per se; the term is only meaningful with regard to specified "external" conditions, for example patient populations, treatment regimens, clinical settings, or outcomes not directly examined in the trial. Internal validity is clearly a prerequisite for external validity: the results of a flawed trial are invalid and the question of its external validity becomes redundant". 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.

While the other three kinds of validity are all relevant for quality, internal validity is by far the most important. Statistical conclusion validity boils down to the appropriate use of statistical methods, or whether the assumptions of the statistical tests or estimation methods are fulfilled. Construct validity can be assessed by checking whether the variables measured reflect the theoretical concepts. Internal validity can only be ensured by eliminating other explanations than the cause stipulated. Internal validity reflects the control or lack of control for potentially important confounders. Assessment of quality must therefore take the control of confounders as the starting point.

This definition of quality excludes some factors that have been used in earlier assessment of quality. To further demarcate the concept of quality for meta-analyses it is probably worthwhile to list factors that are *not* relevant to quality as defined here:

*1. How well the results match the results from other studies*

When results from a study is commensurate with other studies this is often taken as enhancing the trust in the study. To use this principle for quality assessment for meta-analysis would be circular reasoning. Each result in the meta-analysis should be regarded as independent. If several studies show similar results the weighting procedure will ensure that the overall result converges to this value.

*2. How well the study is planned*

A study that has been well planned may be of better quality, because it controls better for confounders. But it is the control for confounders that matters, not the planning process as such.

One factor that has been considered as relevant for the quality assessment is whether the sample size for achieving the necessary power has been estimated in advance. This is completely irrelevant. The fact that the sample may be too small to achieve a significant result is of no importance to the meta-analysis. The actual sample size determines the statistical weight of the study so a small study gets less weight. But its quality (internal validity) is no lower.

*3. The thoroughness of the discussion*

The fact the author of a study is aware of the weaknesses of the study and discusses them shows the author has insight but does not remove the weaknesses. Whether the weaknesses of the study are discussed or not is therefore irrelevant for the quality of a study.

*4. How well the report is written*

Moher et al (1995) emphasise that it is important to distinguish between assessing the quality of a trial and the quality of its report. The quality of reporting has no necessary relation with controls for confounders.

Unfortunately, assessing a study for quality depends on the description of the study and a badly written report may mean that information on the control for confounders is missing. Dickersin and Berlin (1992) say that whatever the scoring system, it provides an estimate of the quality of both the report and the study itself in that well-designed studies that are poorly reported may score poorly. However, missing information should not be given the benefit of the doubt, ie if no control for confounders is described, no control should be assumed. Bad reporting will then lead to a lower quality score. Oxman and Guyatt (1991) believe that if what was done is not reported, there is a good chance that it was not done rigorously. They find support for this belief in an empirical study.

Incorporating features discussed above lowers the validity of the scales. Many of the scales constructed suffers from this. Detsky et al (1992) believe that because of its extreme detail, the Chalmers scale may be more susceptible to confusing deficiencies in reporting with deficiencies in conduct and design than are other scales. Jüni et al (1999) state that the scale developed by Jadad et al, which has been widely advocated, gives more weight to the quality of reporting than to actual reported methodological quality . They also give examples of items in scales, in addition to those discussed above, that lowers the validity of the scale, items that are more closely related to reporting quality, ethical issues or the interpretation of results rather than to the internal validity of trials. Some scales assessed whether the rationale for conducting the trial was clearly stated, whether the trialists' conclusion were compatible with the results obtained, or whether the report stated that participants provided written informed consent (Jüni et al,1999).

### 5.3.2 Ways of assessing 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 (RCT) 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. Checklists, therefore, are of no use to meta-analysis and will not be discussed further.

A scale is constructed by giving the components a numerical value and then add (possibly weighted) the values for all components. To create a scale it is not only necessary to select the components to use but also to decide the values or weights given to each component, ie to determine the relative importance of the components. Since different researchers will have different views with regard to the relative importance of components, this process will necessarily be subjective. This may explain some of the differences between the scores for the different scales described in the previous section. However, as pointed out then, some of the differences may be explained by the inclusion in some scales of items that are not relevant to quality as narrowly defined. If agreement can be reached on the demarcation of the concept of quality, this kind of problems will be reduced.

Whether individual markers, checklist or scales are used, it is always necessary to start with the individual markers. An alternative to this is a "holistic" approach, ie a subjective

general score for the quality of a study. Since the validity of a subjective overall score is impossible to judge, this approach is of dubious value. It is also reasonable to believe that the more specific and less complex the rating, the more likely it can be made reliably. Coding the presence or absence of a study feature, for example, should be more reliable than determining subjective, general scores for internal validity. The literature, however, suggests this is not always the case. Bangert-Drowns, Wells-Parker and Chevillard (1997) refer to two studies. Stock et al. (1982) found that changing from a univariate to a summative strategy increased interrater reliability (measured correlationally) only by a small amount and Wells-Parker et al. (1995) found there was greater interrater reliability for an overall quality rating than for several carefully selected anchored scales.

Even though a subjective approach seem to achieve a reasonable reliability, the problems with judging the validity makes this method less appealing than using individual markers. The holistic approach will not be discussed any further.

Evaluating the validity of individual markers or scales based on individual markers is not simple, either. A scale constructed for assessing randomised controlled trials (Chalmers et al, 1981) is one of the best known of this kind. In fact, Sindu, Carpenter and Seers (1997) used this scale a criterion for validating their own tool for rating quality. However, Chalmers has stated in an interview (Barley, 1988) that their scale has never been validated.

There is no criterion to validate a scale or individual markers against. Some other way of validating must therefore be found. The face validity of individual markers can be evaluated but it is more difficult to determine whether the list of individual markers is the best one. Before possible ways of validating a scale are discussed, however, principles for constructing a scale will be discussed.

As discussed above, the quality of a study is taken as mainly equivalent to the internal validity of a study, ie how well the study controls for possible confounders. To measure quality it is necessary to express the degree of control. This can be done in two different ways, either by using the study design or by considering the confounders directly.

It is well known that a randomised experimental study carried out properly controls for every possible confounder. Quasi-experimental studies do not. Various designs, from simple before-and-after designs without a control group to complicated time-series studies with several control groups, vary in their ability to control for confounding factors. 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. A randomised experimental study carried out properly would get the highest possible quality score. If the implementation has flaws, for example if the randomisation procedure is unsuitable, the score would be reduced. At the other and of the scale would be before-and-after studies without a control group.

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. A possible starting point for a scale of this type is the methodological work of Shadish, Cook and Campbell (2002). This book discusses a large number of experimental and quasi-experimental designs and discusses the so-called threats to validity that are controlled for and not controlled for by the various designs. Threats to validity are the same as confounders.

It may also be possible to combine the two methods. Cooper (1989) referred in Bangert-Drowns, Wells-Parker and Chevillard (1997) argued that an integrated approach, coding both threats to validity and methodological features, is an optimal strategy because specific methodological features may not always represent adequately general threats to

validity and vice versa. However, there seems to be a danger of counting the same weaknesses twice. This idea will not be pursued.

Not all threats to validity are relevant in all contexts. Irrelevant threats should be considered as controlled for. This means that all kinds of studies can be scored using a general list of threats to validity. Downs and Black (1998) chose another path. They considered the possible confounders as topic sensitive and provided the raters with information on known confounders. Khan et al (2001), too, believe that research in a particular topic area may be susceptible to specific biases and one should be prepared to modify the selected generic quality instrument. Including appropriate additional items or deleting irrelevant items may be considered.

This kind of tweaking the scale depending on possible confounders will not be feasible if a scale is constructed so that it gives a quality score between 0 and 1 (see below). The score will then depend on the number of possible confounders defined.

It may seem simpler to code designs rather then threats to validity because there are fewer designs than threats to validity. However, by refining the definitions of designs the number of designs can be made quite large. The number of different designs described in Shadish, Cook and Campbell (2002) is at least 30. Also, threats to validity are more basic than designs, designs are created to control for confounders. This is an argument for using threats to validity as a starting point.

Another argument for using threats to validity as basic is that threats can be controlled for not only by design but also by statistical analysis. Rather than trying to fit statistical analyses in with designs when evaluating designs, it is more logical to check whether a threat is controlled for regardless of type of control. If control by design is regarded as superior to control by statistical analysis this can be accommodated by giving extra points for control by design.

For the objective of constructing a scale for observational or quasi-experimental studies old scales are not of much help. Downs and Black, 1998 found that at least 25 checklists have been developed that provide a framework for judging the methodological quality of randomised trials. Jüni, Altman and Egger(2001) identified 14 instruments more. Verhagen et al (2001) estimate the number of scales to between 50 and 60. However, Downs and Black (1998) carried out a medline search from 1990 to January 1997 which failed to identify any for the assessment of analytical non-randomised studies (cohort and case-control studies). The authors tried to develop a checklist appropriate for assessing both randomised and non-randomised studies. Their question regarding confounding was, however, too general to be useful for a variety of quasi-experimental studies. The question was: "Was there adequate adjustment for confounding in the analysis from which the main findings were drawn?" The authors comment that this question had to be customized by providing the raters with information on known confounders. For quasi-experimental studies it would be an over-simplification to talk about adequate adjustment. No study controls for everything possible. It is therefore a matter of degree of control, not of adequate or inadequate control.

It does not seem to be fruitful to have to provide information on known confounders. The scale needs to be based on general principles. It is therefore preferable to work with types of confounders rather than specific confounders. An example may make this clearer. When a measure is implemented at a certain time an effect found may be due to something else happening at the same time (Shadish, Cook and Campbell (2002) calls this *history*). It is impossible to list all the things that may have happened. However, use of a control group will control for all the things that are likely to affect both the experiment group and the control group in the same way. Assessment of the quality should, therefore, be based on control for types of confounders, like *history*, than on specific confounders of this type.

### 5.3.3 The reliability and validity of scales

Like all measuring instruments, a quality scale needs to be reliable and valid. Measuring instruments are *reliable* if the same result is obtained for repeated measurements or measurements by different persons. They are *valid* if they actually measure what they purport to measure.

To evaluate the reliability is simple in principle. One measure of reliability, inter-rater correlation can easily be estimated once the scale is constructed. This does not mean that reliability may nor be a problem. Clark et al (1999) found that the the interrater agreement calculated for the Jadad scale ranged from 0.37 to 0.39 depending on two or four raters, blinded or unblinded. However, theoretically, reliability is no problem. Reliability will, therefore, not be discussed further.

To evaluate the validity is more difficult. Nunally (1967) distinguishes between three main forms of validity, predictive validity, content validity and construct validity. Predictive validity is relevant when a criterion measure exists and the objective is to predict the score on the criterion measure. Content validity is relevant in cases where the purpose is not to predict something else but to directly measure some ability or performance. Nunally uses the example of the final examination for a course in introductory psychology. The validity of the examination depends on how well the questions cover the assigned reading, ie whether they cover the content of the course.

Construct validity is relevant when the objective is to measure a theoretical, abstract variable. Such a variable is called a construct. Since the construct is impossible to measure directly no criterion measure exist and the validity of the measure of the construct has to be evaluated some other way.

Establishing construct validity is done through the research process. A construct will normally be part of a theory, which will lead to predictions involving the construct. If the predictions are confirmed using the measurement of the construct, the belief in the validity of the construct is increased. For example, intelligent people are supposed to do well in their studies. If people who do well on intelligence tests are found to do well, this enhances the validity of intelligence tests. Note that this is not a case of predictive validity. The purpose of intelligence tests is not to predict aptitude for studying but the general construct of intelligence.

Which of these forms of validity are relevant for quality scales? Predictive validity is obviously not relevant since there is no criterion for quality. Both other forms may however be relevant. If quality of a study is *defined* as the extent that the study controls for the threats to internal validity discussed by Shadish, Cook and Campbell (2002), content validity is relevant. The validity of a scale could then be assured by having a checklist where all threats are included. If, however, the quality of a study is taken more generally to be the trustworthiness of the results, construct validity will have to be invoked.

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.

## 5.4 Quality assessment. Recommendations

Jüni, Altman and Egger (2001) recommend that the influence of the quality of included studies should routinely be examined in meta-analyses. They believe that the best way of doing this is using sensitivity analysis. They are also find that the use of quality scores is problematic and believe that it is preferable to examine the influence of individual components of methodological quality.

From the discussion earlier it should be clear that this advice is not agreed with here. Quality should be routinely examined but is necessary to go further than sensitivity analysis. A sensitivity analysis does not lead to any conclusion as to the most "correct" estimate of an effect. The quality of studies should somehow be taken into account when the best estimate of the effect is calculated. Using the individual quality components will be feasible when there is a systematic association between the quality components and effect size. However, this is not necessarily the case, methodological flaws may just increase the uncertainty of the estimated effect.

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. The recommendations of Jüni, Altman and Egger may be valid at present when there is no obvious candidate for a quality scale. The recommendation here is to develop such a scale and use it to weight studies. It will be difficult, demanding and time-consuming. But it is necessary.

## 5.5 References

Balk, E M, P A L Bonis, H Moskowitz, C H Schmid, J P A Ioannidis, C Wang C, J Lau. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287:2973-2982.

Bangert-Drowns, Robert, Elisabeth Wells-Parker and Isabel Chevillard. Assessing the methodological quality of research in narrative reviews and meta- analyses. In: Kendall J. Bryant, Michael Windle and Stephen G. West: The science of prevention: Methodological advances from alcohol and substance abuse research. Washington, D.C.: *American Psychological Association*, c1997.

Barley, Zoe Ann. Assessment of quality of studies for inclusion in meta-analysis. *Doctoral thesis,* University of Colorado, 1988.

Berard, A and G Bravo. Combining studies using effect sizes and quality scores: Application to bone loss in postmenopausal women. *J Clin Epidemiol* 1998;51:801-807.

Chalmers, Thomas C, Harry Smith Jr, Bradley Blackburn, Bernard Silverman, Biruta Schroeder, Dinah Reitman and Alexander Ambroz. A Method for Assessing the Quality of a Randomised Control Trial. *Controlled Clinical Trials* 2. 31-49 (1981).

Cho, Milred K and Lisa A Bero. Instruments of assessing the quality of drug studies published in the medical literature. *JAMA*, July 13, 1994, Vol 272, No. 2.

Clark, H D, G A Wells, C Huet, A Finlay, L McAlister, Rachid Salmi, Dean Ferguson, and Andreas Laupacis. Assessing: the quality of randomised trials: Reliability of the Jadad Scale. *Controlled Clin Trials* 1999;20:448-452.

Colditz, Graham A, James N Miller and Frederick Mosteller. How study design affects outcomes in comparisons of therapy. i: medical. *Statistics in medicine*, vol. 8, 441-454 (1989).

Detsky, Allan S, C David Naylok, Keith o'Rourke. Alllson J Mcgeer and Kristian A L'abbe. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J clin Epidiemol* Vol 45, No 3, pp 255~265, 1992.

Dickersin, K and J Berlin. Meta-analysis: state-of-the-science. *Epidemiol Rev* 1992; 14: 154- 76.

Downs, Sara H and Nick Black. 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,* 1998:52:377-384.

Elvik, Rune. *Measuring study quality; Mission impossible? Safety in numbers: critically assessing results of safety studies.* Paper presented at the TRB annual meeting 2002.

Emerson, John D, Elisabeth Burdick, David C Hoaglin, Frederick Mosteller and Thomas C Chalmers. An Empirical Study of the Possible Relation of Treatment Differences to Quality Scores in Controlled Randomized Clinical Trials. *Controlled Clinical Trials* 11:339-352 (1990).

Eysenck, H J. An exercise in mega-silliness..*American psychologist*, May 1978.

Feinstein, Alvan R. Meta-analysis: statistical alchemy for the 21 st century. *J Clin Epidemiol* Vol. 48, No. 1, pp. 71-79, 1995

Glass, GV. (December 1980). On criticism of our class size/student achievement research: No points conceded. Phi Delta Kappan, 242-244.

Greenland, Sander. Invited commentary: A Critical Look at Some Popular Meta.Analytic Methods. *American Journal of Epidemiology* Vol 140, No, 3, 1994.

Jadad, Alejandro R, R Andrew Moore, Dawn Carroll, R G, Crispin Lenkinson, D John M Reynolds, David J Gavaghan and Henry J McQuay. Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary? *Controlled Clinical Trials* 17: 1-12 (1996).

Jüni, Peter, Douglas G Altman and Matthias Egger. Assessing the quality of controlled clinical trials. *BMJ books,* volume 323 7 july 2001, 2001a.

Jüni, Peter, Douglas G Altman and Matthias Egger. Assessing the quality of randomised controlled trials. In: Mathias Egger, George Davey Smith and Douglas G Altman: Systematic reviews in Health care. Meta-analysis in context. *BMJ books*, Second edition 2001, 2001b.

Jüni, Peter, Anne Witschi, Ralph Bloch and Mathias Egger. The Hazards of Scoring the Quality or clinical Trials for Meta-analysis. *JAMA,* September 15, 1999 Vol 282 , No 11.

Khan, Khalid S, Gerben ter Riet, Jennie Popay, John Nixon and Jos Kleijnen. Phase 5 Study quality assessment. In: Undertaking systematic reviews of research on effectiveness. *CRD's guidance for those carrying out or commisioning reviews,* March 2001.

Kunz, Regina and Andrew D Oxman. The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ books*, volume 317 31 october 1993.

Miller, James N, Graham A Colditz and Frederick Mosteller. How study design affects outcomes in comparisons of therapy. ii: *Surgical statistics in medicine*, vol. 8, 455-466 (1989).

Moher, David, Alejandro R Jadad, Graham Nichol, Marie Penman, Peter Tugwell and Sharon Walsh. Assessing the Quality of Randomized Controlled Trials: An Annotated Bibliography of Scales and Checklists. *Controlled Clinical Trials* 16:62-73, 1995.

Moher, David, Alejandro R Jadad and Peter Tugwell. Assessing the quality of randomized controlled trials. Current issues and future directions. *International Journal of Technology Assessment in Health Care* 12:2 (1996), 195-208.

Nunally, J. (1967). *Psychometric Theory*. New York: McGraw Hill Book Company.

Oxman A D and G H Guyatt. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991; 11: 1271-8.

Schulz, Kenneth F, Ian Chalmers, Richard J Hayes and Douglas G Altman. Empirical Evidence of Bias. Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials. *JAMA*, February 1, 1995-Vol 273, No 5.

Shadish, William R, Thomas D Cook and Donald T Campbell. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company, Boston, New York, 2002.

Sindhu, Lucy Carpenter and Kate Seers. Development of a tool to rate the quality assessment of randomised controlled trials using a Delphi technique. *Journal of Advanced nursing*, 1997, 25, 1262-1262.

Slavin, Robert E. Best evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol*, Vol.48. No. 1. pp.9—18. 1995.

Spector T D and S G Thompson. The potential and limitations of meta- analysis. *J Epidemiol Commun Health* 1991; 45: 89-92.

Sutton A J, K R Abrams, D R Jones, T A Sheldon and F Song. *Methods for meta-analysis in medical research*, Chichester: John Wiley, 2000.

Verhagen, Arianne P, Henrica C W de Vet, Robert A de Bie, Alphons G H Kessels, Maarten Boers, Lex M Bouter and Paul G Knipschild. The Delphi List: A criteria list for quality assessment of randomised controlled trials for conducting systematic reviews developed by Delphi Consensus. *J clin Epidemiol* 1998;51:1235-41.

Verhagen A P, H C W de Vet, R A de Bie, M Boers and P A van den Brandt. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol* 2001;54:651-654.

# 6 Further work

Further work on methodological challenges in meta-analysis at the Institute of Transport Economics will continue along three paths:

1. Implementing new methods and features of methods
2. Investing the properties and limitations of the methods by employing them on real data and simulated data
3. Continuing to study the literature

One of the problems posed at the inception of the Strategic Institute Program was how to treat the case of dependent results, for example when there are several results from the same study. As yet, this has not been studied and is one of the main tasks that lie ahead. Dependency may possibly be handled by hierarchical models.

Another main task is to investigate the statistical methods to diagnose publication bias by simulation. The simulations cited in the chapter on publication bias gave partly contradictory results. A new simulation is therefore worthwhile. This will use a finer grid of values for the parameters that was the case in the earlier simulation studies. A multivariate analysis will then be carried out on the simulated data to determine how the significance level and the power depend on the variables that are investigated.

The correlation between the various methods will also be studied and if possible analysed with a multivariate method.

While most clinical studies and also road safety studies employ the odds ratio as the effect measure, in many contexts relevant at the Institute of Transport Economics the employed measure is a share or a percentage. How to handle percentages as the effect measure in meta-analyses will also be investigated.

Work with constructing a quality scale for observational studies will also continue.