Biostatistics (2011), **12**, 1, *pp.* 122–142
doi:10.1093/biostatistics/kxq046
Advance Access publication on July 22, 2010

Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis

GERTA RÜCKER*

Institute of Medical Biometry and Medical Informatics, University Medical Center, 79104 Freiburg, Germany

ruecker@imbi.uni-freiburg.de

GUIDO SCHWARZER

Institute of Medical Biometry and Medical Informatics, University Medical Center, 79104 Freiburg, Germany

JAMES R. CARPENTER

Medical Statistics Unit, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK

HARALD BINDER

Institute of Medical Biometry and Medical Informatics, University Medical Center, Freiburg, Germany and Freiburg Center for Data Analysis and Modeling, University of Freiburg, 79104 Freiburg, Germany

MARTIN SCHUMACHER

Institute of Medical Biometry and Medical Informatics, University Medical Center, 79104 Freiburg, Germany

SUMMARY

Statistical heterogeneity and small-study effects are 2 major issues affecting the validity of meta-analysis. In this article, we introduce the concept of a limit meta-analysis, which leads to shrunken, empirical Bayes estimates of study effects after allowing for small-study effects. This in turn leads to 3 model-based adjusted pooled treatment-effect estimators and associated confidence intervals. We show how visualizing our estimators using the radial plot indicates how they can be calculated using existing software. The concept of limit meta-analysis also gives rise to a new measure of heterogeneity, termed G^2, for heterogeneity that remains after small-study effects are accounted for. In a simulation study with binary data and small-study effects, we compared our proposed estimators with those currently used together with a recent proposal by Moreno *and others*. Our criteria were bias, mean squared error (MSE), variance, and coverage of 95% confidence intervals. Only the estimators arising from the limit meta-analysis produced approximately unbiased treatment-effect estimates in the presence of small-study effects, while the MSE was acceptably small, provided that the number of studies in the meta-analysis was not less than 10. These

*To whom correspondence should be addressed.

limit meta-analysis estimators were also relatively robust against heterogeneity and one of them had a relatively small coverage error.

Keywords: Empirical Bayes; Heterogeneity; I^2; Meta-analysis; Publication bias; Radial plot; Small-study effects.

1. Introduction

Systematic reviews and meta-analysis are invaluable tools of collating and synthesizing evidence in the life sciences. However, 2 main threats exist to the validity of meta-analysis, heterogeneity, and small-study effects. Heterogeneity may have different sources. One is "clinical heterogeneity" between patients from different studies, measured, for example, in patient baseline characteristics and not necessarily reflected in the outcome measure. There may also be "heterogeneity related to study design" or other study-level characteristics. In this article, we are interested in "statistical heterogeneity," quantified on the effect measurement scale. That is, we look at the extent of treatment-by-study interaction (Senn, 2000). Heterogeneity on this scale essentially measures remaining between-study variation, the clinical implications of which are often context specific.

There is a substantial literature on statistical heterogeneity in meta-analysis, see, for example, DerSimonian and Laird (1986), Hardy and Thompson (1998), Thompson and Sharp (1999), Senn (2000), Engels *and others* (2000), Higgins and Thompson (2002), Sidik and Jonkman (2005), Knapp *and others* (2006), Mittlböck and Heinzl (2006), Jackson (2006), Viechtbauer (2007) and Rücker *and others* (2008b).

"Small-study effects" is a generic term for a phenomenon sometimes observed in meta-analysis that small studies have systematically different (often stronger) treatment effects compared to large ones (Sterne *and others*, 2000). Reasons for this may be publication bias, heterogeneity, selective outcome reporting bias, a mathematical artifact (Schwarzer *and others*, 2002), or genuine random variation (Rothstein *and others*, 2005). There is a vast range of tests for small-study effects, most of them based on funnel plots (Begg and Mazumdar, 1994; Egger *and others*, 1997; Harbord *and others*, 2006; Peters *and others*, 2006; Schwarzer *and others*, 2007; Rücker *and others*, 2008a). Copas and Malley (2008), building on radial plots, developed robust *P*-values adjusting for small-study effects. Stanley (2008) was probably the first who proposed a regression-based treatment-effect estimate adjusting for small-study effects. Moreno *and others* (2009a) systematically evaluated adjusted estimates of a similar type in a comprehensive simulation study, including estimates derived from various linear regression tests and the trim-and-fill method introduced by Duval and Tweedie (2000).

The starting point of the present work is that small-study effects cannot easily be separated from heterogeneity. Rather, they can be seen as a particular case of heterogeneity. Consequently, this article has 2 objectives. First, we develop a new model-based method of calculating adjusted treatment-effect estimates for a meta-analysis potentially affected by heterogeneity, including small-study effects. This is done via a so-called limit meta-analysis, a concept developed in Section 2. Second, we use this concept of limit meta-analysis to introduce a new measure of heterogeneity, called G^2, which measures only systematic heterogeneity that is not accounted for by small-study effects.

The article is organized as follows. In Section 2, we introduce the limit meta-analysis model. In Section 3, we derive the limit meta-analysis using an empirical Bayes argument. In Section 4, maximum likelihood (ML) estimates of the model parameters are derived. In Section 5, we describe how the model can be interpreted using radial plots. In Section 6, we report results of a simulation study, comparing 3 estimates, based on the limit meta-analysis, with established methods. In Section 7, we build on the limit meta-analysis to derive the new measure, G^2. Its properties are explored with real data examples. We conclude with a discussion in Section 8.

2. LIMIT META-ANALYSIS

The concept of limit meta-analysis is based on increasing the precision of a given meta-analysis using a random-effects model that allows for small-study effects. Let k be the number of studies in a meta-analysis, and let x_i be the within-study treatment-effect estimate (e.g. a log-odds ratio), σ_i^2 the (true) within-study variance of x_i (estimated by s_i^2), and $w_i = 1/s_i^2$ the estimated inverse variance (also called precision) used as the weight of study $i\,(i = 1, \ldots, k)$ in the usual fixed-effect model.

We start from the random-effects model in meta-analysis. It assumes that the true treatment effects in the k trials vary around a global mean treatment effect μ_{R} with an underlying between-study variance τ^2, representing heterogeneity:

$$x_i = \mu_{\mathrm{R}} + \sqrt{\sigma_i^2 + \tau^2}\epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1). \tag{2.1}$$

The fixed-effect model is the special case of $\tau^2 = 0$. In the next step, we extend the random-effects model to take account of possible small-study effects by allowing the effect to depend systematically on the standard error:

$$x_i = \beta_{\mathrm{R}} + \sqrt{\sigma_i^2 + \tau^2}(\alpha_{\mathrm{R}} + \epsilon_i), \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1). \tag{2.2}$$

Here, β_R corresponds to the treatment effect (adjusted for small-study effects and therefore usually different from μ_{R}) and α_{R} represents a potential small-study effect. This model is motivated by the test by Egger *and others* (1997) which assumes an additive effect (intercept) α_{R} representing "publication bias." We call it the "extended random-effects model" and will come back to this in Section 5. We now follow an idea used earlier (Rücker *and others*, 2008b) and consider a setting where each study has an M-fold increased precision:

$$x_{M,i} = \beta_{\mathrm{R}} + \sqrt{\sigma_i^2/M + \tau^2}(\alpha_{\mathrm{R}} + \epsilon_i), \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1).$$

Under this setting, all studies are assumed to be more precise ("larger"), but each within-study standard error is still proportional to σ_i. Letting $M \to \infty$, we obtain

$$x_{\infty,i} = \beta_{\mathrm{R}} + \tau(\alpha_{\mathrm{R}} + \epsilon_i), \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1). \tag{2.3}$$

In this model, the random variation within the studies has been removed, but there is still variation between study means. Note that

$$E(x_{\infty,i}) = \beta_{\mathrm{R}} + \tau\alpha_{\mathrm{R}} = \beta_0, \text{ say}, \tag{2.4}$$

and $\mathrm{Var}(x_{\infty,i}) = \tau^2$. We term β_0 the limit meta-analysis expectation.

Based on this model, we construct a "limit meta-analysis" derived from the original meta-analysis. To this end, first write the random errors of the studies in the extended random-effects model (2.2) as

$$\epsilon_i = \frac{x_i - \beta_{\mathrm{R}}}{\sqrt{\sigma_i^2 + \tau^2}} - \alpha_{\mathrm{R}}. \tag{2.5}$$

We assume ϵ_i is fixed for study i and substitute it into (2.3). Then the α_{R} terms cancel out and we get

$$x_{\infty,i} = \beta_{\mathrm{R}} + \sqrt{\frac{\tau^2}{\sigma_i^2 + \tau^2}}(x_i - \beta_{\mathrm{R}}).$$

By plugging in the study-specific standard errors s_i for σ_i and $\hat{\beta}_{\mathrm{R}}$, $\hat{\tau}^2$ from fitting (2.2), we define the "limit meta-analysis" with study-specific treatment-effect estimates

$$y_i = \hat{\beta}_{\mathrm{R}} + \sqrt{\frac{\hat{\tau}^2}{s_i^2 + \hat{\tau}^2}}(x_i - \hat{\beta}_{\mathrm{R}}) \tag{2.6}$$

and standard errors s_i. The y_i are interpreted as new study means, adjusted for small-study effects and shrunken toward a new common mean. In the next paragraph, we use an empirical Bayes argument to interpret the y_i.

3. Relationship between limit meta-analysis and standard empirical Bayes estimates

We apply the idea of the empirical Bayes estimate, equivalently the best linear unbiased predictor, given, for example, in Higgins *and others* (2009), Raudenbush and Bryk (1985), Stijnen and Houwelingen (1990), Greenland and O'Rourke (2001), Verbeke and Molenberghs (2000, page 81), and Rabe-Hesketh and Skrondal (2005, page 22) to the setting of our extended random-effects model (2.2). Let u_i be the trial-specific heterogeneity residual for trial i, including a potential small-study effect. By (2.2), we have

$$u_i = \sqrt{\sigma_i^2 + \tau^2}(\alpha_{\mathrm{R}} + \epsilon_i)$$

with $\mathrm{E}(u_i) = \alpha_{\mathrm{R}}\sqrt{\sigma_i^2 + \tau^2}$ and $\mathrm{Var}(u_i) = \sigma_i^2 + \tau^2$. Thus, we can split up x_i into

$$x_i = \beta_{\mathrm{R}} + u_i = \beta_{\mathrm{R}} + \beta_i + \delta_i,$$

where β_{R} is the treatment-effect parameter, β_i is a random variable following $N(0, \tau^2)$, representing the heterogeneity of the "true" study means as in the usual random-effects model, and δ_i is a "biased" random error variable with expectation $\mathrm{E}(\delta_i) = \alpha_{\mathrm{R}}\sqrt{\sigma_i^2 + \tau^2}$ and variance $\mathrm{Var}(\delta_i) = \sigma_i^2$, with $\mathrm{E}(\delta_i) = 0$ only when $\alpha_{\mathrm{R}} = 0$. The empirical Bayes method estimates the trial-specific heterogeneity residual β_i by its posterior mean $\mathrm{E}(\beta_i)$, given the observed x_i, the prior distribution $\beta_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$, with estimates substituted for τ^2, σ^2, β_{R}, and α_{R}. Using Bayes formula, we obtain for β_i a normal posterior distribution with expectation

$$\widehat{\mathrm{E}}(\beta_i) = \frac{\hat{\tau}^2}{\hat{\tau}^2 + s_i^2}\left(x_i - \hat{\beta}_{\mathrm{R}} - \hat{\alpha}_{\mathrm{R}}\sqrt{s_i^2 + \hat{\tau}^2}\right) \tag{3.1}$$

and variance

$$\widehat{\mathrm{Var}}(\beta_i) = \frac{\hat{\tau}^2 s_i^2}{\hat{\tau}^2 + s_i^2}.$$

Estimating β_i by $\hat{\beta}_i = \widehat{\mathrm{E}}(\beta_i)$, interpreting the square root of the variance as its standard error $\widehat{\mathrm{SE}}_{\hat{\beta}_i}$, and furthermore using (2.4), we see that the limit meta-analysis (2.6) can be written as

$$y_i = \hat{\beta}_{\mathrm{R}} + \hat{\tau}\hat{\alpha}_{\mathrm{R}} + s_i\hat{z}_i = \hat{\beta}_0 + s_i\hat{z}_i,$$

where $\hat{z}_i = \hat{\beta}_i / \widehat{\mathrm{SE}}_{\hat{\beta}_i}$. This justifies our choice of s_i as standard errors of y_i. We can view $\hat{z}_i$ as the "specific z-score" or "specific standardized residual" for trial i. One can think of setting up a new fixed-effects model, where the population mean is $\hat{\beta}_0$ and the (usually unknown) measurement error is now a shrunken estimate of its typical magnitude ($\hat{z}_i$) scaled by the trial-specific standard error s_i. This gives us

y_i, which can be viewed as an estimate of the effect from trial i that is more robust with respect to random as well as systematic error. We can then use the (y_i, s_i), $i \in 1, \ldots, k$ to estimate the treatment effect (fixed or random-effects model), adjusted for small-study effects, and assess and investigate heterogeneity. Compared to standard empirical Bayes estimation, the limit meta-analysis (2.6) has a shrinkage factor of $\sqrt{\hat{\tau}^2/(s_i^2 + \hat{\tau}^2)}$, means less shrinkage than with empirical Bayes $(\hat{\tau}^2/(s_i^2 + \hat{\tau}^2))$ (3.1).

4. MODEL FITTING USING ML

In this section, we simultaneously estimate the parameters of model (2.2), the treatment-effect β_R and the small-study effect α_R, given an estimate of the underlying between-study variance, $\hat{\tau}^2$, representing heterogeneity. As usual in meta-analysis, the observed data are the within-study means x_i and their within-study variance estimates $s_i^2 = \hat{\sigma}_i^2 (i = 1, \ldots, k)$. Using the expectation $\mathrm{E}(x_i) = \beta_R + \alpha_R\sqrt{\sigma_i^2 + \tau^2}$ and variance $\mathrm{Var}(x_i) = \sigma_i^2 + \tau^2$ of x_i and inserting method-of-moment estimates s_i^2 and $\hat{\tau}^2$ for σ_i^2 and τ^2, we obtain the log-likelihood contribution of study i (omitting summands not depending on α_R or β_R)

$$l(\alpha_R, \beta_R | x_i) = -\frac{1}{2(s_i^2 + \hat{\tau}^2)} \left(x_i - \beta_R - \alpha_R\sqrt{s_i^2 + \hat{\tau}^2} \right)^2. \tag{4.1}$$

Writing $w_i = 1/(s_i^2 + \hat{\tau}^2)$, summing up over all studies and setting the partial derivatives to zero yields the estimates

$$\hat{\beta}_R = \frac{\sum_{i=1}^k w_i x_i - \frac{1}{k}\sum_{i=1}^k \sqrt{w_i} \sum_{i=1}^k \sqrt{w_i} x_i}{\sum_{i=1}^k w_i - \frac{1}{k}\left(\sum_{i=1}^k \sqrt{w_i}\right)^2}, \tag{4.2}$$

$$\hat{\alpha}_R = \frac{1}{k}\sum_{i=1}^k \sqrt{w_i}(x_i - \hat{\beta}_R). \tag{4.3}$$

As we will see in Section 5, $\hat{\beta}_R$ (the treatment-effect estimate) and $\hat{\alpha}_R$ (the small-study effect estimate) can be interpreted as slope and intercept in linear regression on so-called generalized radial plots.

Variance estimators are derived from the expected Fisher information (McCullagh and Nelder, 1989, p. 472)

$$I = \begin{pmatrix} k & \sum\sqrt{w_i} \\ \sum\sqrt{w_i} & \sum w_i \end{pmatrix}$$

with inverse

$$I^{-1} = \frac{1}{k\sum w_i - \left(\sum\sqrt{w_i}\right)^2} \begin{pmatrix} \sum w_i & -\sum\sqrt{w_i} \\ -\sum\sqrt{w_i} & k \end{pmatrix},$$

leading to the variance estimates:

$$\widehat{\mathrm{Var}}(\hat{\beta}_R) = \frac{1}{\sum w_i - \frac{1}{k}\left(\sum\sqrt{w_i}\right)^2}, \tag{4.4}$$

$$\widehat{\mathrm{Var}}(\hat{\alpha}_R) = \frac{\frac{1}{k}\sum w_i}{\sum w_i - \frac{1}{k}\left(\sum\sqrt{w_i}\right)^2}. \tag{4.5}$$

Both variance estimates are inversely proportional to the sampling variance of the observed study precisions $1/s_i$. This means that estimation is the more precise, the more study precision varies. By contrast,

when standard errors are similar or even equal for all studies ($s_i = s$), the limit meta-analysis is degenerate since regression of y_i/s on a constant $1/s$ is infeasible, independently of whether heterogeneity or small-study effects may or may not be present. This restriction is a common feature of all funnel plot methods (Higgins and Green, 2009).

5. Model fitting and interpretation using radial plots

In this section, we show how to fit and interpret the extended random-effects model (2.2) in practice. The idea underlying the extended model (2.2) is best motivated by looking at radial and generalized radial plots, see Galbraith (1988), Copas and Malley (2008), and Copas and Lozada-Can (2009).

5.1 *Radial plot*

A radial plot is a scatterplot, using $\sqrt{w_i} = 1/s_i$ as x-coordinates and $\sqrt{w_i}x_i = x_i/s_i$ as y-coordinates. As easily seen (and well known), (i) the slope of the regression line through the origin is the treatment-effect estimate of the fixed-effect model, denoted $\hat{\mu}_{\mathrm{F}}$ and (ii) the sum of squared residuals with respect to this line is Q, which measures the weighted squared deviation of study treatment effects from the overall fixed-effect estimate:

$$Q = \sum_{i=1}^{k} w_i \left(x_i - \frac{\sum w_j x_j}{\sum w_j} \right)^2.$$

Under the null hypothesis of no between-study heterogeneity, Q follows a χ^2 distribution with $k-1$ degrees of freedom (Cochran, 1954). Figure 1 (bottom left panel) shows the radial plot for an example, a meta-analysis on thrombolytic therapy in acute myocardial infarction (Olkin, 1995). The line through the origin is dashed.

Now we look at the best fitting line (not necessarily through the origin, solid line in bottom left panel of Figure 1). The test for small-study effects by Egger *and others* (1997) uses this line, testing the null-hypothesis that its intercept is zero. The intercept $\hat{\alpha}_{\mathrm{F}}$ represents small-study effects. In general, the intercept differs from zero and thus the slope of the line, denoted $\hat{\beta}_{\mathrm{F}}$, differs from the fixed treatment-effect estimate. Following Copas and Malley (2008), we can interpret this slope as a fixed treatment-effect estimate, when allowing for small-study effects. This adjustment often results in smaller estimates of the overall treatment effect (Stanley, 2008; Moreno *and others*, 2009a). Estimates of slope and intercept are identical to the ML estimates (4.2) and (4.3). In analogy to Q, we can define a measure Q' of heterogeneity with respect to the best fitting line:

$$Q' = \sum_{i=1}^{k} w_i \left(x_i - \hat{\beta}_{\mathrm{F}} - \frac{\hat{\alpha}_{\mathrm{F}}}{\sqrt{w_i}} \right)^2, \tag{5.1}$$

where the $w_i = 1/s_i^2$ denote the fixed-effect model weights, as above. We have

$$Q' \leqslant Q$$

since the best fitting line minimizes the residual sum of squares. Again, following Cochran (1954), it can be shown that under the null hypothesis of no between-study heterogeneity and normal assumption, Q' follows a χ^2 distribution with $k-2$ degrees of freedom.

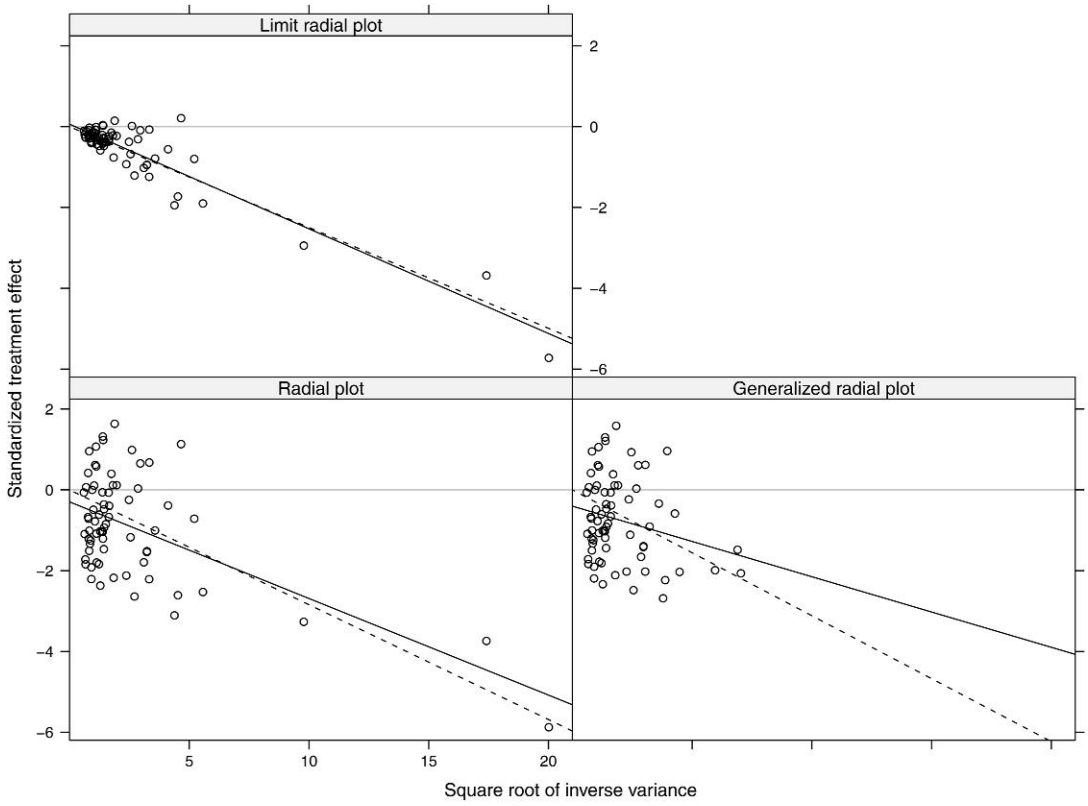

Fig. 1. Thrombolytic therapy data example (Olkin, 1995): radial plot (bottom left panel), generalized radial plot (bottom right panel), and limit radial plot (top panel). Dashed: line through the origin. Solid: best fitting line. Details in text.

5.2 *Generalized radial plot*

Following Copas and Malley (2008), the radial plot can be generalized to incorporate between-study heterogeneity, taking now the values $\sqrt{w_i} = 1/\sqrt{s_i^2 + \hat{\tau}^2}$ as x-coordinates and $\sqrt{w_i}x_i = x_i/\sqrt{s_i^2 + \hat{\tau}^2}$ as y-coordinates (Figure 1, bottom right panel). We do not use different notation for fixed and random-effects weights. This generalized radial plot represents the random-effects model since the slope $\hat{\mu}_{\mathrm{R}}$ of the line through the origin in the generalized radial plot is the random-effects model (2.1) estimate (dashed line). We propose to complement the plot by a regression line that allows for an intercept (bottom right panel of Figure 1, solid line). Its slope $\hat{\beta}_{\mathrm{R}}$ is the treatment-effect estimate of the extended model (2.2), allowing for small-study effects. The intercept corresponds to $\hat{\alpha}_{\mathrm{R}}$, the bias introduced by small-study effects, interpreted as the expected shift in the standardized treatment-effect estimate for a hypothetical "small" study with zero precision.

The generalized radial plot provides an interpretation for $\beta_0 = \beta_{\mathrm{R}} + \tau\alpha_{\mathrm{R}}$, (2.4). This can be interpreted as the expected adjusted treatment effect of a hypothetical study with infinite precision ($s = 0$) and coordinates $1/\tau$ and β_0/τ since inserting $1/\tau$ in the regression equation provides $\beta_0/\tau = \beta_{\mathrm{R}}/\tau + \alpha_{\mathrm{R}}$. This regression problem is essentially equivalent to model (1c) by Moreno *and others* (2009a), see also Stanley (2008).

5.3 *Limit radial plot*

For $M \to \infty$, the radial plots (representing the fixed-effect model) approach a stable limit with x-coordinates $1/s_i$ and y-coordinates y_i/s_i (Figure 1, top panel). This is equivalent to shrinking the points of the funnel plot toward a "new" mean, see Section 3. In our data example, both lines become very similar. One may add a generalized limit radial plot, representing the random-effects model for the limit meta-analysis (not shown). As heterogeneity is much reduced by the shrinkage process, this is mostly very similar to the limit radial plot.

5.4 *Adjusted treatment-effect estimates*

The basis of estimation is either a meta-analysis with raw data (x_i, s_i) or the corresponding limit meta-analysis (y_i, s_i), calculated using (2.6). We use the convention that treatment-effect parameters denoted by the letter β indicate models including an intercept, while parameters denoted by μ indicate models without an intercept. We consider the treatment-effect estimates $\hat{\mu}_{\mathrm{F}}$ (fixed-effect model), $\hat{\mu}_{\mathrm{R}}$ (random-effects model), and $\hat{\mu}_{\mathrm{lim}}$ (limit meta-analysis, fixed-effect model) from a line through the origin, and $\hat{\beta}_{\mathrm{R}}$, $\hat{\beta}_{\mathrm{F}}$, $\hat{\beta}_{\mathrm{lim}}$, and $\hat{\beta}_0$ from a best fitting line for the respective models. Our consideration in terms of radial plots shows that all parameters can be estimated using standard meta-analysis software as follows:

1. Given x_i and s_i for each trial, calculate $\hat{\mu}_{\mathrm{F}}$, $\hat{\mu}_{\mathrm{R}}$, and an estimate for $\hat{\tau}^2$, for example, the method-of-moments estimate (DerSimonian and Laird, 1986).
2. Using the radial plot, determine the slope $\hat{\beta}_{\mathrm{F}}$.
3. Using $\hat{\tau}^2$, construct the generalized radial plot and determine the slope $\hat{\beta}_{\mathrm{R}}$ and the intercept $\hat{\alpha}_{\mathrm{R}}$.
4. Compute $\hat{\beta}_0 = \hat{\beta}_{\mathrm{R}} + \hat{\tau}\hat{\alpha}_{\mathrm{R}}$.
5. Use (2.6) for calculating the limit meta-analysis treatment-effect estimates y_i, construct the limit radial plot and compute the slopes $\hat{\mu}_{\mathrm{lim}}$ of the line through the origin and $\hat{\beta}_{\mathrm{lim}}$, allowing for an intercept.

Suppose a radial plot (that can also be a generalized radial plot) has x-coordinates $\sqrt{w_i}$ and y-coordinates $\sqrt{w_i}x_i$, where the weights w_i correspond to the chosen model. The slope of the line through the origin is

$$\hat{\mu} = \frac{\sum_i^k w_i x_i}{\sum_i^k w_i}.$$

The slope of the best fitting line is given as in (4.2), its intercept by (4.3). Analogous equations hold for $\hat{\mu}_{\mathrm{lim}}$ and $\hat{\beta}_{\mathrm{lim}}$, with x_i replaced with y_i. Standard errors and confidence intervals are calculated in 2 different ways, both based on the study weights w_i. For models without an intercept, the usual meta-analytic approach is used, which takes $1/\sqrt{\sum w_i}$ as standard error of the pooled estimate. For models including an intercept, the standard errors are derived from the variances given in (4.4) and (4.5). The same standard errors are used for $\hat{\mu}_{\mathrm{lim}}$ (version without an intercept) and $\hat{\beta}_0$ and $\hat{\beta}_{\mathrm{lim}}$ (including intercept). Notice that this approach, in line with the usual random-effects model, treats τ^2 as fixed.

6. Simulation study

In this section, we report results of a simulation study. We computed estimates of μ_{F} (usual fixed-effect model), μ_{R} (random-effects model), β_{F} (fixed-effect model, allowing for an intercept), β_{R} (random-effects model, allowing for an intercept), μ_{lim} (fixed-effect model for the limit meta-analysis, no intercept), β_{lim} (fixed-effect model for the limit meta-analysis, allowing for an intercept), and $\beta_0 = \beta_{\mathrm{R}} + \tau\alpha_{\mathrm{R}}$. These estimates were compared to the Mantel–Haenszel and the Peto estimate (Greenland and Robins, 1985;

Yusuf *and others*, 1985). Moreover, we included one of the most successful adjusting methods identified by the recent study of Moreno *and others* (2009a), the so-called Peters method, see also Peters *and others* (2006). Criteria were (i) the absolute bias of the treatment-effect estimate, (ii) the MSE, (iii) the observed variance between treatment-effect estimates of the same scenario, and (iv) the coverage of 95% confidence intervals. In particular, we were interested to know to what extent the methods were able to reduce bias in the treatment-effect estimates in the presence of small-study effects.

6.1 *Design*

We evaluated a number of scenarios, all based on binary response data, with varying number of trials in the meta-analysis (5, 10, 20), control group event probability (0.05, 0.10, 0.20, 0.30), true odds ratio (0.5, 0.667, 0.75, 1), and heterogeneity variance ($\tau^2 = 0, 0.05, 0.10, 0.20$). Small-study effects were simulated on the basis of the Copas selection model, see Copas and Shi (2000a, 2001). This procedure is extensively described in Rücker *and others* (2008a). The selection parameter of the model, ρ^2, was varied from 0 (no selection), 0.36 (low selection), 0.64 (moderate selection), to 1 (strong selection). Trial sizes were drawn from a log-normal distribution that was fitted to the sample sizes of the rosiglitazone meta-analysis (Nissen and Wolski, 2007). The parameters were 6.056 (mean) and 0.69 (variance); quartiles were 244 (25%), 427 (50%), and 747 (75%). Each of 768 scenarios was repeated 1000 times to provide Monte–Carlo estimates of bias and coverage probabilities.

6.2 *Results*

Results for bias and MSE are shown in Figures 2 and 3 for moderate number of studies ($k = 10$, fixed) and substantial heterogeneity ($\tau^2 = 0.1$), which we believe are representative for many real meta-analyses. Results were similar for the "classical" methods (fixed/random-effects model, Mantel–Haenszel method, and Peto method) with respect to all criteria. Of these, the Peto method had both the least absolute bias and the least MSE. Therefore, only the Peto method was chosen to represent the classical methods on the plots. Not unexpectedly, in general, the absolute bias was larger for smaller number of trials, larger heterogeneity, higher selection (small-study effects), and smaller event rates in the control group. The influence of the true odds ratio was weak. All (unadjusted) classical methods produced estimates biased downward (that is, odds ratios more distant from one) if there were small-study effects, and the MSE increased. The fixed-effect model allowing for an intercept (β_F) tended to a small positive bias (that is, an odds ratio nearer to one) and large MSE and variance. For the random-effects model allowing for an intercept (β_R), we often found markedly biased estimates, for which reason it is not shown.

By contrast, the estimates based on the limit meta-analysis, $\hat{\beta}_{lim}$, $\hat{\mu}_\mathrm{lim}$ and $\hat{\beta}_0$ were nearly unbiased if there were small-study effects. $\hat{\beta}_\mathrm{lim}$ had the smallest bias, but a larger variance of estimates between runs of the same scenario than $\hat{\mu}_\mathrm{lim}$ and $\hat{\beta}_0$. Thus, $\hat{\mu}_\mathrm{lim}$ and $\hat{\beta}_0$ had the smallest MSE, together with the Peto method. The Peto method had the smallest variance, but tended to bias if there was selection, particularly for small event proportions. For meta-analyses with a very small number of trials ($k = 5$) or without any selection, however, the new methods were markedly inferior to the classical methods (results not shown). This is no surprise, as it is well known that for small meta-analyses, all funnel plot methods work poorly with respect to both size and power. We thus recommend not to use adjusting methods if the number of trials is less than 10. Moreover, the extended model has an additional parameter which must be estimated. Bias and MSE of the Peters method were small, often ranging between that of the unadjusted methods and those based on the limit meta-analysis.

Coverage was generally poor, if there was heterogeneity, see Figure 4. This is particularly true for large heterogeneity ($\tau^2 = 0.2$, results not shown) or strong selection. Naturally, the random-effects model

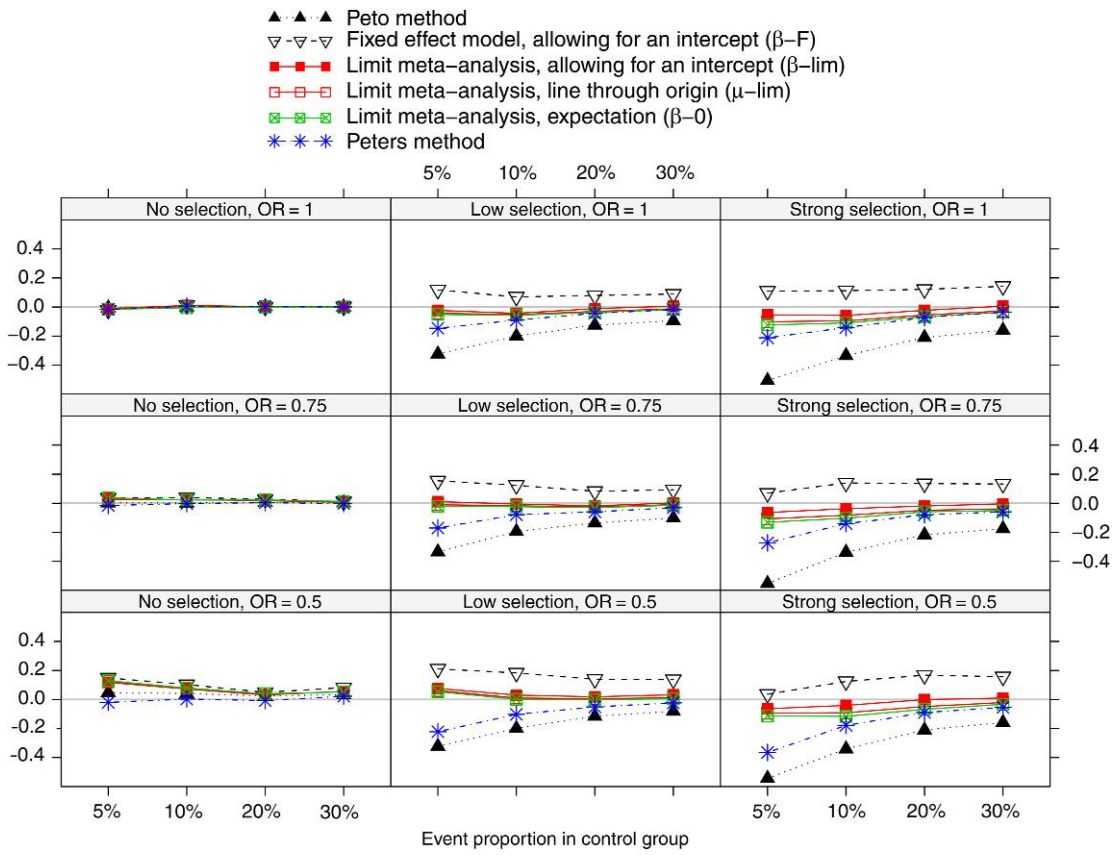

Fig. 2. Bias of treatment-effect estimate on log-odds ratio scale ($\log \widehat{\mathrm{OR}} - \log \mathrm{OR}$) for $\tau^2 = 0.1$ and 10 studies per meta-analysis, 6 models. Various scenarios.

was best within the classical models. The larger the small-study effect, the greater was the superiority of the new methods over the random-effects model. The limit meta-analysis expectation $\hat{\beta}_0$ and the Peters method had best coverage within all adjusted models. There was almost no dependence on the true odds ratio but a strong dependence on the control event rate, with an interaction between this and the model used: If there was selection, coverage increased with increasing control event rate for the classical models, while it always decreased for both the new methods and the Peters method. The poor coverage of $\hat{\mu}_{\mathrm{lim}}$ seems to be due to underestimation of its standard error.

7. A NEW MEASURE OF HETEROGENEITY

The promising results for adjusted treatment-effect estimates based on the limit meta-analysis motivated us to derive a new measure for heterogeneity, called G^2. It is determined on a percentage scale, following the established measure I^2 introduced by Higgins and Thompson (2002). G^2 assesses the proportion of the variance that is unexplained after we have allowed for possible small-study effects in the limit meta-analysis.

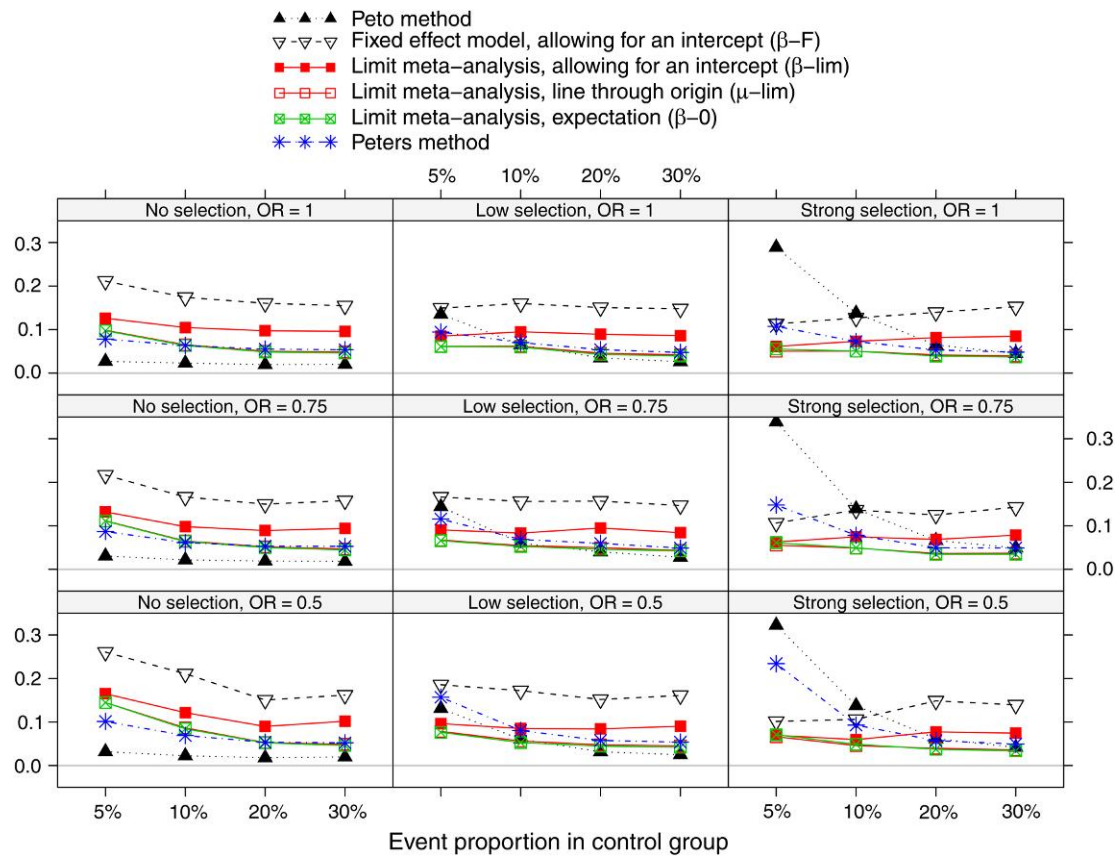

Fig. 3. MSE of treatment-effect estimate on log-odds ratio scale for $\tau^2 = 0.1$ and 10 studies per meta-analysis, 6 models. Various scenarios.

We note first that in a regression model there is a natural measure of the proportion of the unexplained variance:

$$1 - R^2_{\text{reg}} = \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}},$$

where R^2_{reg} is consistently estimated by the squared Pearson correlation coefficient. In analogy to this, G^2 is defined with respect to the fixed-effect model fitted to the limit meta-analysis and allowing for small-study effects (corresponding to parameter β_{lim}). Specifically, G^2 is defined as $1 - R^2_{\text{reg}}$ when regressing the standardized treatment-effect estimates y_i/s_i, obtained from the limit meta-analysis, on $1/s_i$. This is easily done using standard linear regression software. Thus G^2 is given by

$$G^2 = 1 - \frac{\left[\sum w_i y_i - \frac{1}{k}\left(\sum \sqrt{w_i}\right)\left(\sum \sqrt{w_i} y_i\right)\right]^2}{\left[\sum w_i - \frac{1}{k}\left(\sum \sqrt{w_i}\right)^2\right]\left[\sum w_i {y_i}^2 - \frac{1}{k}\left(\sum \sqrt{w_i} y_i\right)^2\right]}.$$

G^2 is closely related to the heterogeneity statistic Q', defined in Section 5 and measuring residual variation with respect to a fixed-effect model allowing for small-study effects, while G^2 is based on the

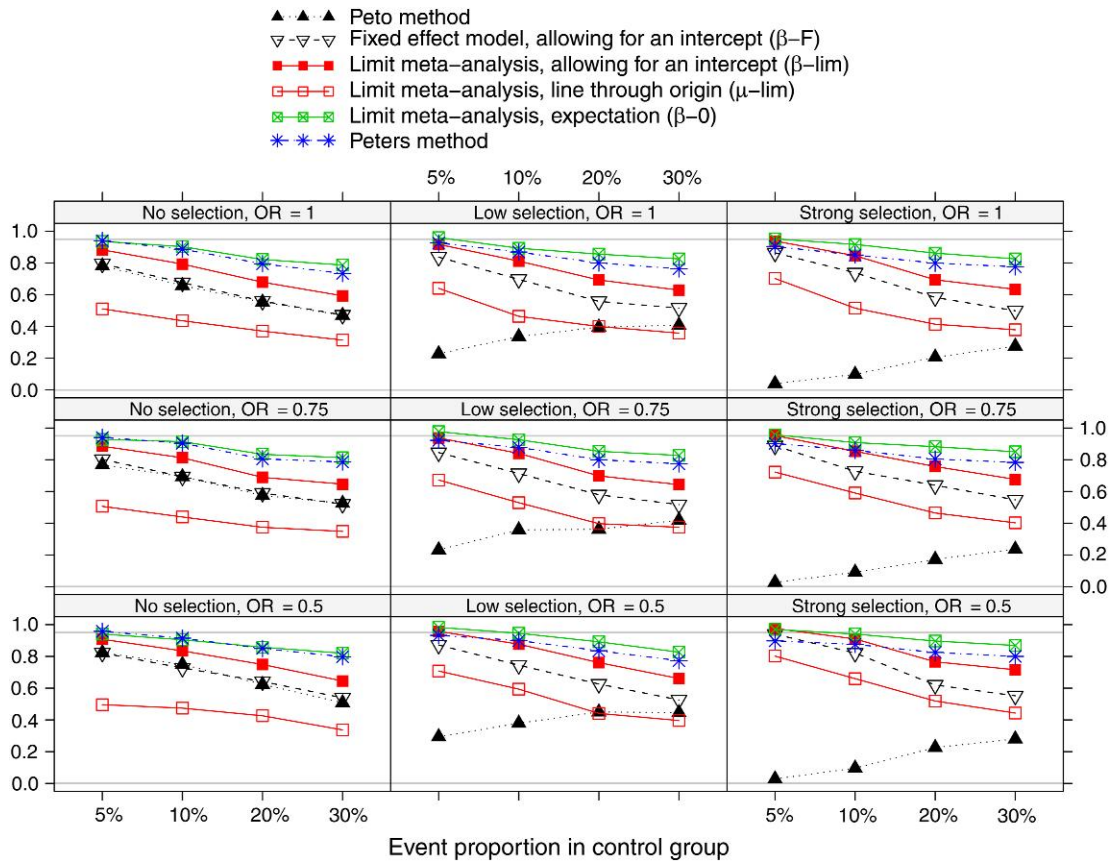

Fig. 4. Coverage of 95% confidence intervals of treatment-effect estimate on log-odds ratio scale for $\tau^2 = 0.1$ and 10 studies per meta-analysis, 6 models. Various scenarios.

limit meta-analysis and scaled to [0,1] by dividing the residual sum of squares by the total sum of squares:

$$G^2 = \frac{Q'_{\text{Limit meta-analysis}}}{\text{Total Sum of Squares in limit meta-analysis}}. \tag{7.1}$$

For testing for residual heterogeneity after allowing for small-study effects, we propose to apply the Q' statistic to the given analysis instead of the limit meta-analysis. The reason is that the shrinkage process leading to the limit meta-analysis removes random error and thus causes both Q and Q' to decrease very much. Thus, we propose the following test procedure (using the same level for all 3 tests):

1. For a test of heterogeneity in the usual sense (that is, not distinguishing between heterogeneity caused by small-study effects and heterogeneity from other causes), take Q, which under the null hypothesis of no heterogeneity follows a χ^2_{k-1} distribution. Stop if the test is not significant.
2. If the null hypothesis of no heterogeneity is rejected, carry out an appropriate test of small-study effects, for example, by taking $Q - Q'$, which under the null-hypothesis of no small-study effects follows a χ^2_1 distribution.

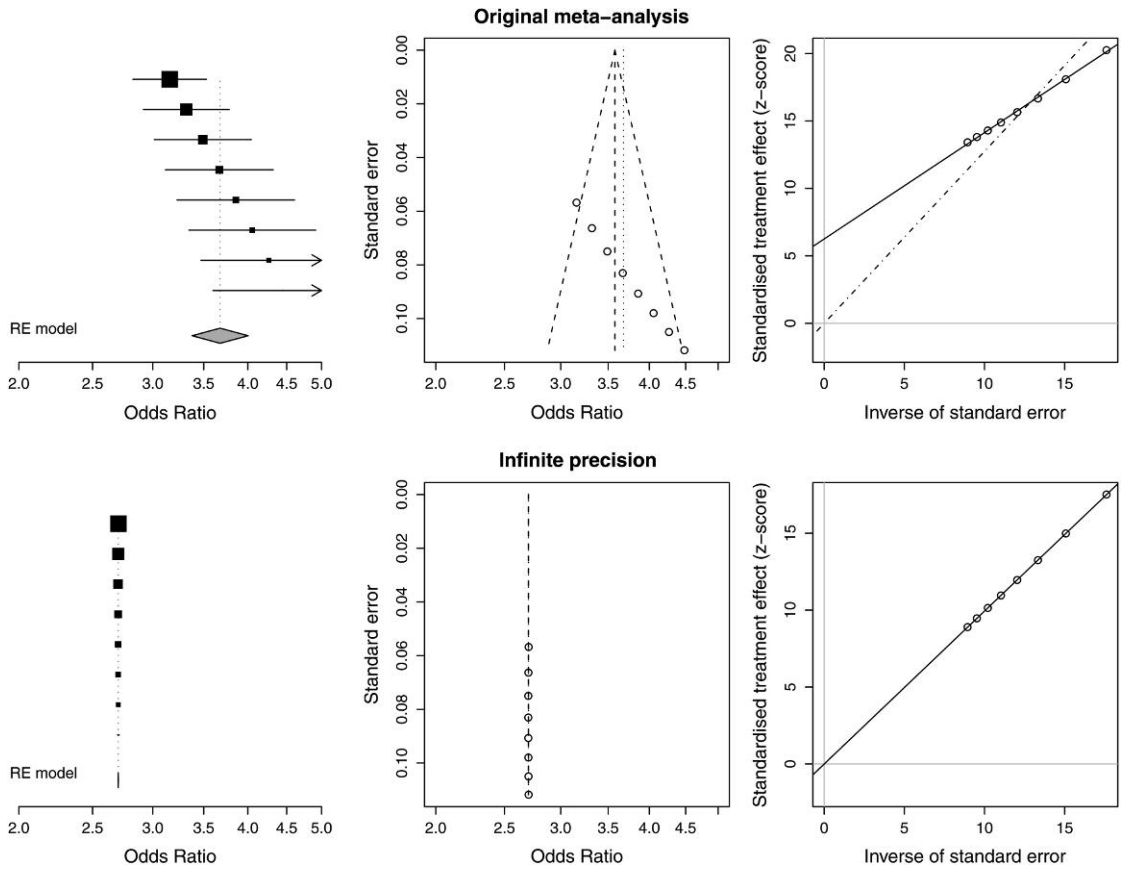

Fig. 5. Forest plot (left), funnel plot (middle), and radial plot (right) for original (top) and limit (bottom) meta-analysis (fictional example, see text).

3. If the null hypothesis of no heterogeneity is rejected, test for residual heterogeneity beyond small-study effects, using Q', which under the null hypothesis of no residual heterogeneity follows a χ^2_{k-2} distribution. Residual heterogeneity can be quantified by G^2.

7.1 *Examples*

In this subsection, we look at 4 examples, representing typical settings: no heterogeneity; small-study effects without additional heterogeneity; and heterogeneity including minor or major small-study effects.

7.1.1 *No statistical heterogeneity.* If no statistical heterogeneity is found in the given meta-analysis (that is, $\hat{\tau}^2 = 0$), the limit meta-analysis yields equal $y_i = \hat{\beta}_F$ for all trials, and the limit radial plot is perfectly fitted by the regression line, that is, $G^2 = 0$. Though a test on potential small-study effects may be significant, notice that—as an immediate consequence of the fixed-effect model—all deviations from the mean are random by definition, which means that any apparent small-study effect must be spurious as well. In this case, funnel plot asymmetry disappears with increasing precision. Note, in passing, that the requirement of homogeneity when testing for funnel plot asymmetry—see, for instance, Ioannidis and Trikalinos (2007)—may be reduced *ad absurdum*.

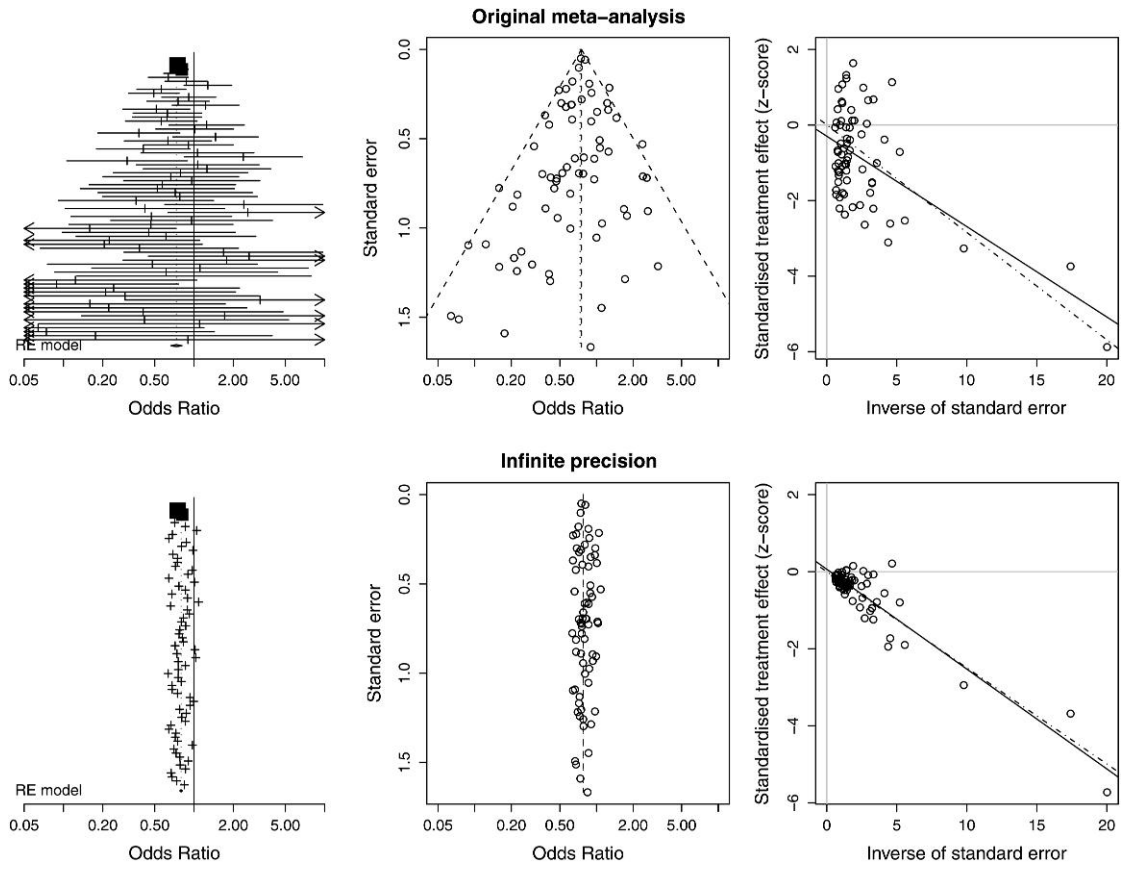

Fig. 6. Thrombolytic therapy data (Olkin, 1995). Forest plot (left), funnel plot (middle), and radial plot (right) for original (top) and limit (bottom) meta-analysis.

7.1.2 *Small-study effects without additional heterogeneity.* G^2 is based on a notion of heterogeneity different from that usually used, in the sense that it does not incorporate potential small-study effects. Thus, it is possible that $G^2 = 0$ while $\hat{\tau}^2 > 0$, that is, there is no other heterogeneity apart from that due to small-study effects (to which τ^2 is sensitive but not G^2). This can be illustrated by a fictional, somewhat pathological example (Figure 5). Here, we see a striking small-study effect so that all dots in the radial plot lie more or less exactly on one line not going through the origin. Therefore, all variation is explained by a fixed-effect model that allows for small-study effects, and residual heterogeneity measured by G^2 is almost zero, though $\tau^2 > 0$ ($Q = 15.46(p = 0.031)$, $I^2 = 54.7\%[0\%; 79.5\%]$, $G^2 = 0.01\%$). Adjusting for small-study effects yields residual heterogeneity $Q' = 0.052(p = 1.000)$, that is, there is no heterogeneity left beyond the small-study effect, which itself is significant: $Q - Q' = 15.41$ ($p \leqslant 0.001$).

7.1.3 *Heterogeneity: thrombolytic therapy data.* We use the example introduced above, see Figure 1 and also Figure 6. The various estimates are given in Table 1. This is an example showing some heterogeneity ($\hat{\tau}^2 = 0.018$). The limit radial plot looks similar to the original one and shows some residual variation. We find $I^2 = 18.6\%$, whereas $G^2 = 15.4\%$. For this meta-analysis, most tests indicate a minor small-study effect (e.g. $p = 0.075, 0.088, 0.091, 0.063$ for Egger's test, Harbord's test, Peters' test, and the arcsine

Table 1. *Estimated odds ratios from different models for the thrombolytic therapy data example (Olkin, 1995)*

	Original meta-analysis		Limit meta-analysis
Model	Fixed-effect model	Random-effects model	Fixed-effect model
Plot	Radial plot	Generalized radial plot	Limit radial plot
Model without intercept	$\exp(\mu_F)$ 0.753 [0.710; 0.798]	$\exp(\mu_R)$ 0.732 [0.664; 0.808]	$\exp(\mu_{lim})$ 0.779 [0.735; 0.826]
Model with intercept	$\exp(\beta_F)$ 0.787 [0.731; 0.847]	$\exp(\beta_R)$ 0.840 [0.710; 0.993]	$\exp(\beta_{lim})$ 0.772 [0.717; 0.831]
Expectation			$\exp(\beta_0)$ 0.796 [0.739; 0.857]

test, respectively, see Egger *and others*, 1997; Harbord *and others*, 2006; Peters *and others*, 2006; Rücker *and others*, 2008a). The value of G^2 indicates residual heterogeneity beyond this, not explained solely by a fixed-effect model allowing for small-study effects. However, it is is not significant, if tested via Q' ($Q' = 80.84$, $p = 0.137$). Further, we find $Q = 84.73 (p = 0.096)$ and $Q - Q' = 3.884$, again consistent with a small-study effect ($p = 0.049$).

7.1.4 *Heterogeneity and small-study effects: passive smoking data.* This example by Hackshaw *and others* (1997) was intensively discussed in the literature for several reasons (for details, see Copas and Shi, 2000b; Senn, 2009). It was also used by Copas and Malley (2008) when deriving a robust *P*-value for the treatment effect in meta-analysis. We find $Q = 47.52 (p = 0.095)$, $Q' = 40.96 (p = 0.225)$, and therefrom $Q - Q' = 6.55 (p = 0.010)$, indicating a small-study effect. The plots are shown in Figure 7, the estimates given in Table 2. Both the fixed-effect model (μ_F) and the random-effects model (μ_R) find an odds ratio of about 1.2 and thus a significant excess risk of lung cancer for persons exposed to passive smoking. The effect is reduced when using the limit meta-analysis but still significant (μ_{lim}). By contrast, it vanishes completely if adjusting for small-study effects, with nearly concordant estimates for β_F, β_R, β_{lim}, and β_0. For this meta-analysis, heterogeneity seems moderate, measured by $\hat{\tau}^2 = 0.0168$ and also $I^2 = 24.2\%$. However, at first glance surprisingly, G^2 is very large ($G^2 = 94.6\%$), indicating much residual heterogeneity, caused by the fact that just 2 of the 3 large dominating studies show reciprocal effects with mutually exclusive confidence intervals. Since these studies are large, they are relatively insensitive to the shrinkage process, see the bottom left and bottom right panels of Figure 7. The passive smoking data example therefore shows that G^2 is particularly sensitive to heterogeneity in large trials.

8. DISCUSSION

We have introduced the concept of limit meta-analysis, derived from an extended random-effects model for meta-analysis which includes a parameter for the bias introduced by potential small-study effects. The limit meta-analysis takes into account small-study bias to yield shrunken estimates of individual study effects. We showed that these shrunken estimates can also be justified from an empirical Bayesian viewpoint. Our approach is thus consistent with the philosophy of random-effects modeling that "inference for each particular study is performed by "borrowing strength" from the other studies" (Higgins *and others*, 2009). We are essentially correcting for possible small-study effects before we "borrow strength" from other studies. The limit meta-analysis gives rise to 3 possible estimators for the overall intervention

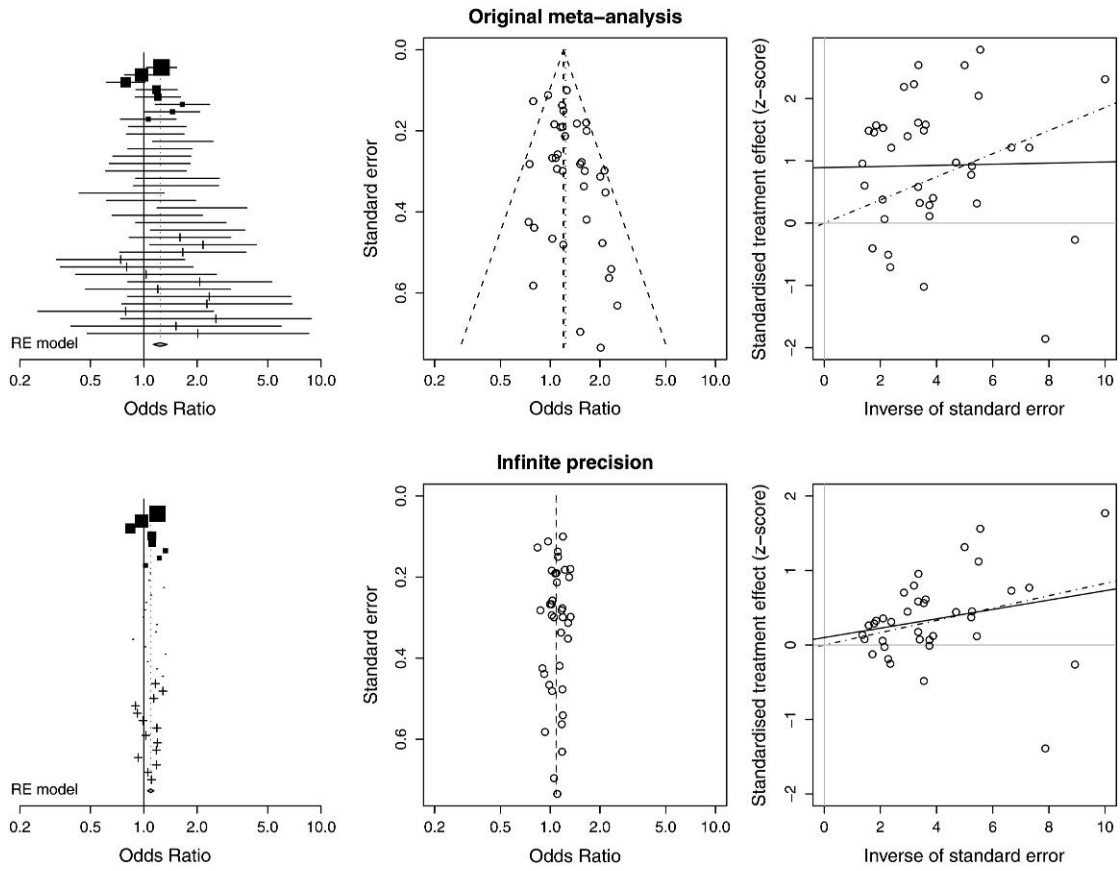

Fig. 7. Passive smoking data (Hackshaw *and others*, 1997). Forest plot (left), funnel plot (middle), and radial plot (right) for original (top) and limit (bottom) meta-analysis.

Table 2. *Estimated odds ratios from different models for the passive smoking data (Hackshaw and others, 1997)*

	Original meta-analysis		Limit meta-analysis
Model	Fixed-effect model	Random-effects model	Fixed-effect model
Plot	Radial plot	Generalized radial plot	Limit radial plot
Model without intercept	$\exp(\mu_F)$ 1.204 [1.120; 1.295]	$\exp(\mu_R)$ 1.238 [1.129; 1.357]	$\exp(\mu_{lim})$ 1.086 [1.010; 1.168]
Model with intercept	$\exp(\beta_F)$ 1.009 [0.865; 1.176]	$\exp(\beta_R)$ 0.973 [0.757; 1.250]	$\exp(\beta_{lim})$ 1.065 [0.913; 1.242]
Expectation			$\exp(\beta_0)$ 1.094 [0.939; 1.276]

effect, adjusting for small sample effects, as well as a measure of heterogeneity after accounting for small sample effects.

We derived ML estimates of the quantities in the limit meta-analysis in Section 4. However, we show—as a direct consequence of interpreting the limit meta-analysis in terms of the radial plot in Section 5—that

parameter estimation in limit meta-analyses can be carried out using existing software. This removes one hurdle to its use in practice.

The ultimate aim of quantitative meta-analysis is to arrive at a pooled estimate of the intervention effect. We thus performed a comprehensive simulation study with binary response data to compare the 3 proposed estimators of the pooled effect that emerged from our limit meta-analysis with currently used estimators. The simulation study explored a range of effect sizes, underlying event probabilities, heterogeneity (unrelated to small-study effects) and selection (as a surrogate for small-study effects). All 3 proposed estimators had small bias and comparable mean square error in the presence of small-study effects. However, the "β_0" estimator (2.4) gave the best confidence interval coverage and is thus our preferred estimator in the presence of small-study effects.

Unfortunately, neither the proposed nor existing estimators performed acceptably in all situations—that is, in both situations where small-study effects were present and situations where small-study effects were not present. In practice, analysts must therefore decide which estimator to use. To support this decision, we advocate one of the more recent tests for publication bias. A useful summary is given in Chapter 10, Section 4, of the Cochrane Handbook for Systematic Reviews of Interventions (Higgins and Green, 2009). We acknowledge that some authors take a censorious attitude to such tests, believing them misleading (Lau *and others*, 2006; Terrin *and others*, 2003; Tang and Liu, 2000), stigmatizing them as "pseudo tests" (Ioannidis, 2008), and questioning whether funnel plots are a suited means at all for judging small-study effects (Terrin *and others*, 2005). However, when applied following a prespecified analysis protocol, with their limitations duly acknowledged, we argue that such concerns (Ioannidis and Trikalinos, 2007) are minimized and that testing is a useful aide to researchers in judging funnel plots. After all, adjusted treatment-effect estimates were used successfully for predicting the effect of the whole database of antidepressant trials in the food and drug administration registry from a biased subset of published trials (Moreno *and others*, 2009b).

In this paper, we have not considered the source of small-study effects, be it publication bias or heterogeneity arising from differing patient, or other study specific, characteristics. Specifically, when adjusting the treatment-effect estimate for small-study effects, it does not matter where the small-study effect comes from (Moreno *and others*, 2009a). The limit meta-analysis can be readily extended to adjust for any covariates which explain heterogeneity; it then would address remaining unexplained small-study effects.

An even more provoking question was raised by Stanley *and others* (2010), whether it "could be better to discard 90% of the data," arguing that in the presence of small-study effects all adjusting methods lead to estimates that are very similar to the results of the one or 2 largest studies. However, these may also disagree, as illustrated by the passive smoking data example.

Of course, the above process may not explain all the heterogeneity, and we propose the test statistic Q' and the measure G^2 to assess and quantify, respectively, the remaining heterogeneity after adjusting for small-study effects. If we believe the principal source of small-study effects is publication bias, then detecting and investigating heterogeneity "after this has been accounted for" is arguably of greater scientific relevance—as it relates directly to factors affecting the efficacy of the intervention in practical settings.

A potential drawback of our approach is its dependence on the estimation of τ^2, for which a number of competing estimators are given in the literature. In this article, we have used the methods-of-moments estimator (DerSimonian and Laird, 1986). This estimator is both the most widely accepted and used. It is implemented in the Review Manager software for Cochrane reviewers (The Cochrane Collaboration, 2009). Unfortunately, the difference between these estimators tends to be greater the smaller the number of studies in the meta-analysis and the smaller the true heterogeneity. We therefore revisited our analysis of both the thrombolytic therapy and the passive smoking meta-analyses, using 7 options for estimating τ^2 available in the R package metafor (R Development Core Team, 2008). We found that while our 3 pooled effect estimators were relatively robust, the estimate of G^2 varied considerably. Thus, we prefer

to use the test statistic Q' to assess heterogeneity and report G^2 as a measure of such heterogeneity, possibly also reporting the latter for a range of estimates of τ^2.

Another issue is the use of the Copas selection model for generating the data in the simulation study. Strictly, in using this model, we are generating data from a slightly different model than we are fitting to the data. However, if a method is reliable in this setting, this provides reassurance for its use in practice, where we cannot know the data generation model.

To conclude, we have introduced the idea of a limit meta-analysis which we believe is a promising approach for finding "shrunk," empirical Bayes, estimates of study effects in the presence of small sample bias. This led to 3 proposed estimators for an overall effect in the presence of small sample bias. Our simulation study suggested all 3 methods had smaller bias and mean square error than estimators which did not account for small sample bias. One of these 3 methods, the "expected limit estimate," also had good confidence interval coverage and is our preferred method for use in practice. We have also described an approach for assessing heterogeneity after accounting for small-study effects, and illustrated its utility with a reanalysis of data on the effects of passive smoking.

Software

All calculations were carried out using the freely available software R, version R-2.10.1, particularly using the packages meta (Schwarzer, 2007) and metafor (R Development Core Team, 2008). R code for calculation of all estimates given in this paper can be obtained from the first author.

Acknowledgment

Conflict of Interest: None declared.

Funding

Deutsche Forschungsgemeinschaft (FOR 534 Schw 821/2-2 to G.R. and J.R.C).

References

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

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10**, 101–129.

Copas, J. and Lozada-Can, C. (2009). The radial plot in meta-analysis: approximations and applications. *Applied Statistics* **58**, 329–344.

Copas, J. and Shi, J. Q. (2000a). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* **1**, 247–262.

Copas, J. B. and Malley, P. F. (2008). A robust p-value for treatment effect in meta-analysis with publication bias. *Statistics in Medicine* **27**, 4267–4278.

Copas, J. B. and Shi, J. Q. (2000b). Reanalysis of epidemiological evidence on lung cancer and passive smoking. *British Medical Journal* **320**, 417–418.

Copas, J. B. and Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research* **10**, 251–265.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

Duval, S. and Tweedie, R. (2000). A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* **95**, 89–98.

EGGER, M., SMITH, G. D., SCHNEIDER, M. AND MINDER, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629–634.

ENGELS, E. A., SCHMID, C. H., TERRIN, N., OLKIN, I. AND LAU, J. (2000). Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* **19**, 1707–1728.

GALBRAITH, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine* **7**, 889–894.

GREENLAND, S. AND O'ROURKE, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* **2**, 463–471.

GREENLAND, S. AND ROBINS, J. M. (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics* **41**, 55–68.

HACKSHAW, A. K., LAW, M. R. AND WALD, N. J. (1997). The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical Journal* **315**, 980–988.

HARBORD, R. M., EGGER, M. AND STERNE, J. A. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine* **25**, 3443–3457.

HARDY, R. J. AND THOMPSON, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* **17**, 841–856.

HIGGINS, J. P. AND GREEN, S. (2009). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2*. http://www.cochrane-handbook.org.

HIGGINS, J. P., THOMPSON, S. G. AND SPIEGELHALTER, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A* **172**, 137–159.

HIGGINS, J. P. T. AND THOMPSON, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539–1558.

IOANNIDIS, J. P. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice* **14**, 951–957.

IOANNIDIS, J. P. A. AND TRIKALINOS, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Canadian Medical Association Journal* **176**, 1091–1096.

JACKSON, D. (2006). The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine* **25**, 2911–2921.

KNAPP, G., BIGGERSTAFF, B. J. AND HARTUNG, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal* **48**, 271–285.

LAU, J., IOANNIDIS, J. P. A., TERRIN, N., SCHMID, C. H. AND OLKIN, I. (2006). The case of the misleading funnel plot. *British Medical Journal* **333**, 597–600.

MCCULLAGH, P. AND NELDER, J. (1989). *Generalized Linear Models*. London: Chapman and Hall.

MITTLBÖCK, M. AND HEINZL, H. (2006). A simulation study comparing properties of heterogeneity measures in meta-analyses. *Statistics in Medicine* **25**, 4321–4333.

MORENO, S., SUTTON, A., ADES, A., STANLEY, T., ABRAMS, K., PETERS, J. AND COOPER, N. (2009a). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology* **9**, 2. http://www.biomedcentral.com/1471-2288/9/2/abstract.

MORENO, S. G., SUTTON, A. J., TURNER, E. H., ABRAMS, K. R., COOPER, N. J., PALMER, T. P. AND ADES, A. E. (2009b). Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *British Medical Journal* **339**, 494–498.

NISSEN, S. E. AND WOLSKI, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular diseases. *The New England Journal of Medicine* **356**, 2457–2471.

OLKIN, I. (1995). Statistical and theoretical considerations in meta-analysis. *Journal of Clinical Epidemiology* **48**, 133–146.

PETERS, J. L., SUTTON, A. J., JONES, D. R., ABRAMS, K. R. AND RUSHTON, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association* **295**, 676–680.

RABE-HESKETH, S. AND SKRONDAL, A. (2005). *Multilevel and Longitudinal modeling using STATA*. College Station, TX: Stata Press.

RAUDENBUSH, S. W. AND BRYK, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics* **10**, 75–98.

R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org.

ROTHSTEIN, H. R., SUTTON, A. J. AND BORENSTEIN, M. (2005). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Chichester, UK: Wiley.

RÜCKER, G., SCHWARZER, G. AND CARPENTER, J. R. (2008a). Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* **27**, 746–763.

RÜCKER, G., SCHWARZER, G., CARPENTER, J. R. AND SCHUMACHER, M. (2008b). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology* **8**, 79. http://www.biomedcentral.com/1471-2288/8/79.

SCHWARZER, G. (2007). Meta: An R package for meta-analysis. *R News* **7**, 40–45. http://cran.r-project.org/doc/Rnews/Rnews_2007-3.pdf.

SCHWARZER, G., ANTES, G. AND SCHUMACHER, M. (2002). 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**, 2465–2477.

SCHWARZER, G., ANTES, G. AND SCHUMACHER, M. (2007). A test for publication bias in meta-analysis with sparse binary data. *Statistics in Medicine* **26**, 721–733.

SENN, S. (2000). The many modes of meta. *Drug Information Journal* **34**, 535–549.

SENN, S. J. (2009). Overstating the evidence: double counting in meta-analysis and related problems. *BMC Medical Research Methodology* **9**, 10.

SIDIK, K. AND JONKMAN, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical, Series C (Applied Statistics)* **54**, 367–384.

STANLEY, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics* **70**, 105–127.

STANLEY, T. D., JARRELL, S. B. AND DOUCOULIAGOS, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician* **64**, 70–77.

STERNE, J. A. C., GAVAGHAN, D. AND EGGER, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* **53**, 1119–1129.

STIJNEN, T. AND HOUWELINGEN, J. C. V. (1990). Empirical Bayes methods in clinical trials meta-analysis. *Biometrical Journal* **32**, 335–346.

TANG, J.-L. AND LIU, J. L. Y. (2000). Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology* **53**, 477–484.

TERRIN, N., SCHMID, C. H. AND LAU, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology* **58**, 894–901.

TERRIN, N., SCHMID, C. H., LAU, J. AND OLKIN, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine* **22**, 2113–2126.

THE COCHRANE COLLABORATION (2009). Review manager (RevMan) [computer program]. version 5.0.23. http://ims.cochrane.org/revman.

THOMPSON, S. G. AND SHARP, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* **18**, 2693–2708.

VERBEKE, G. AND MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer series in statistics. New York: Springer.

VIECHTBAUER, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine* **26**, 37–52.

YUSUF, S., PETO, R., LEWIS, J., COLLINS, R. AND SLEIGHT, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* **27**, 335–371.

[*Received July 28, 2009; revised April 13, 2010; accepted for publication June 16, 2010*]