Combining Statistical Evidence

Elena Kulinskaya Stephan Morgenthaler Robert G. Staudte *

SUMMARY

The combination of evidence from independent studies has a curious history. The origins reach back at least to the beginning of the twentieth century. Since the mid-seventies the term meta-analysis (G. Glass, 1976 presidential address to the American Educational Research Association) has become popular in several fields, among them medical statistics and the behavioral sciences. The most widely used procedures were perfected in early papers and subsequently a kind of groupthink has taken hold of meta-analysis. This explains the need for a review in a statistics journal, destined for a statistical audience. Meta-analysis is not a hot research topic among graduate students in statistics and by writing this article we hope to change this. We wish to point out the shortcomings of the mainstream view and exhibit some of the open problems that await the attention of statistical researchers.

A host of competent reviews of meta-analysis have been published and several book-length treatments are also available. We have listed many of these in the bibliography, but cannot guarantee completeness.

Keywords: review, meta-analysis, effect size, random effects, meta-regression, software

*Elena Kulinskaya is Professor, School of Computing Sciences, University of East Anglia Norwich NR4 7TJ, United Kingdom (e-mail: e.kulinskaya@uea.ac.uk); Stephan Morgenthaler is Professor, Ecole polytechnique fédérale de Lausanne (EPFL), Station 8, 1015 Lausanne, Switzerland (e-mail: stephan.morgenthaler@epfl.ch); and Robert G. Staudte is Professor, La Trobe University, Department of Statistics and Mathematics, Melbourne, Australia 3086 (e-mail: r.staudte@latrobe.edu.au).

RÉSUMÉ

La mise en commun des résultats de plusieurs études individuelles portant sur la même question possède une drôle d'historique. Les débuts remontent au début du 20^e siècle. Depuis le milieu des années septantes le terme méta-analyse (G. Glass, 1976 discours presidentiel lors de la conférence de l'American Educational Research Association) est devenu populaire dans plusieurs domaines, entre autres la statistique médicale et les sciences du comportement et de psychologie. Les méthodes les plus utilisées ont été élaborées assez vite et ensuite une sorte de pensée unique s'est emparée de la méta-analyse. Cela explique la nécessité d'un bilan dans un journal déstiné à un publique de statisticien(ne)s. La méta-analyse n'est pas un sujet apprécié par nos étudiants doctorants et nous espérons de changer cela avec cet article. Nous voulons montrer les points faibles de l'approche habituelle et explorer quelques problèmes ouverts qui posent des défis aux chercheurs en statistique.

Une série de bilans compétents de la méta-analyse ont été publiés et plusieurs livres sont également à disposition. Notre article fait référence à beaucoup d'entre eux, mais nous avons sûrement oublié quelques-uns.

Mots clés: bilan, méta-analyse, ampleur de l'effet, effets aléatoires, méta-régression, logiciels

1 Introduction

The broad aim of a meta-analysis (MA) is to provide a review of the literature on some scientific question and to summarize the information in a quantitative manner. It is hoped that by combining the statistical evidence from all the available studies a stronger consensus view can emerge. Meta-analysis taken in this sense is merely a part of a systematic review, which involves an extensive methodology for research synthesis going beyond meta-analysis. A systematic review requires the identification of the relevant studies, the determination how each study was conducted and under what precise circumstances, the collection of the associated data , the evaluation of the study quality, etc. But it is fair to say that unless the statistical methods being used are sound, the numerical summaries from a meta-analysis may be misleading and no benefits may accrue from combining the studies.

Each study or trial used in a meta-analysis is summarized by an estimate of an appropri-

ate *effect*, for example, the odds-ratio when comparing two samples with binary outcomes, or the mean, when the study consists of a single sample with continuous outcomes, and so on. The meta-analysis attempts to provide a more precise estimate. The single most important reason for failures in meta-analyses are biases. Many estimators, for example, are biased when applied to finite samples. Combining even a very large number of finite sample estimates does not make this bias disappear. On the contrary, its relative importance grows, because the variance decreases towards zero by combining studies. Another example are study-related biases. Does it makes sense to estimate an overall effect via a meta-analysis? To what populations will the result apply? Sequences of studies are usually done with differing standards, different definitions and methods, different populations, and so on. Sometimes, covariates may help in explaining the differences at least partially. In other instances it might be better to refrain from attempting a combination. Yet other biases are more subtle. Difficulties in publishing papers with insignificant findings, for example, will result in publication bias which exaggerates the true effect. The history of science provides many examples in which statistically significant effects were identified, published and even celebrated, only to be refuted later on. It seems all too easy to arrive at such wrong conclusions, particularly if the effect is believed to be real by the scientific community. Statisticians have already contributed to the resolution of publication bias (see Section 3), but much more needs to be done.

The second order effects, such as variances or lengths of confidence intervals, are of lesser importance, but nevertheless deserve attention. One should make sure that all sources of variation are accounted for when computing the variance of an estimated effect. The resulting confidence intervals will be wider, but also more realistic. It clearly would also be desirable to include potential biases when calculating confidence intervals. But this requires a rethink of our current theories of testing and inference.

In view of the unmet needs in the area of bias and variance, it seems almost unnecessary to talk about the higher order effects due to deviations from Gaussianity. Clearly, however, more accurate models will lead to more accurate inferences and it presumably would be useful for meta-analysis to use more sophisticated models.

1.1 A very brief history of meta-analysis

An early paper is an investigation reported in Simpson and Pearson (1904) of correlation coefficients in 2×2 contingency tables of incidence or survival vs. inoculation. Data sets from South Africa and India are taken into consideration and the paper's main achievements are the organization and reduction of the data into comparable form and the summary of the data via a correlation coefficient with accompanying probable error. The only meta-statistic Simpson and Pearson present is the mean across the data sets. One of the conclusions of the paper states that "To sum up, it seems that, while most of the correlations both for immunity and recovery are distinctly sensible, having regard to their probable error, yet they are so irregular that little reliance can be placed upon them as representing a definitive uniform effect." (Simpson and Pearson, 1904, p. 1244). This remains a valid statement of one of the main difficulties one encounters when combining estimated effects.

The analysis of variance, a technique developed by R.A. Fisher in the nineteen-thirties, can be seen as a basic tool for meta-analysis. The simplest case concerns several parallel samples (1-way ANOVA), with the factor denoting the studies. The common mean represents the common effect, while the individual study effects can be used to assess the variation between studies. The residual variation, finally, is a measure of the within study variation. The first papers written from this point of view appeared in the late thirties (Cochran, 1937; Yates and Cochran, 1938). In Cochran's paper a setup is introduced that is still in use today. The individual samples – or centres as he calls them – are summarized by the effect estimate x_i (the mean) and its standard error s_i based on a known number of degrees of freedom. An even simpler summary – merely a p-value of a significance test for a null hypothesis common to all centers — was discussed in Tippett et al. (1931) and Fisher (1932). They proposed methods for computing a combined p-value.

For most statisticians, meta-analysis became an area to avoid, perhaps because they mistakenly thought that such analyses were straightforward. Notable exceptions include Gene V. Glass, Larry Hedges, and Ingram Olkin.

1.2 Data in a typical meta-analysis

It is of interest to understand what data constitutes a 'typical' meta-analysis. We analysed all meta-analyzes using the difference of means as an effect measure from issue 4 of the

Cochrane database (2004, compact disk edition). All these studies have two arms: a treatment and a control arm with sample sizes denoted be n_T and n_C, respectively, with the overall sample size $n = n_T + n_C$. The data consists of sample means and sample variances in the two arms. There were 4,585 meta-analyses having $K > 1$ studies and positive variances in both arms. Interestingly, the numbers of studies per meta-analysis are small: 47% with $K = 2$, 21% with $K = 3$, and in total 95% with $K \leq 8$, though the maximum was $K = 58$.

An assumption often taken for granted in theoretical work is that of large sample sizes. In fact, the majority of meta-analyses include some small studies. The minimum study size is 20 or less in 25% of meta-analyses; ≤ 33 in 50%, and ≤ 70 in 75%. Often all studies in a meta-analysis are rather small: maximum study size is 50 or less in 25% of meta-analyses and less than ≤ 110 in 50%. Only 10% of the meta-analyses include one or more studies of 490 or more patients. The majority of the studies are fairly balanced: in 75% of the meta-analyses $\min\{n_T/n, n_C/n\} \geq 0.44$, and only in 10% $\min\{n_T/n, n_C/n\} \leq 0.33$.

1.3 What is to come

The paper offers a review of the procedures and open problems in statistical meta-analysis. This is a topic of growing importance, because in many areas of application the need for combining different sources of data and different sources of information in order to reach an overall assessment manifests itself. The classical material on meta-analytic statistical procedures are reviewed in Section 2. These include the distinction between fixed and random effects models, tests of homogeneity, meta-regression, observational studies, and a discussion of the types of data typically available for a meta-analysis. Section 3 investigates two sources of bias in meta-analysis, the bias due to the systematic selection of studies showing stronger than average effects, and the meta-analysis of smallish studies combined with the use of estimators with appreciable small sample biases. In Sections 4 and 5 generalizations and extensions to the standard procedures are discussed. They include multivariate responses and sequential procedures. A non-exhaustive list of software tools closes out the review.

2 Models for meta-analysis and meta-regression

For a statistician, standard meta-analysis is very close to fixed or random effects 1-way ANOVA under heteroscedasticity, complicated by mere asymptotic rather than exact normality of the participating statistics. A closely linked area is the analysis of interlaboratory studies. Exact distributional results do not exist in a closed form, and various approximations are in use. There are three possible ways to derive these approximations: increasing within-study sample sizes $n_k \to \infty$ for a fixed number of studies K; increasing number of studies $K \to \infty$ for fixed or bounded study sizes; and also for both K and n_k increasing simultaneously. These three options are often erroneously interchanged. As we shall see, they result in very different inferential procedures.

A reader who is interested in more detailed information about meta-analysis, can consult one of the recently published books on the subject. Here is a selection in chronological order: Hedges and Olkin (1985), Wolf (1986), Sutton et al. (2000), Whitehead (2002), Schulze et al. (2003), Rothstein et al. (2006), Kulinskaya et al. (2008), Cooper et al. (2009), Higgins and Green (2011), Borenstein et al. (2011), Stanley and Doucouliagos (2012), Pigott (2012), Koricheva et al. (2013).

2.1 Fixed effects model

We are given K studies, each trying to measure some effect θ. The effects can be measured by a variety of statistics, such as sample means, correlation coefficients and, for studies in which there are treatment and control arms, difference of sample means, standardized mean differences, odds ratios, and differences or ratios of binomial probabilities known as risk differences and relative risks. In the kth study or trial there are n_k observations yielding an estimator $\hat{\theta}_k$, which is asymptotically normal in the sense that $\sqrt{n_k}\,(\hat{\theta}_k - \theta_k) \to \mathbb{N}(0, v_k)$ in distribution for some unknown parameters (θ_k, v_k). A second assumption is that for each k a consistent estimator $\hat{v}_k$ of v_k exists in the sense that $\hat{v}_k/v_k \to 1$ in probability. This justifies large sample confidence intervals for θ_k of the form $\hat{\theta}_k \pm z_{1-\alpha/2}\{\hat{v}_k/n_k\}^{1/2}$, with confidence coefficient $1-\alpha$. It further follows for the *known* inverse variance weights $w_k = n_k/v_k$, that a large sample confidence interval for the combined or meta effect $\theta_w = \sum_k w_k\theta_k/W$, where $W = \sum_k w_k$, is given by $\hat{\theta}_w \pm zW^{-1/2}$.

For the fixed effects model (FEM) it is assumed all $\theta_k = \theta$, say. It is widely recognized

that this assumption is an over-simplification of reality, but nevertheless an analysis for it is usually given for the sake of comparison with a random effects model analysis, to be discussed shortly. For the FEM, $\theta_w = \theta$ for any set of weights. The weights that minimize the standard error of $\hat{\theta}_w$ for estimating θ are the inverse variance weights $w_k = n_k/v_k$. So the 'conventional' meta-analysis now estimates θ by $\hat{\theta}_{\hat{w}} = \sum_k \hat{w}_k \hat{\theta}_k / \hat{W}$, where $\hat{w}_k$ is the consistent estimator of w_k in the kth study. Figure 1 is an example of a meta-analysis of seven two-armed binomial studies. The effect is the difference of the two probabilities. The data is from Fleiss (1993) and concerns the use of aspirin to prevent death following a myocardial infarction. The plot is a common technique for visualizing the results of a meta-analysis.

The variance of $\hat{\theta}_{\hat{w}}$ is no longer W^{-1}, because the weights are estimated. In fact, W^{-1} underestimates this variance (Li et al., 1994; Rukhin, 2009), so that the coverage of the conventional interval $\hat{\theta}_{\hat{w}} \pm z_{1-\alpha/2}\hat{W}^{-1/2}$ is lower than the nominal level $1-\alpha$, and the conventional Wald test, $\hat{W}^{1/2}\hat{\theta}_{\hat{w}}$, for $H_0 : \theta_w = 0$ is too liberal. For example, for normally distributed data with $n_i = 3$ for each study, the variance of $\hat{\theta}_{\hat{w}}$ is K times larger than W^{-1} (Rukhin, 2009)!

Another shortcoming of the variance estimation by W^{-1} is its over-sensitivity to the minimum of the estimates of the variances in the K studies, which follows from the inequality $W^{-1} \leq \min(v_k, k = 1, \cdots, K)$ (Li et al., 1994).

An approximate unbiased variance of $\hat{\theta}_{\hat{w}}$ up to any order was obtained for the normal model by Sinha (1985). There are no similar results in the general case.

The maximum likelihood estimator (MLE) of θ does not have a closed form, although it is a weighted means statistic with the weights inversely proportional to the MLEs of the within-study variances. Several tests for the common mean, including a test based on the first-order approximate variance by Sinha (1985) and on the MLE combined with a parametric bootstrap procedure were studied by simulation in Chang and Pal (2008) for $K = 2$ and $K = 5$. Both tests performed well, with the MLE based test being somewhat more powerful.

It can be seen that proper statistical inference under the FEM is difficult even under normality; it becomes even more so in the general case. The problem is exacerbated by possible correlations between the effects and the weights. Higher order inference needs to be developed for many non-normal effects used in meta-analysis.

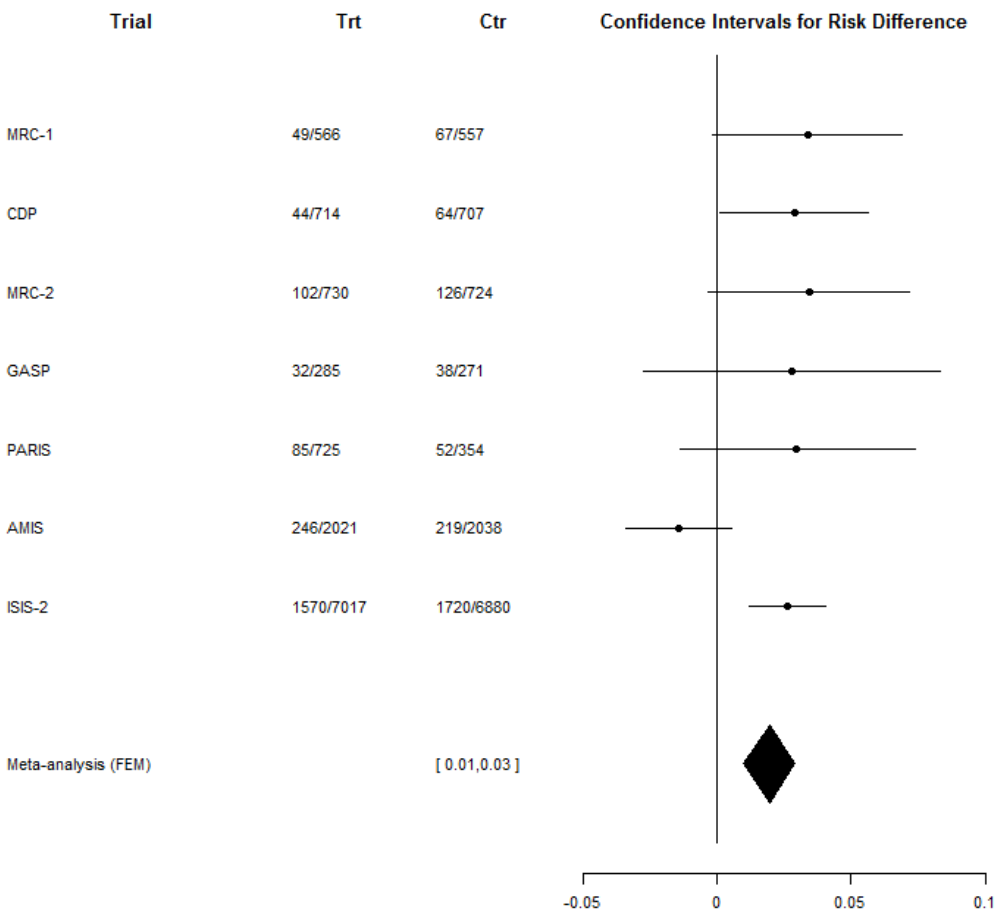

Figure 1: *A forest plot of seven studies involving two binomial arms. For each trial, the summary of the data and the confidence interval for the difference of the probabilities is shown. The first five studies are too small to show a significant effect. The bigger studies give contradictory outcomes. The final meta interval based on the FEM is to a large extend determined by the largest study.*

2.2 Random effects model

For the random effects model (REM) it is assumed that the $\theta_1, \ldots, \theta_K$ are a random sample from a $\mathbb{N}(\theta, \tau^2)$ distribution, with both parameters unknown. When combined with the assumption that $\hat{\theta}_k|\theta_k \sim \mathbb{N}(\theta_k, v_k/n_k)$, this leads to the unconditional distribution for the estimated effect $\hat{\theta}_k^* \sim \mathbb{N}(\theta, \tau^2 + v_k/n_k)$. Note that we have changed notation from $\hat{\theta}_k$ to $\hat{\theta}_k^*$ to reflect the change in distribution. There are now $K + 2$ unknown parameters: the overall or representative effect θ, the inter-study variance τ^2, and the unknown v_k, $k = 1, \ldots, K$. There are two important issues here, how to interpret θ, and how to estimate it. The

main advantage of the REM over the FEM is that because the effects themselves are a random sample with mean θ, an estimate of the latter could provide insight into the larger family of studies which might be carried out under similar conditions. At its most extreme interpretation, there is an urn of possible studies, and the ones that have been selected have randomly chosen sample sizes and within-study variances as well as effects, see Shuster (2010); Buonaccorsi (2006). The choice of the random urn model greatly affects statistical inference under REM. The conventional estimator $\hat{\theta}^*_{\hat{w}}$ is biased, and the unweighted mean is recommended in Shuster (2010). However, a more modest interpretation is that the REM is chosen for mathematical convenience, and allowing for only one more (nuisance) parameter τ^2. An alternative, multiplicative model for REM based on overdispersion due to within-studies correlations was recently proposed by Kulinskaya and Olkin (2013). The real advantage of detecting a positive τ^2 is that it suggests finding explanatory variables (or moderators) which explain the variation between study outcomes, as in meta-regression, see Section 2.5. For a recent review of various issues involving the REM, the reader is referred to Higgins et al. (2009). See also Sutton and Higgins (2008) for a more general discussion of a variety of meta-analytic methods.

Statistical methods for the REM just described has been the subject of much research (see Section 2.4.), especially with regard to the nefarious τ^2. This parameter cannot be estimated to any reasonable precision without a large number K of studies, and hence neither can θ. The traditional approach of substituting an estimator $\hat{\tau}^2$ for τ^2 into the asymptotic variance formula $\mathrm{Var}[\hat{\theta}^*_k] = \tau^2 + v_k/n_k$ and applying the estimated inverse variance weights combination suffers from all the problems of this methodology for the FEM, further aggravated by neglecting the variability of $\hat{\tau}^2$. It has been applied to thousands of data sets, despite the lack of theoretical or simulation studies to confirm when it does work, if ever.

If a prior distribution for θ and τ^2 is chosen, the REM becomes a Bayesian hierarchical model. Much has recently been written about Bayesian methods in statistical modelling and meta-analysis is a popular area of application. A very readable introduction to the methods can be found in Koricheva et al. (2013). Bayesian methods are particularly popular when performing network meta-analyses, in which studies comparing several treatments are combined. Both direct comparisons between two treatments used together in a trial as well as indirect comparisons of the two treatments used in different studies are combined.

Xie et al. (2011) propose a frequentist method that relies on summarizing the results of the meta-analysis in the form of a confidence distribution. This is related to the Bayesian approach using non-informative priors.

2.3 Testing for homogeneity

In a meta-analysis it is usual to conduct a homogeneity test to determine if the effects measured by the included studies are sufficiently similar. For the FEM, the homogeneity test is a test of the hypothesis that the underlying effects are all equal, $\theta_k = \theta$. For the REM, this is the test of the null hypothesis $H_0 : \tau^2 = 0$.

The most commonly used test statistic is Cochran's Q (Cochran, 1937) which is defined by $Q = \sum_k \hat{w}_k(\hat{\theta}_k - \hat{\theta}_{\hat{w}})^2$ with the inverse variance weights $\hat{w}_k$. Several other tests were compared by Takkouche et al. (1999) and Viechtbauer (2007), but it was concluded that Q is the best choice. A number of results have been published about the distribution of Q for the case in which the effects are normally distributed sample means and the weights are inverses of sample variances. Under these normality assumptions, the chi-square distribution is an exact distribution of the Q statistic if the variances are assumed to be *known*, resulting in known weights. We denote this statistic with known weights by Q_w, to distinguish it from $Q = Q_{\hat{w}}$. Since the randomness of the weights is traditionally ignored in meta analysis, a number of publications provided further distributional results for Q_w, among them Biggerstaff and Tweedie (1997), Jackson (2006), Biggerstaff and Jackson (2008).

For estimated weights, there is no exact analytic expression for the distribution of Q, and so an approximation must be used in order to conduct the homogeneity test. Further, the distribution of Q will vary depending on the effect measures. The chi-square distribution is asymptotically valid as the sizes n_k of the studies become large, but the approximation is less accurate for small and medium sample sizes, see the simulation studies by Hedges and Olkin (1985), Viechtbauer (2007) and the references therein. James (1951) and Welch (1951) proposed separate order $O(1/n_k)$ corrections to the null distribution of Q for the normal case. Welch's proposal (more commonly used and known as the Welch test) refers Q to a rescaled F-distribution ($cF_{I-1,\nu}$) with $I - 1$ and ν degrees of freedom where ν and c are estimated from the data. Kulinskaya et al. (2003) dealt with the improved approximation under alternatives for the FEM.

Kulinskaya et al. (2011b,a) found $O(1/N)$ improvements to the chi-square approximation to the null distribution for Q applicable to non-normal effect measures, where $N = \sum_k n_k$ is the total sample size. Kulinskaya et al. (2011b) dealt with the situation in which both the effect and the weight from an individual study depend on a single parameter, with principal application to the standardized mean difference between treatment and control arms of a study. Kulinskaya et al. (2011a) provides expansions for the first two moments of Q when the effect and weight for an individual study depends on two parameters, the effect θ_k and a nuisance parameter ζ_k. These expansions were applied to the difference of binomial probabilities (risks) from treatment and control arms of the studies. In this context, a two-moment gamma approximation was recommended as an approximate null distribution of Q. The resulting homogeneity test is substantially more accurate than the standard chi-square test, especially when the sample sizes are small or moderate.

An asymptotic distribution of Q_w for non-normal effects when the sample sizes are finite and $K \to \infty$ is discussed in Demidenko (2004, Section 5.1.3). This distribution is asymptotically $\mathbb{N}(K, (\kappa - 1)K)$, where κ is the kurtosis of the underlying distribution.

Asymptotics in the case of $K \to \infty$ is discussed by Akritas and Papadatos (2004) both for Q and for a new unweighted statistic T_K which can be used with small sample sizes. They do not require normality, but some standard moment conditions, and demonstrate that an asymptotic approximation to the distribution of Q is possible only if the within-study sizes $n_k \to \infty$ suitably fast in relation to K. In that case $Q_{\hat{w}}$ is asymptotically equivalent to Q_w and $K^{-1/2}(Q_{\hat{w}} - (K-1)) \sim \mathbb{N}(0, 2)$ under the null. They also show that Q is very unstable for small sample sizes. The unweighted statistic proposed by Akritas and Papadatos (2004) is given by

$$T_K = K^{-1/2} \sum_{k=1}^{K} \left[n_k (\bar{X}_{k\cdot} - \bar{X}_{\cdot\cdot})^2 - (1 - \frac{n_k}{N}) S_k^2 \right],$$

where X_{ki} is the i-th observation in the k-th study and S_k^2 are sample variances. This statistic, equivalent to the standard F-statistic in the case of equal study sizes, is yet to be tried in meta-analytic applications. Akritas and Papadatos (2004) derive its distribution under local alternatives in FEM. The rates that local alternatives must converge to the null are $K^{-1/4}$ for bounded study sizes and $K^{-1/4} n_k^{-1/2}$ for $n_k \to \infty$. These rates resemble those for lack-of-fit testing in nonparametric regression.

There are so far no results on the distribution of Q or T_K for the REM. These results

would be important for derivation of the distribution of $\hat{\tau}$.

There is no consensus on the appropriateness of conducting a homogeneity test. If the choice of model in further analysis depends on the Q test as recommended by Normand (1999) and as is often done in applications, then the two-stage procedure is in use and the level needs to be adjusted accordingly (Hartung and Knapp, 2003). Another view is that the heterogeneity should not be tested but quantified by some effect measure (Higgins et al., 2009), the most popular being the $I^2 = (Q - (K-1))/Q$ proposed by Higgins and Thompson (2002). I^2 is an increasing function of Q, which means that the statistical properties of I^2 can be deduced from those of Q. If there is no additional variance, the expectation of Q is approximately equal to $K-1$ and I^2 will be close to zero. If the alternative is true, Q grows with the total sample size $N = \sum_k n_k$ and I^2 tends to 1 as $N \to \infty$, unless the number of studies K grows with N in such a way that the average study size N/K remains bounded. Appropriately standardized effect measures for heterogeneity are yet to be derived. The paper by Demidenko et al. (2012) takes a step in this direction.

2.4 Inference for the REM

There are several statistical problems in REM, to do with inference for θ and for τ. Usually, the former is of primary interest, but the variance of $\hat{\theta}$ depends on τ, so the latter cannot be easily bypassed.

The most popular estimator of τ is the moment estimator of DerSimonian and Laird (1986)

$$\tau^2_{DL} = \max\big(0, \frac{Q-(n-1)}{S_1 - S_2/S_1}\big),$$

where $Q = \sum_1^K w_k(\hat{\theta}_k - \bar{\theta})^2$ is the Q statistic (Cochran, 1937) and $S_r = \sum_1^K w_i^r$. Instead, the MLE of τ^2, or the restricted MLE (REML) can be used. Yet another possibility is to use profile likelihood (Hardy and Thompson, 1996; Malloy et al., 2013). All these methods require numerical maximization. An easier option is to use the Mandel-Paule (MP) algorithm (Mandel and Paule, 1970). Given that the weights $w_k = w_k(y) = (y + \hat{v}_k/n_k)^{-1}$, the Mandel-Paule estimator is found from the estimating equation

$$F(y) = \sum_1^K w_k(y)(\hat{\theta}_k - \bar{\theta}_w)^2 = K - 1\,.$$

The motivation comes from the fact that, for the true weights, $F(y) \sim \chi^2_{K-1}$ and therefore the first moment is $K-1$. The solution is unique because $F(y)$ is the a convex monotonically

decreasing function of $y \geq 0$ (Rukhin, 2009). The MP algorithm was initially introduced in the context of inter-laboratory studies, but was then adopted for meta-analysis by Rukhin (2003) and DerSimonian and Kacker (2007). In the modified Mandel-Paule procedure (MMP) $K-1$ is replaced by K in the right-hand side of the above estimating equation. As was shown in Rukhin and Vangel (1998), under normality, the MLE of τ^2 coincides with the MMP estimate if the weights w_k admit the representation $w_k = w_k(y)$. The original MP is similarly related to REML (Rukhin et al., 2000). The MP algorithm is also a generalized Bayes procedure (Rukhin et al., 2000). Unfortunately, there is no uniformly MSE optimal estimator of τ^2 over the whole range of τ^2, even if the distribution is normal. Under normality, improved quadratic estimators of the random variance component τ^2 and within-study variances σ_k^2 in the spirit of Stein are given in Mathew et al. (2010).

Given an estimate $\hat{\tau}^2$, the weights $w_k^* = 1/(\hat{\tau}^2 + \hat{v}_k/n_k)$ can be used to obtain the combined effect estimate $\hat{\theta}_{w^*}$. Inference for $\hat{\theta}_{w^*}$ has the same problems as the inference in FEM, additionally the variability in $\hat{\tau}^2$ is often neglected. The coverage of the conventional confidence intervals is considerably below nominal. A further problem with these random effects confidence intervals is that they can be very sensitive to publication bias. If smaller studies (with larger variances) are less likely to be published than larger studies (with smaller variances), then the coverage probabilities of these confidence intervals can rapidly decrease as the degree of heterogeneity, or the number of studies or both, becomes large (Henmi and Copas, 2010).

An alternative to the standard inverse weights method accounts for the variability in τ by replacing the approximating normal distribution for the weighted effect by a Student-t distribution; theoretical and simulation studies justifying this approach are in Hartung (1999), Hartung and Makambi (2003), Sidik and Jonkman (2002, 2003, 2006, 2007).

Hardy and Thompson (1996) obtained the ML confidence interval for θ by inverting the likelihood ratio test for θ combined with the profile likelihood estimate of τ^2. However, Sørensen (2008) showed that, for small values of K, the distribution of the likelihood ratio test, and therefore related p-values and confidence intervals, strongly depend on the true value of τ^2. A higher order asymptotic procedure was recently developed by Sharma and Mathew (2011), but the regularity conditions for its applicability were not established. Henmi and Copas (2010) proposed a new interval centered at the FEM combined effect $\hat{\theta}_{\hat{w}}$ and based on the gamma-approximation to the conditional distribution of Q given

$\hat{\theta}_{\hat{w}}$. The coverages of the intervals by Sidik and Jonkman (2002); Hardy and Thompson (1996); Sharma and Mathew (2011); Henmi and Copas (2010) are considerably better than the coverage of conventional intervals for small K (starting from $K = 5$). Additionally the interval by Henmi and Copas (2010) is designed to perform well under publication bias. All these methods assume that normality of estimated effects is already reached and therefore require considerable within-study sample sizes. As an example, Henmi and Copas (2010) simulated the odds ratios (their effect size of interest) from the normal distribution, and therefore have no information on the within-studies sample sizes required for good coverage. Our experience is that for large K, larger sample sizes are often required for asymptotic results to hold. Applicability of the above methods to various effect measures should be explored by extensive simulations. It may be necessary to develop second-order asymptotic methods for non-normal effect measures such as odds ratios or standardized mean differences.

A recent addition to this area by Rukhin (2013) proposes new estimators of the between-study variability which are linear functions of quadratic statistics $(X_i - X_j)^2$ and the sample variances. This class includes the τ^2_{DL} estimator (DerSimonian and Laird, 1986) among others. The proposed estimators perform well in the case of small to moderate numbers of studies. An important conclusion is that different estimators of τ^2 should be used depending on the inferential task: "one to minimize the variance of the treatment effect statistic; another to construct a reliable confidence interval for this parameter; yet another to estimate τ^2 itself! "

Another approach by Malloy et al. (2013) for random effects having the Student-t distribution uses variance stabilization before finding maximum likelihood and profile estimates of both θ and τ^2; this approach yields explicit formulae for the standard errors of each estimator in terms of the number of studies K. In addition, these authors show that the best performing confidence interval for δ is a simple t-interval which does not require an estimate of τ^2. Further, both traditional and new methods for the FEM are shown to be robust to small $\tau^2 > 0$, so that if K is too small to estimate τ^2, a practical solution is to revert to the FEM, even though one might prefer the REM. These results should be able to be extended to the standardized mean difference, see Malloy et al. (2013, Section 1.4).

2.5 Meta-regression

Let $\hat{\theta}_k$ denote the estimated effect for the kth study which is based on n_k observations, and suppose that these estimates satisfy

$$\hat{\theta}_k = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_k + \varepsilon_k \tag{2.1}$$

where $\boldsymbol{x}_k = [x_{k1}, \ldots, x_{kp}]'$ is a vector of study covariates, called *moderators* in the meta-analytic literature, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]$ is a vector of unknown coefficients, and $\varepsilon_k \sim N(0, v_k/n_k)$, where v_k are known constants. This is the *fixed effects meta-regression model*. Tests for non-zero coefficients can be found using traditional weighted least squares. Examples are given in Hedges and Olkin (1993), where variance stabilization of the effect estimates precedes weighted least squares in order to obtain $\hat{\theta}_k$'s (at least approximately) satisfying the normality assumption with known variances. When the variances v_k/n_k are unknown, it is tempting to use weighted least squares with estimated variances $\hat{v}_k/n_k$, but this can lead to highly biased estimates of the coefficients, as demonstrated in Malloy et al. (2011), where it is compared with a generalized linear model approach.

It is unlikely that all moderators would be identified in advance, so that there would be unknown heterogeneity in studies that should be accounted for. A *mixed effects meta-regression model* is then appropriate. The simplest version of this model, the one where only the intercept is random, was introduced by Colditz et al. (1994); Berkey et al. (1995), and assumes that

$$\hat{\theta}_k = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_k + b_k + \varepsilon_k\,, \tag{2.2}$$

where the b_k's are assumed to be independent of each other, the moderators and of the ε_k's; further they are assumed to be distributed $b_k \sim N(0, \tau^2)$, where τ^2 is the unknown inter-study variance. Berkey et al. (1995) did assume that the v_k/n_k's were approximately equal to within study estimates $\hat{v}_k/n_k$, and proceeded to use weighted least squares with inverse variance weights $1/\{\hat{\tau}^2 + \hat{v}_k/n_k\}$, where $\hat{\tau}^2$ is obtained by an iterative procedure. Little progress has been made since then in dealing with two outstanding problems, namely the usually unwarranted assumptions of normality of effect size estimates and of known within-study variances, see Huizenga et al. (2011) for a thorough discussion.

There are at least two promising avenues for further research in mixed effects meta-regression. First, maximum likelihood estimation of the parameters $\boldsymbol{\beta}$ and τ^2 using the full likelihood, not the conditional one based on a false assumptions of known variances

or normality where they are clearly not satisfied. Second, parameter estimation using generalized linear models after variance stabilization, with the then reasonable assumption of approximate normality.

2.6 Individual participant data versus summary statistics

Although considered the gold standard, individual participant data (IPD) and the expertise required for its analysis are rarely both available, and IPD meta analysis is used only in a small minority of the MAs. 9.5% of the publications between 1990 to 2004 according to Simmonds et al. (2005).

Aggregated (summary) data (AD) may be a matter of analytic choice or be necessitated by lack of the original data or lack of permission to use it, possibly because of ethical/confidentiality issues.

For the most common case of studies of a continuous outcome with two arms (treatment and control), Mathew and Nordstroöm (1999) have shown that AD analysis results under FEM are equivalent to those based on IPD regardless of the covariance structure within the studies. More recently, the same authors Mathew and Nordstroöm (2010) considered the weighted estimation of a linear function of the mean, based on linear models for summary data and for IPD. Within-studies covariances were assumed known. They derived a condition under which the IPD and AD meta-analysis estimators coincide. This condition always holds for the FEM, and for the REM it holds when the proportion in the treatment arm is the same across studies. They also show that when covariates are present, the two estimators coincide only under an extra simplifying assumption that represents homogeneity of the covariates across studies. When the condition is not satisfied, the one-step analysis is always more efficient. The results by Mathew and Nordstroöm (2010) are not valid when within-trial variances and covariances are not known but estimated from the data. An example of this is shown in Jones et al. (2009), where the AD meta-analysis estimates have a very slightly smaller standard error than the IPD meta-analysis estimates.

Lin and Zeng (2010) studied the relative efficiency of AD versus IPD meta-analysis in the multivariate FEM with general, not necessarily the same, distributions across studies. They show that, for all commonly used parametric and semi-parametric models, there is no asymptotic efficiency gain by analyzing the original data if the parameter of main interest has a common value across studies, the nuisance parameters have distinct values

among studies, and the summary statistics are based on maximum likelihood. They also consider the case of estimated within-study covariance matrices, and then the equivalence holds asymptotically. They also studied the relative efficiency of the two methods when the parameter of main interest has different values among studies, i.e. under the UFEM (unequal but fixed effects model), or when there are common nuisance parameters across studies. Their examples include the linear model, logistic and Cox regression.

Lin and Zeng (2010) comment that one reason for obtaining original data is to model individual-level covariates. They have shown that there is no bias or efficiency loss if the effect estimates are properly adjusted for individual-level covariates within each study and then combined. They also claim that their results hold also for the REM, but these results are unpublished.

2.7 Meta-analysis of observational studies

Randomized controlled trials (RCTs) are considered the gold standard in medicine and health sciences, and restriction of meta-analyses to synthesis of their findings is strongly supported by the influential Cochrane Collaboration (Higgins and Green, 2011). The main advantage of randomisation of allocation of participants into control and treatment arms of a study is that the difference between the summary effects in the two arms is an unbiased estimate of the true effect. Randomization takes care of the various observed and unobserved prognostic factors given that their differences between the arms are truly random. No further statistical modelling is necessary. In real life, randomization may not be practicable or ethical, and observational studies (experimental or not) constitute the bulk of the body of evidence in epidemiology, life and social sciences (Stroup et al., 2000; Konnerup and Kongsted, 2009). Hence a large and fast growing number of meta-analyses of observation studies. The standards of reporting and main issues in the meta-analysis of observational studies in epidemiology are summarized in the MOOSE consensus statement, Stroup et al. (2000, for the meta-analysis of observational studies in epidemiology (MOOSE) group). The authors recommend in-depth investigation of issues to do with heterogeneity of populations, designs and outcomes across studies, and formal evaluation of the study quality. Reeves and Wells (2013) is a recent special issue covering the topic of the use of non-randomized studies in systematic reviews.

The main concern in meta-analysis of observational studies is the existence of various

biases due to non-random allocation. The main danger is the existence of a systematic selection bias due to systematic differences between the groups at the baseline (Deeks et al., 2003). These baseline differences may be related to outcomes; differences in severity of a condition of interest as the reason for treatment allocation is an example of such a relationship. This introduces systematic bias leading to over- or under-estimation of treatment effects. Other biases in non-randomized studies include attrition bias (due to drop-out or non-compliance), detection bias (because of non-standardized assessment of outcomes) and performance bias (errors and inconsistencies in the allocation, application and recording of interventions) (Deeks et al., 2003).

Selection bias can potentially be adjusted for by statistical modelling. This includes modelling the effects of interest controlling for known prognostic factors, or modelling the allocation mechanism itself and then using the results for the effect adjustment (propensity score analysis). The empirical modelling performed in Deeks et al. (2003) showed inadequacy of both approaches; the main reason is the omission of important unknown confounders.

Due to the high risk of biases and typically large sample sizes in observational studies, meta-analysis may reach spurious conclusions with very tight confidence intervals around biased summary effects. Different biases can be modelled explicitly (Wolpert and Mengersen, 2004), but this requires some strong assumptions about the nature of biases. An alternative approach is the elicitation of expert opinions on the nature and size of biases. In a recent paper by Thompson et al. (2011) the biases are separated into internal biases reflecting the study quality and external biases reflecting generalisability to a target setting. Next, subjective opinions on the size of biases and their type (additive vs multiplicative) are elicited from several experts and are used for bias adjustment.

Perhaps a safer option is to consider the robustness of the combined effects to different levels of bias in point estimates or their variances. An interesting proposal by Salanti and Ioannidis (2009) is to assume a limit to the chance that an effect is in a particular direction and not in the other one. They call this the *credibility ceiling*. For the kth study with the effect θ_k and variance v_k, consider a random variable $u \sim N(\theta_k, v_k)$. Then the credibility is defined as $P(u < 0|\theta_k > 0)$ or $P(u > 0|\theta_k < 0)$. If this probability is less than the credibility limit c, the variance is recalculated as $v_k^* = \max\{(\theta_k/z_c)^2, v_k\}$, where z_c is the percentage point from the standard normal distribution. Meta-analysis is repeated for a

range of values of c, using these inflated variances. Various issues to do with biases in meta-analysis are considered in more detail in Section 3.

Finally, it is possible to design an unbiased observational study. Available methods include matching, natural experiments and the use of instrumental variables; see Konnerup and Kongsted (2009, Section 4) for more details.

3 Publication and other biases

As mentioned in the introduction, bias is the most immediate danger when combining the estimated effects from K trials or studies. In this section we touch on two sources of such bias. First, the selection bias resulting from the inability of accessing studies with negative outcomes and, second, the effect of small sample biases on a meta-analysis.

3.1 Publication bias

Publication bias refers to the case in which the K studies we are given access to for our meta-analysis are those with a relatively large estimated effect. Other studies, with smaller effects may, however, also exist. This is a special case of selection bias. We are handed K estimated effects and want to combine them to produce an overall estimate. If the large estimated effects are over-sampled, then the naive, uncorrected combination overestimates the true effect.

Example 3.1. *For a formal discussion of this bias, we will consider the case, where the effect estimates in each trial are variance stabilized. To distinguish these estimates from the more general statistics $\hat{\theta}_k$, we denote them by $\hat{\kappa}_k$ and assume that they have variance equal to $1/n_k$, that is, the test statistic $\sqrt{n_k}\hat{\kappa}_k$ has constant variance 1. With a fixed common effect κ we then have approximately $\hat{\kappa}_k \sim \mathbb{N}(\kappa, 1/n_k)$. The combined estimated effect from K randomly selected studies is $\hat{\kappa}_{meta} = \sum_{k=1}^{K} n_k\hat{\kappa}_k/N \sim \mathbb{N}(\kappa, 1/N)$, where $N = \sum_{k=1}^{K} n_k$ is the total sample size. This represents the ideal situation, where the meta-analysis produces an estimated effect which is equal to what one would get if all the data were combined and a single trial of size N were performed. Since the weights depend on the known n_k, variance stabilization leads to more interpretable and more stable meta-analysis procedures when compared to the usual fixed common effects estimates. For details, see Kulinskaya et al. (2008).*

If only the studies which reach a minimal level of significance are entered into the meta-analysis, we no longer get to work with a random sample of K studies and a selection bias is created. Because those studies resulting in insignificant effects are filed away and forgotten, selection based on significance is sometimes referred to as the file drawer effect.

A well-written account of selection bias and its consequences can be found in an article by J. Lehrer that appeared in the New Yorker magazine (Lehrer, 2010). He points out that selection bias, if ignored, can have costly consequences both in terms of ill-advised research activity and in practical terms, by influencing, for example, medical decision making. Selection bias occurs not only because the results contained in a submitted article are negative, but also due to the reviewers likes and dislikes, the withholding of data for commercial purposes, the choice of design of the clinical trial, and so on. Completely different sources of bias exist as well. As mentioned by Lehrer, researchers themselves may be prone to reach firmer conclusions than warranted due to preconceived notions and subtle biases.

3.1.1 Assessing publication bias

If the number of studies K entering a meta-analysis is large, there is a reasonable hope in determining whether a selection as described in the above example has occured and possibly even in estimating the number of missing studies.

Example 3.2. *We continue with the example above in order to determine the consequences of selecting only certain studies. Suppose that the true (fixed) effect is κ and that the null hypothesis being tested is κ_0. If only the studies which reach significance are selected, only the studies with $\sqrt{n_k}(\hat{\kappa}_k - \kappa_0) > b = z_{1-\alpha}$ are included in the meta-analysis. In our model, $\sqrt{n_k}\hat{\kappa}_k \sim \mathbb{N}(\sqrt{n_k}\kappa, 1)$, which implies that the probability for a study of size n to pass the selection is $1 - \Phi(b + \sqrt{n}(\kappa_0 - \kappa))$. This does not depend on n if $\kappa_0 = \kappa$, that is, the null hypothesis being tested is true. The selection favors small studies if $\kappa_0 > \kappa$ and large studies otherwise. Being selected means that $\sqrt{n_k}\hat{\kappa}_k$ has a truncated normal density equal to*

$$f(x) = \frac{\varphi(x - \sqrt{n_k}\kappa)}{1 - \Phi(b + \sqrt{n_k}(\kappa_0 - \kappa))} \text{ for } x > b + \sqrt{n_k}\kappa_0\,,$$

where Φ, φ denote the standard normal distribution and density, respectively.

If a non-random selection has produced the studies available in a meta-analysis, similar to the above example, then for large K the histogram of the standardized effect estimates or of the raw effect estimates would show the missing parts. If K is small, the task of assessing missingness is essentially impossible, unless additional information is provided through other means, for example if all the studies had to be announced via a database.

To make informal inferences about missingness, the funnel plot is popular. This graphical display was introduced in Light and Pillemer (1984) and has been reviewed in Egger et al. (1997). The funnel plot replaces the histograms mentioned above and brings into consideration also the trial sizes or the precision of the effect estimates. For K studies with effect estimates $\hat{\kappa}_k$ and sizes n_k, one could produce

- either a scatter plot of n_k vs. $\hat{\kappa}_k$ for $k = 1, \ldots, K$,
- or a scatter plot of $-1/\sqrt{n_k}$ vs. $\hat{\kappa}_k$ for $k = 1, \ldots, K$, or
- or a scatter plot of $-\sqrt{\hat{v}_k}$ vs. $\hat{\kappa}_k$ for $k = 1, \ldots, K$, with $-\hat{v}_k$ being the estimated variance of $\hat{\kappa}_k$.

If the studies are randomly selected, the point scatter will exhibit a funnel shape, with the large studies tightly clustered around the common effect and the smaller studies showing more variation. Publication bias as discussed in the example will lead to holes or asymmetry in the funnel plot. Duval and Tweedie (2000b) and Duval and Tweedie (2000a) describe a relatively simple and intuitive method based on the expected symmetry of the funnel plot for estimating the number of missing studies and for creating imputed estimated effects and standard errors for the missing studies. Their method is based on the idea that the missing studies are missing because the effects found in the studies were too small.

Of course, the asymmetry in the point cloud of a funnel plot can have other causes. If, for example, a single large study has been performed and it contradicts a host of previous smaller studies, the funnel plot can take on an unexpected shape.

3.1.2 Correcting the bias

In order to estimate the bias caused by selection bias, various models can be used. The statistical literature on estimation based on samples subject to selection bias focuses on procedures like the EM-algorithm or data augmentation, censoring and truncation as well as various biasing mechanisms. An example of a fairly general estimation aim and a fairly

general selection mechanism is Vardi (1985). An example that is more specific to meta-analysis is Givens et al. (1997).

Example 3.3. *This example is again based on normally distributed variance stabilized effect estimates with variance* $1/n_k$*. To make the calculations easier, suppose that the true value of the fixed effect is* $\kappa = 0$ *and that the null hypothesis being tested is also* $\kappa_0 = 0$*, so that the null hypothesis being tested is true. Suppose further, that only studies with* $\sqrt{n_k}\hat{\kappa}_k > b$ *are included in the meta-analysis. In our model,* $\sqrt{n_k}\hat{\kappa}_k \sim \mathbb{N}(0,1)$*, which implies that a fixed fraction of all studies will be selected and that this fraction is equal to* $1 - \Phi(b)$*. The combined effect* $\hat{\kappa}_{meta} = \sum_{k=1}^{K} n_k\hat{\kappa}_k/N$ *based on* K *such studies, is not any longer distributed as* $\mathbb{N}(\kappa, 1/N)$*. Being selected means that* $\sqrt{n_k}\hat{\kappa}_k$ *has a truncated normal density*

$$f(x) = \frac{\varphi(x)}{\int_b^\infty \varphi(u)\,du} = \frac{\varphi(x)}{1-\Phi(b)}\,.$$

An elementary calculation shows that the expectation and variance of any selected $\sqrt{n_k}\hat{\kappa}_k$ *are* $0 < \mu(b) = \varphi(b)/(1-\Phi(b))$ *and* $\sigma^2(b) = 1 - \mu(b)(\mu(b) - b) < 1$*. For the combined effect* $\hat{\kappa}_{meta}$*, this implies an expectation of* $0 < \sum_{k=1}^{K} \sqrt{n_k}\mu(b)/N$ *and a variance of* $\sigma^2(b)/N < 1/N$*. Were we to ignore the bias, we would often observe a significant effect, because in order to compute a p-value, we would compare the combined value* $\sqrt{N}\hat{\kappa}_{meta}$ *with a standard normal distribution. In reality, though, the mean of this random variable is equal to* $\mu(b)\sum_{k=1}^{K} \sqrt{n_k}\mu(b)/\sqrt{\sum_{k=1}^{K} n_k} \geq \mu(b)\sqrt{K}$*, where the inequality follows from the concavity of the square root. As the number of studies grows and if* b *is not too small, it becomes very likely that the combined effect is judged to be significantly different from* $\kappa = 0$ *and one would falsely declare the discovery of a significant positive effect.*

If the selection of available studies is not based on significance, but rather on the observed raw effect, then the chance of being included in the meta-analysis depends on the sample size. Suppose only those studies with $\hat{\kappa}_k > b$ *are selected. Under our model, the probability of selection is then equal to* $\mathbb{P}[\hat{\kappa}_k > b] = 1 - \Phi(\sqrt{n_k}b)$*, which means that the selection of a trial for inclusion in a meta-analysis is correlated with the trial size. If* $b > 0$*, large studies are often selected out and do not make it into a meta-analysis, while for* $b < 0$ *all the studies have a good chance of being selected. This type of selection bias leads to more complex properties of* $\hat{\kappa}_{meta}$ *than in the previous case. Given that a trial of size* n_k *has passed selection, the cumulative distribution of* $\hat{\kappa}_k$ *is* $F_k(x) = (\Phi(\sqrt{n_k}x) - \Phi(\sqrt{n_k}b))/(1-\Phi(\sqrt{n_k}b))$ *which has density* $f_k(x) = \varphi(\sqrt{n_k}x)\sqrt{n_k}/(1-\Phi(\sqrt{n_k}b))$*. It follows*

that $\mathbb{E}[\hat{\kappa}_k] = (\sqrt{n_k}\varphi(\sqrt{n_k}b)/[(1-\Phi(\sqrt{n_k}b))] \approx b$*, where the approximation holds for* $b > 0$ *and large* n_k *and follows from the asymptotic equivalence* $1 - \Phi(x) \sim \varphi(x)/x$ *as* $x \to \infty$*. The expected value of* $\hat{\kappa}_{meta}$ *cannot be computed without more knowledge about the choice of* n_k*.*

Shuster (2010) has put forward the idea that the correlation between sample size and effect size is a common phenomenon in meta-analysis. Any weighted estimate using weights that depend on the sample size then causes a bias. However, in the absence of selection bias, the arithmetic mean of the trial effects, $\hat{\kappa}_{meta} = \sum_{k=1}^{K} \hat{\kappa}_k/K$*, remains an unbiased and, in this sense, viable estimate.*

If an effect estimate $\hat{\kappa}$ is included in a meta-analysis only if $\sqrt{n}\hat{\kappa} > b$, one can model the $\sqrt{n_k}\hat{\kappa}_k$ included in the analysis as a left-censored sample with a fixed censoring bound b. By stipulating values for the number of censored studies, M, the bound, b, and the sample sizes of the missing studies n and then investigating what biasing effect such a scenario would have on the meta-analysis, we could gain a better understanding of the bias. This is easy to make precise in the case of variance stabilized effect estimates with (approximate) normal distribution. In this case, we have $\hat{\kappa}_k \sim \mathbb{N}(\kappa, 1/n_k)$, where κ is the fixed common effect.

Suppose K studies are available for a meta-analysis and M studies are missing (or rather censored!) and suppose that the selection is based on significance when testing $H_0 : \kappa = \kappa_0$, which means that a trial is censored when $(\hat{\kappa} - \kappa_0)\sqrt{n} \leq b$, that is, when $\hat{\kappa} \leq b/\sqrt{n} + \kappa_0$. The likelihood function for the unknown common effect κ is then equal to

$$L(\kappa) = \prod_{k=1}^{K} \varphi\left((\hat{\kappa}_k - \kappa)\sqrt{n_k}\right) \quad \prod_{m=1}^{M} \Phi\left(b - \sqrt{n_m}(\kappa - \kappa_0)\right) . \tag{3.1}$$

In order to compute this likelihood, we thus need to know not only M and the censoring mechanism (that is, b), but also the trial sizes of the missing studies $n_1, \ldots, n_M$. The log-likelihood is

$$\log(L(\kappa)) = \ell(\kappa) = -(K/2)\log(2\pi) - \sum_{k=1}^{K} -\frac{n_k}{2}(\hat{\kappa}_k - \kappa)^2 + \sum_{m=1}^{M} \log\left(\Phi\left(b - \sqrt{n_m}(\kappa - \kappa_0)\right)\right) . \tag{3.2}$$

The last term in this log-likelihood is due to the missing studies and its derivative, $\sum_{m=1}^{M} -\left(\frac{\sqrt{n_m}\varphi(b-\sqrt{n_m}(\kappa-\kappa_0))}{\Phi(b-\sqrt{n_m}(\kappa-\kappa_0))}\right)$, is everywhere negative, that is, the term is decreasing in κ. When $\kappa \to \infty$, the derivative tends to $-\infty$ and behaves asymptotically linear in κ, while

for small values of κ, it quickly approaches zero, because the numerator dominates. Its second derivative,

$$-\sum_{m=1}^{M} \left[\frac{n_m \varphi\left(b - \sqrt{n_m}(\kappa - \kappa_0)\right)}{\Phi\left(b - \sqrt{n_m}(\kappa - \kappa_0)\right)} \left(b - \sqrt{n_m}(\kappa - \kappa_0) + \frac{\varphi\left(b - \sqrt{n_m}(\kappa - \kappa_0)\right)}{\Phi\left(b - \sqrt{n_m}(\kappa - \kappa_0)\right)} \right) \right],$$

is negative, which shows that this term is strictly concave in κ. The first two terms in $\log(L(\kappa))$ form the usual normal log-likelihood, a strictly concave quadratic polynomial in κ, whose maximum is at $\hat{\kappa}_{\text{meta}} = \sum_{k=1}^{K} n_k \hat{\kappa}_k / N$ and whose second derivative is $-\sum_{k=1}^{K} n_k$. Substituting the sum of maximum likelihood estimator of κ in the formula for the second derivative of the log-likelihood and taking the negative inverse of this value, allows us to estimate the variance of the maximum likelihood estimate.

The censored log-likelihood (3.2) is a well-behaved, strictly concave function with a unique maximum. Since it is the sum of a concave quadratic with a function close to zero for small values of κ and quickly tending to $-\infty$ for large values, the quadratic first part is modified as shown in Figure 2.

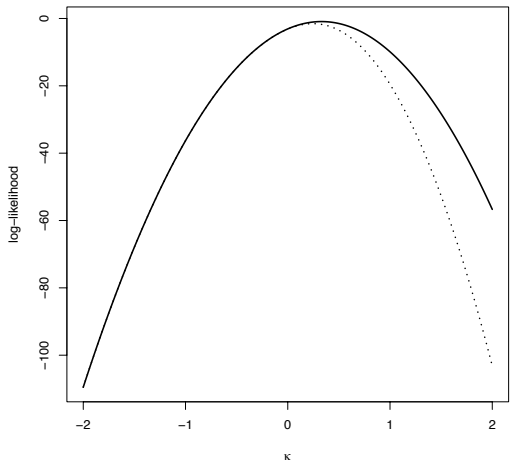

Figure 2: *The log-likelihood based on a single study with* $\hat{\kappa} = 0.33$ *and* $n = 40$ *is shown. The dotted line shows the log-likelihood when adding a missing study of size* $n = 30$*, with hypothesized value* $\kappa_0 = 0$ *and cutoff* $b = z_{0.95} = 1.645$*. The effect of adding the censored part makes large values of* κ *much less likely, whereas the small values are left untouched. As a result, the maximum likelihood of the censored log-likelihood moves towards smaller* κ *values. The maximum likelihood estimate of* κ *based on the censored likelihood becomes 0.245 with an estimated variance of roughly* $1/57$*. The p-value before inclusion of the censored study is 0.018 and becomes 0.031 upon inclusion of the missing study.*

Since the log-likelihood (3.2) is a nice concave function, it is easy to determine the maximum likelihood estimator through the use of the Newton-Raphson algorithm (see Kulinskaya et al. (2008), Chapter 26), with a starting value equal to the meta effect estimated with the observed studies.

With regard to the number of missing studies, several values should probably be tried. If one believes that the significance of the published studies is entirely due to chance, one might be inclined to add the 95% missing studies to the 5% published ones as in Rosenthal (1979). In the above example, this would mean that 19 missing studies existed. If we add these 19 studies with smallish sample sizes of 10 for each, the combined effect is estimated to be 0.08 with a p-value of 0.21. If we take the 19 studies to be small screening studies of size $n = 2$ each, the effect estimate remains quite high (0.21) but is also insignificant (p-value of 0.067). This illustrates the difficulties one faces and which are unavoidable. One has to make quite strong assumptions about the selection mechanism and one has to guess or estimate the number and sample sizes of the missing studies. The statistical analysis we outlined above does not take into account the uncertainty inherent in these choices. All it can do is to tell us under what assumptions the resulting meta effect estimate becomes insignificant.

3.2 Small Sample Biases

The deleterious effect on coverage probability of centering a confidence interval on a biased estimator of the parameter seems to be relatively unknown and is widely ignored in the traditional approach to meta-analysis. We illustrate by a simple example how this bias undermines the coverage probabilities of fixed common effects as the number of studies increases from 1 to 5.

Comparison of four confidence intervals for p based on binomial observations

Confidence intervals for the binomial parameter p have long been of interest to statisticians, see the discussion paper Brown et al. (2001) and literature referred to therein. Amongst the best performers are the Wilson interval (Wilson, 1927), and a similar, but simpler to explain interval by Agresti and Coull (1998). The latter (with $c = 2$) falls in the class of intervals centered on $\tilde{p}_c = (X + c)/(n + 2c)$, where X has a binomial $B(n, p)$ distribution. These estimators are studied by Böhning and Viwatwongkasen (2005). It is easy to see

that $\tilde{p}_c$ is for all $c > 0$ biased towards 1/2 and satisfies $\mathrm{E}[\tilde{p}_c] = p + c(1-2p)/(n+2c)$ and $\mathrm{Var}[\tilde{p}_c] = np(1-p)/(n+2c)^2$.

Agresti and Coull (1998) suggest the interval $\tilde{p}_2 \pm z_{0.975}\sqrt{\tilde{p}_2(1-\tilde{p}_2)/(n+4)}$, which is slightly more conservative than the one obtained using the standard error of $\tilde{p}_2$. This is the first of the four intervals we study and we denote it I_{AC}. Amongst the intervals of the form $\tilde{p}_c \pm z_{0.975}\sqrt{n\tilde{p}_c(1-\tilde{p}_c)}\,/(n+2c)$, Böhning and Viwatwongkasen (2005) recommends the use of $c = 1$ for a number of reasons, so this is the second interval I_{BV} we will include in our comparison.

The third interval I_{AS} to be considered is the traditional arcsine interval which is obtained in two steps: first $\tilde{p}_c$ is transformed to $h(\tilde{p}_c) = 2\arcsin(\sqrt{\tilde{p}_c})$, which is asymptotically normal with mean $h(p)$ and variance $1/n$. This yields the approximate 95% confidence interval $h(\tilde{p}_c) \pm z_{0.975}/\sqrt{n}$ for $h(p)$. This interval, call it $[l, u]$, is truncated to lie within $[0, \pi]$, the range of $h(p)$ for $p \in [0, 1]$, and then back-transformed via $h^{-1}([l, u]) = \sin([l, u]/2)$ to an interval of the same coverage for p. While Anscombe (1948) found that $c = 3/8$ provided the best variance stabilization, here we use $c = 1/2$, which yields similar results.

The above three intervals are biased towards 1/2. We now give a bias correction for the arcsine interval. By expanding $h(\tilde{p}_c)$ about $h(p)$ in a Taylor series and taking the expectation, one finds the bias in $h(\tilde{p}_c)$ for estimating $h(p)$ to be

$$\mathrm{E}[h(\tilde{p}_c) - h(p)] \doteq \frac{\{h'(p)\}^3(1-2p)}{4(n+2c)^2}\left[a_0c^2 + a_1c + a_2\right],$$

where $a_0 = 12p(1-p) - 1$, $a_1 = 4p(1-p)$ and $a_2 = -np(1-p)$. The quadratic in brackets is 0 when $c = 1/4 + O(1/n)$, for p bounded away from 0 and 1. The fourth interval in our study is the arcsine interval based on $h(\tilde{p}_c)$, where $c = 1/4$. We call this interval I_{ASBC}. We note that Böhning and Viwatwongkasen (2005, Theorem 4.1) found that $c = 1/4$ minimizes the average bias in $\tilde{p}_c(1-\tilde{p}_c)$ for estimating $p(1-p)$.

3.2.1 Simulation study comparing the four estimators

In Figure 3 are shown the results of 10,000 simulations of $X \sim B(20, p)$ at increments of 0.01. The upper left plot gives the estimated bias for $\tilde{p}_2$ (solid line) and the bias in $\tilde{p}_1$ (dashed line) for estimating p. These graphs are linear in p; they correspond to the biases in the centers of the AC interval and the BV interval, respectively. Also shown are the biases of $h(\tilde{p}_{0.5})$ (dot-dashed line) and $h(\tilde{p}_{0.25})$ (dotted line) for estimating $h(p)$; they

correspond to centers of the arcsine and bias-corrected arcsine intervals for estimating $h(p)$. Note that the bias-correction works quite well for all p except near the boundaries.

The top right plot in Figure 3 shows the empirical standard deviation of each of the four estimators, when divided by its estimated standard deviation. For example, the solid line depicts the graph of

$$\sqrt{\widehat{\mathrm{Var}}\left(\tilde{p}_2/\sqrt{\tilde{p}_2(1-\tilde{p}_2)/(n+4)}\right)} \text{ versus } p\,,$$

where $\widehat{\mathrm{Var}}$ is the estimate based on the simulation. This would be close to 1 if the standard deviation were known. However, this plot shows that this ratio is far from 1, and the situation does not improve with increasing n, as other simulations, not shown, demonstrate. Similar remarks apply to the standardized $\tilde{p}_1$ shown in the dashed line. The standardized version of $h(\tilde{p}_{0.5})$, namely $\sqrt{n}\,h(\tilde{p}_{0.5})$ in the dot-dashed line has empirical standard deviation near 1, as one would expect from a variance stabilized statistic. Somewhat surprisingly, the bias-corrected version $\sqrt{n}\,h(\tilde{p}_{0.25})$ in the dotted line reveals an empirical standard deviation even closer to 1 over the range of p.

The left-hand plot in the second row of Figure 3 shows the empirical coverages of the 4 intervals, again as a function of p. Note that neither the BV or the bias-corrected arcsine interval provides adequate coverage over the range of p. However, the AC interval and the arcsine interval have coverage between 94% and 97% for most values of p.

What are the implications for meta-analysis? We illustrate the phenomenon of how persistent bias undermines coverage probabilities in the bottom right plot of Figure 3, where the intervals are based on $K = 5$ studies, each of size $n = 20$. The AC and BV intervals are centered on the weighted average of the 5 individual estimates of p, namely $\tilde{p}_{c,w} = \sum \tilde{w}_k \tilde{p}_{c,k}/(\sum \tilde{w}_k)$, where $\tilde{w}_k$ is the estimated inverse variance of $\tilde{p}_{c,k}$. To obtain the meta-interval of nominal coverage 95%, one takes $\tilde{p}_{c,w} \pm z_{0.975}/(\sum \tilde{w}_k)^{1/2}$.

The arcsine interval $[L_{\mathrm{AS}}, U_{\mathrm{AS}}]$ for $h(p)$ is obtained by finding the weighted average $\sum_k n_k h(\tilde{p}_{0.5,k})/(\sum_k n_k)$ and adding and subtracting $z_{0.975}/(\sum_k n_k)^{1/2}$. This is then back-transformed to an interval for p via $h^{-1}([L_{\mathrm{AS}}, U_{\mathrm{AS}}])$. The bias-corrected version is found in the same way, starting with the bias-corrected estimators $\tilde{p}_{0.25,\,k}$, $k = 1, \ldots, 5$.

The results are plain to see in the bottom-right hand plot of Figure 3. The estimated inverse weights coverage probabilities are now unacceptably low, even though these intervals are based on more information than the bottom-left hand plot which is based on a single study. Of course the AC and BV intervals were not designed for meta-analysis, but

they were used here to illustrate the point that what works well for one study, will not necessarily work at all in a meta-analysis using estimated inverse variance weights, especially if the center of the interval is biased for the parameter of interest. The plots in Figure 3 are typical of what we found for many more choices of sample sizes $n_1, \ldots, n_K$ for this fixed effects model. Increasing the study sample sizes does not remove the problem; bias will cause significant loss of coverage with K only 5 or 10 and the situation deteriorates with increasing K. On the other hand, the bias-corrected arcsine interval maintains its good coverage for a large number of studies, even $K = 100$.

4 Multivariate meta-analysis and meta-regression

Development of multivariate methods is necessitated by an abundance of studies reporting a number of correlated outcomes which are habitually meta-analyzed independently using univariate methods. Multiple outcomes appear by design in such areas as diagnostic test meta-analysis and network meta-analysis. The former usually models sensitivity and specificity of diagnostic tests requiring bivariate analysis, see van Houwelingen et al. (1993); Harbord et al. (2007); Putter et al. (2010); Paul et al. (2010). The latter aims to compare multiple treatments, see the recent reviews Jansen et al. (2011); Hoaglin et al. (2011) for more details. Multiple outcomes, such as disease-free and all-cause survival, also appear routinely in both clinical trials and observational studies. See Jackson et al. (2011) and its discussion for comprehensive reviews.

The potential applications of multivariate analysis are found in the collection Arends (2006), which summarizes most of the material in Arends et al. (2000); van Houwelingen et al. (2002); Arends et al. (2003, 008a,b). These papers lead the way in demonstrating what one could do if normality of effects sizes with known variances is assumed. But there are neither sensitivity analyses nor mathematical arguments nor simulation studies to support the models. Applications include comparing relative risks of treatment and control arms, comparing ROC curves, comparing risks with baseline risks, and comparison of survival curves. A recent detailed review of Arends (2006) is given in Staudte (2010).

4.1 Multivariate meta-analysis

For the multivariate random effects model it is assumed that vectors of estimates

$$\boldsymbol{y}_k = \boldsymbol{\mu} + \boldsymbol{b}_k + \boldsymbol{\epsilon}_k, \quad k = 1, \cdots, K,$$

where now $\boldsymbol{\mu}$, $\boldsymbol{b}_k$ and $\boldsymbol{\epsilon}_k$ are p-vectors, and $\boldsymbol{b}_k$ and $\boldsymbol{\epsilon}_k$ have zero means and covariance matrices $\boldsymbol{\Sigma_b}$ and $\boldsymbol{\Sigma}_k$. Usually $\boldsymbol{b}_k$ and $\boldsymbol{\epsilon}_k$ are assumed to be independent and multivariate normal. These terms have the same interpretation as in the univariate REM. In the normal case, marginally

$$\boldsymbol{y}_k \sim \mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma_b} + \boldsymbol{\Sigma}_k),$$

where vectors of estimates $\boldsymbol{y}_k$ are further assumed to be independent because they come from different studies. Exactly as in the univariate meta-analysis, the within-study variances $\boldsymbol{\Sigma}_k$ are assumed to be known, and the goal is to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma_b}$. When not all outcomes are observed in all studies, a slightly more general model has $\mathrm{Cov}(\boldsymbol{b}_k) = \boldsymbol{V}_k = \boldsymbol{H}_k \boldsymbol{\Omega} \boldsymbol{H}_k'$ for design matrices $\boldsymbol{H}_k$. Here $\boldsymbol{\Omega}$ is called the heterogeneity matrix by Demidenko (2004, Section 5.3.1). For simplicity, we shall assume that all studies provide all effects.

The pooled estimate of $\boldsymbol{\mu}$ is given by a familiar weighted mean

$$\hat{\boldsymbol{\mu}} = \Big(\sum_{k=1}^{K} (\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma_b})^{-1}\Big)^{-1} \Big(\sum_{k=1}^{K} (\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma_b})^{-1} \boldsymbol{Y}_k\Big)\,.$$

The standard approach to inference about the effects substitutes estimated $\widehat{\boldsymbol{\Sigma_b}}$ instead of true between-studies variance. Then the variance of $\hat{\boldsymbol{\mu}}$ is given by

$$\mathrm{Var}(\hat{\boldsymbol{\mu}}) = \Big(\sum_{k=1}^{K} (\boldsymbol{\Sigma}_k + \widehat{\boldsymbol{\Sigma_b}})^{-1}\Big)^{-1}\,.$$

How large should be the number of studies K and the within studies sample sizes for this to work is not known. This covariance matrix can be used to provide further inference such as univariate and joint confidence regions.

Statistical issues related to the actual randomness of the estimated within-studies covariance matrices $\boldsymbol{\Sigma}_k$ have not yet been addressed in the multivariate meta-analysis. Estimation of $\widehat{\boldsymbol{\Sigma_b}}$ can be achieved by maximum likelihood, REML, profile likelihood, and the method of moments (unweighted or weighted) (Jackson et al., 2011). Chen et al. (2012) developed a noniterative method of moments matrix estimator for the between-study covariance matrix. This estimator is a multivariate extension of DerSimonian and Laird's

univariate method of moments estimator, and it is invariant to linear transformations. Attention should be given to the constraint that this matrix should be positive-definite. The simplest way to achieve this is to use a projection of $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{b}}}$ on the set of nonnegative definite matrices found by $\widetilde{\boldsymbol{\Sigma}_{\boldsymbol{b}}} = \boldsymbol{P}\boldsymbol{\Lambda}_{+}\boldsymbol{P}'$, where $\boldsymbol{P}$ is the matrix of eigenvectors and $\boldsymbol{\Lambda}_{+}$ is the matrix of truncated at 0 eigenvalues of matrix $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{b}}}$, Demidenko (2004, Section 5.3.3). To relax the assumption of normality, many of these methods can be cast in an unbiased estimating equations (EE) format (Ritz et al., 2008). These EE methods are asymptotically ($K \to \infty$) equivalent to ML for distributions with zero third moment, and are much simpler computationally. Bayesian methods are available but they may be particularly sensitive to the choice of prior with the increase in the dimensionality, see Jackson et al. (2011) for discussion. Rukhin (2007) investigated the multivariate REM under normality. He does not consider within-study covariance matrices Σ_k known; under normality their sample counterparts have the scaled Wishart distribution. He discussed ML and REML estimation, but the main focus of his paper is the generalization of DerSimonian-Laird and Mandel-Paule techniques to the multivariate setting, see also Jackson et al. (2010). Similar to the univariate case, the Mandel-Paule estimator is close to the REML estimator. Rukhin (2007) also provides two residuals-based estimators of the covariance matrix of the weighted mean $\hat{\boldsymbol{\mu}}$. He showed that, for a general matrix of weights, an unbiased quadratic estimator of $\mathrm{Var}(\hat{\boldsymbol{\mu}})$ does not exist, and derived an almost unbiased estimator which is considerably better than the standard inverted sum of weights.

4.2 Multivariate meta-regression

Multivariate meta regression further considers $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_k})'$ is an $n \times p$ design matrix, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. This problem is very similar and requires the same techniques. For any weight matrix $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{W_1}, \cdots, \boldsymbol{W_k})$ and a covariance matrix $\boldsymbol{S} = \mathrm{diag}(\boldsymbol{S_1}, \cdots, \boldsymbol{S_k})$, the weighted least-squares estimator of $\boldsymbol{\beta}$ is

$$\boldsymbol{b_W} = \Big(\sum_{k=1}^{K} \boldsymbol{X_k'} \boldsymbol{W_k} \boldsymbol{X_k}\Big)^{-1} \Big(\sum_{k=1}^{K} \boldsymbol{X_k} \boldsymbol{W_k} \boldsymbol{Y_k}\Big)\,.$$

The covariance matrix of $\boldsymbol{b}$ is

$$\mathrm{Cov}(\boldsymbol{b_W}) = \Big(\sum_{k=1}^{K} \boldsymbol{X_k'} \boldsymbol{W_k} \boldsymbol{X_k}\Big)^{-1} \Big(\sum_{k=1}^{K} \boldsymbol{X_k} \boldsymbol{W_k} \boldsymbol{S_k} \boldsymbol{W_k} \boldsymbol{X_k}\Big) \Big(\sum_{k=1}^{K} \boldsymbol{X_k'} \boldsymbol{W_k} \boldsymbol{X_k}\Big)^{-1}\,.$$

The main difficulty in the multivariate meta-analysis and meta-regression is the lack

of knowledge of the within-study correlations. These correlations are not reported in the underlying studies. Thus the matrices $\boldsymbol{\Sigma}_k$ are not known beyond the estimated main diagonals, unless the full within-studies data are available. Correlations can be calculated analytically for mutually exclusive binary outcomes from a multinomial distribution, such as death from different causes (Trikalinos and Olkin, 2008), but this is not possible in general. One possibility is to make some assumptions about the within-study correlations, the simplest being the constant within-study correlation ρ. In this vein, Riley et al. (2008) proposes to replace the marginal covariance matrix $\boldsymbol{\Sigma_b} + \boldsymbol{\Sigma}_k$ by a matrix with constant correlation instead of separate within- and between-studies correlations. This approach was used only in a bivariate setting, and its generalizability is not clear.

A recent paper (Hedges et al., 2010) proposed to estimate the covariance matrix $\mathrm{Cov}(\boldsymbol{b_W})$ of a weighted estimator in a meta-regression context by substituting empirical within-study covariance matrices $\boldsymbol{e_k e_k'}$, i.e. the matrices of weighted cross-products of within-study residuals $\boldsymbol{e_k} = \boldsymbol{Y_k} - \boldsymbol{X_k b}$ for $\boldsymbol{S_k}$ in the above equation. Thus, the covariance is estimated by

$$\widehat{\boldsymbol{V}} = \Big(\sum_{k=1}^{K} \boldsymbol{X_k' W_k X_k}\Big)^{-1} \Big(\sum_{k=1}^{K} \boldsymbol{X_k W_k e_k e_k' W_k X_k}\Big) \Big(\sum_{k=1}^{K} \boldsymbol{X_k' W_k X_k}\Big)^{-1}.$$

For a sequence of independent random matrices $\boldsymbol{X_k}$, $\boldsymbol{W_k}$, $\boldsymbol{\epsilon_k}$, with diagonal matrices of weights $\boldsymbol{W_k}$, when $K \to \infty$ Hedges et al. (2010) showed that under some regularity conditions $\boldsymbol{b}$ is approximately $N_p(\boldsymbol{\beta}, \boldsymbol{V})$. These conditions include uniform boundedness of the weights, the error variances and various covariances, and some moment conditions. Further Hedges et al. (2010) propose taking equal (inverse sum of variances) fixed weights within a study, and to specify a single value of within-study correlation ρ. Then the random variance component τ^2 can be estimated from the weighted residual sum of squares. Hedges et al. (2010) recommend sensitivity analysis in respect of the influence of ρ values on the estimated τ^2. They also provide a simulation study and an R program. Recommendation to use inverse variance weights in the asymptotics $K \to \infty$ with bounded sample sizes seems to be rather risky. On the other hand, equal weights within a study would result in an unbiased estimate of the covariance of $\boldsymbol{b_W}$ (Rukhin, 2007), as long as the maximum study weight $w_k < 1/2$, and a simple correlation structure is a necessity in multivariate meta-analysis when the covariances are not known, so overall this approach is a promising development.

5 Sequential meta analysis

The goal of meta-analysis is to combine evidence from various studies of the same effect to provide a more reliable basis for decision making. However, any given meta-analysis provides just a snapshot of the available evidence at a given point in time. At first there may not be sufficient evidence for an effect, but as additional studies become available, the significance of the new effect can be established. Detecting the moment when significance is reached is a statistical problem also found in group sequential trials.

Further, temporal changes in effect sizes of substantial magnitude may occur, leading to the loss or gain of statistical significance or even a change in the sign of the cumulative mean effect. Such changes have been reported in medicine, ecology, the social sciences and evolutionary biology. Factors initiating such change include time-lag bias, publication bias, heterogeneity among studies, more effective research methods, paradigm shifts or underlying changes in effects over time due to temporal changes in baseline values or strength of causal agents (Koricheva and J., 2011).

These results suggest that results of meta-analyses conducted early in the process of research accumulation should be interpreted with caution. Detection of temporal trends in effect sizes is therefore an important methodological issue. Once temporal changes are detected, one needs a model for this trend which will enable one to measure it. Cumulative meta-analysis (CMA) was introduced in medical research to detect the earliest date for which a treatment effect became statistically significant, and for which its clinical efficiency could be evaluated (Lau et al., 1992). In a CMA studies are sorted in chronological order and combined estimates of the effect $\hat{\theta}_k$ and their confidence intervals are plotted and scrutinized visually for possible temporal trends. As with all visual tools, CMA are subject to misinterpretation and need to be supplemented by formal statistical methods. Such methods should also take into account multiple testing inherent in CMA. Thus sequential statistical methods are in order.

A recently proposed class of methods for detecting the moment when the significance is reached is based on statistical methods developed for sequential clinical trials, Pogue and Yusuf (1997); Brok et al. (2008); Wetterslev et al. (2008); van der Tweel and Bollen (2010); Higgins et al. (2011). Group sequential methods define the Trial Sequential Monitoring Boundaries based on prespecified type I and II errors and on clinically important effect. Different types of monitoring boundaries can be used to stop the meta-analysis.

Pogue and Yusuf (1997) adapted the Lan and DeMets (1983) alpha spending function with O'Brien-Fleming boundaries to cumulative fixed effects meta-analysis. They defined the Information Size (IS) as the sample size required to achieve power for one study under the fixed effects model.

Brok et al. (2008) and Wetterslev et al. (2008) inflated these boundaries to accommodate the random effects meta analysis. This method is called Trial Sequential Analysis (TSA). The inflation coefficient is the ratio of sample sizes required under the REM and FEM. Given that the sample size calculation under REM requires, in general, an increase in the number of trials K (Goudie et al., 2010), this method appears too simplistic.

Several recent papers (van der Tweel and Bollen, 2010; Higgins et al., 2011) used group sequential procedures by Whitehead (1999) and his package PEST in application to random effects CMA. These procedures are designed to satisfy a pre-specified power requirement. The main difference between TSA and these methods is the use of the sum of weights W_K as the IS instead of the sample size. Here the scaled cumulative effect $Z_k = W_k\theta_k$ is plotted against W_k, for an accumulated sum of weights W_k. A decision is made when the trajectory crosses a monitoring boundary. Whitehead (1997) used the triangular design and van der Tweel and Bollen (2010) used the double-triangular design. Higgins et al. (2011) use the restricted procedure by Whitehead, equivalent to an O'Brian and Flemming stopping rule.

The theory of group sequential methods is based on an approximation of the trajectory $\{Z_k, W_k\}$ by a Brownian motion. For the FEM, the increments $Z_k - Z_{k-1}$ are independent, and this approximation is valid. For REM, these increments have a complicated correlation structure resulting from dependencies between estimated random variance components $\hat{\tau}_k^2$. Thus an application of group sequential methods to random effects CMA is not justified theoretically.

Lan et al. (2003) proposed to penalize the test statistic using the law of iterated logarithm to account for multiple tests. Specifically, they propose to use $Z_k^* = \sqrt{W_k}\hat{\theta}_k/\sqrt{\lambda \ln\ln(W_k)}$ for testing in the FEM, and $Z_k^{**} = \sqrt{W_k^*}\hat{\theta}_k^*/\sqrt{\lambda \ln\ln(W_k^*)}$ for the REM, where W_k^* and $\hat{\theta}_k^*$ correspond to accumulated sum of weights and the combined effect for the REM. The penalizing constants λ are chosen from simulations to be 1.5 in the FEM, and 2 in the REM.

A difficult related problem is an estimation of the variance component τ^2 when the number of studies is small in the early stages of a cumulative meta-analysis. A semi-

Bayesian method was proposed by Higgins et al. (2011). Lan et al. (2003) used the sample variance of effects which is clearly overestimating τ^2 but helps to control the overall level α.

Kulinskaya and Koricheva (2010) proposed the use of standard quality control (QC) charts, in particular $\bar{X}$ charts and CUSUM charts to detect possible outliers and trends over time in meta-analysis. The CUSUM charts, equivalent to sequential likelihood ratio tests, seem to be especially well suited for use in the CMA, both to assess significance, and for detection of temporal trends. So far they were used only in the fixed effects model, but a generalization of CUSUM charts to REM seems to be straightforward.

In order to model temporal trends, a number of papers used various regression approaches. Linear regression is used in Gehr et al. (2006) and Kampichler and Bruckner (2009). Regression applied to consecutive combined effects is proposed in Bagos and Nikolopoulos (2009). An exponential model of proportional decrease in effect is introduced in Baker and Jackson (2010). So far there is no empirical evidence on the actual shapes of temporal trends in effect sizes, and this is one more possible area of future work.

6 Available software packages

There are numerous packages, commercial, shareware and free, containing some meta-analytic capabilities. Here we mention just several main contenders and provide some further references.

The majority of systematic reviews in medicine and health sciences use meta-analytic procedures included in Revman, the software by Cochrane Collaboration, now at version 5.2, http://ims.cochrane.org/revman. Description of the statistical methods used can be found in Deeks and Higgins (2010). These are the basic procedures for univariate continuous and binary data, and also some methods for meta-analysis of diagnostic tests accuracy.

The main commercial package is the Comprehensive Meta-Analysis (CMA), developed by a group of experts specialising in meta-analysis. The package includes the standard univariate methods, 1-factor or one covariate meta-regression, and cumulative meta-analysis, along with a number of bias-detection tools. See Bax et al. (2007) for a review of seven specialised software packages including the CMA. Comparison of the CMA to some programs available in SAS, SPSS, Excel and other general statistical software is given in

http://www.meta-analysis.com/pages/comparisons.html.

A recently developed free package, the Meta-analyst (Wallace et al., 2009), includes the same capabilities. The reference also includes a review of the existing software, against which the Meta-analyst was extensively tested.

The main development of new statistical methods of meta-analysis is happening in either Stata or R computational environment.

Stata includes a number of user-written packages, see Sterne (2009) and http://www.stata.com/support/faqs/statistics/meta-analysis/ for the full list. The procedures include both standard methods and various advanced options, such as multivariate random-effects meta-analysis (White, 2009).

R has a similar array of continuously developing packages for meta-analysis. The main three are *meta* by Guido Schwartzer, http://cran.r-project.org/web/packages/meta/meta.pdf, *metafor* by Wolfgang Viechtbauer, http://www.metafor-project.org/, Viechtbauer (2009) and *mvmeta* by Antonio Gasparrini, http://cran.r-project.org/web/packages/mvmeta/mvmeta.pdf, Gasparrini et al. (2012). *meta* includes all standard univariate methods and some extensions; *metafor* additionally includes meta-regression, and a comprehensive array of options for estimation in the random effects model; and *mvmeta* includes multivariate meta-analysis.

7 Summary

The paper reviews the procedures and open problems in statistical meta-analysis. This is a topic of growing importance, because in many areas of application the need for combining different sources of data and different sources of information in order to reach an overall assessment manifests itself. The classical material on meta-analytic statistical procedures are reviewed in Section 2. These include the distinction between fixed and random effects models, tests of homogeneity, and a discussion of the types of data typically available for a meta-analysis. Section 3 investigates two sources of bias in meta-analysis, the bias due to the systematic selection of studies showing stronger than average effects, and the meta-analysis of smallish studies combined with the use of estimators with appreciable small sample biases. In Sections 4 and 5 generalizations and extensions to the standard procedures are discussed. They include meta-regression, multivariate responses and sequential

procedures. A non-exhaustive list of software tools in Section 6 closes out the review.

A list of research problems for statisticians that we have identified follows:

1. Development of methods for the FEM that are based on the 'actual likelihood' (without the assumptions of known within-study variances and normality). By 'development', we mean both theoretical and simulation study analyses, as well as provision of user-friendly software (Section 2.1).

2. The same problems exist for the REM, with the additional complication of estimating inter-study variability. Properly standardized measures of heterogeneity and valid tests for heterogeneity are yet to be developed for the REM. Alternative models such as multiplicative ones should be investigated. (Section 2.4 and 2.3).

3. Investigation of actual likelihood models, or alternatively quasi-likelihood models following transformations, which allow not only for moderators but additional inter-study variation. For conceptual simplicity these models should reduce to the REM when there are no significant moderators. (Section 2.5).

4. Modelling of biases in observational studies and further development of sensitivity measures for such studies. (Section 2.7).

5. Although the qualitative issues regarding publication bias are well-understood, little has been done to assess the degree of, and correct for, this widespread problem. (Some detailed theory in this direction is provided in Section 3).

6. Removing the deleterious effect of small sample biases, as carried out in Section 3.2 for the one-sample binomial effects, for numerous other effects such as the relative risk and odds ratio.

7. Multivariate meta-analysis and meta-regression methods promise to help solve a broad range of problems in the health, social and medical sciences, as detailed in Section 4.1. However, nearly all the research so far is based on the over-optimistic assumptions of known covariance matrices and normality of errors. Much more publicity needs to encourage gathering of covariate sample correlations, and modeling of such to provide justification for assumed correlation structures. Again, sensitivity measures for model miss-specification are sorely needed. As in the univariate case, development of more realistic methodology for actual likelihood models presents many

challenging problems, perhaps best solved on case by case basis in the context of the same multivariate response data sets with similar covariates.

8. Although many of the above problems require more sophisticated attention, their assumptions can be violated by temporal changes in effects, and so monitoring methods and tests for such changes are required. When found, suitable meta-analyses with built-in trends are required, especially to help in planning for future studies. Careful asymptotics with attention to the rates at which within-study sample sizes grow with the number of studies are required here, as in the above-described situations. (Section 5.)

Acknowledgements

We would like to thank the two referees whose comments helped us to improve the manuscript. The work by the second author was supported by a grant from the Swiss National Science Foundation.

References

Agresti, A. and Coull, B. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, 52:119–126.

Akritas, M. and Papadatos, N. (2004). Heteroscedastic one-way ANOVA and lack-of-fit tests. *Journal of the American Statistical Association*, 99:368–382.

Anscombe, F. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika*, 35:266–254.

Arends, L. (2006). *Multivariate meta-analysis: modelling the heterogeneity; Mixing apples and oranges: dangerous or delicious?* Haveka BV, Alblasserdam.

Arends, L., Hamsa, T., van Houwelingen, H., Heijenbrok-Kai, M., Hunink, M., and Stijnen, T. (2008a). Bivariate random effects meta-analysis of ROC curves. *Medicine Decision Making*, 28:621–638.

Arends, L., Hoes, A., Lubsen, J., Grobbee, D., and Stijnen, T. (2000). Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine*, 19:3497–3518.

Arends, L., Hunink, M., and Stijnen, T. (2008b). Meta-analysis of summary survival curve data. *Statistics in Medicine*, 27:4381–4396.

Arends, L., Voko, Z., and Stijnen, T. (2003). Combining multiple outcome measures in a meta-analysis: an application. *Statistics in Medicine*, 22:1335–1353.

Bagos, P. and Nikolopoulos, G. (2009). Generalized least squares for assessing trends in cumulative meta-analysis with applications in genetic epidemiology. *Journal of Clinical Epidemiology*, 62:1037–1044.

Baker, R. and Jackson, D. (2010). Inference for meta-analysis with a suspected temporal trend. *Biometrical Journal*, 52:538–551.

Bax, L., Yu, L.-M., Ikeda, N., and Moons, K. (2007). systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology*, 7:40.

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

Biggerstaff, B. and Jackson, D. (2008). The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine*, 27:6093–6110.

Biggerstaff, B. and Tweedie, R. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, 16:753–768.

Böhning, D. and Viwatwongkasen, C. (2005). Revisiting proportion estimators. *Statistical Methods in Medical Research*, 14:147–169.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2011). *Introduction to meta-analysis*. Wiley.

Brok, J., Thorlund, K., Gluud, C., and Wetterslev, J. (2008). Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *Journal of Clinical Epidemiology*, 61:763–769.

Brown, L., Cai, T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16:101–112.

Buonaccorsi, J. (2006). Estimation in two-stage models with heteroscedasticity. *International Statistical Review*, 74:403–418.

Chang, C.-H. and Pal, N. (2008). Testing on the common mean of several normal distributions. *Computational Statistics and Data Analysis*, 53:321–333.

Chen, H., Manning, A. K., and Dupuis, J. (2012). A method of moments estimator for random effect multivariate meta-analysis. *Biometrics*, 68:1278–1284.

Cochran, W. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, 4:102–118.

Colditz, G., Brewer, T., Berkey, C. S., Wilson, M. E., Burdick, E., Fineberg, H. V., and Mosteller, F. (1994). Efficacy of bcg vaccine in the prevention of tuberculosis. meta-analysis of the published literature. *The Journal of the American Medical Assiciation*, 271:698–702.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2009). *Handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Deeks, J., Dinnes, J., D'Amico, R., Sowden, A., Sakarovitch, C., Song, F., Petticrew, M., and Altman, D. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, 7:1–173.

Deeks, J. and Higgins, J. (2010). Statistical algorithms in review manager 5. Technical report, Cochrane Collaboration. on Behalf of the Statistical Methods Group of the Cochrane Collaboration.

Demidenko, E. (2004). *Mixed Models. Theory and applications*. John Wiley & Sons.

Demidenko, E., Sargent, J., and Onega, T. (2012). Random effects coefficient of determination for mixed and meta-analysis models. *Communications in Statistics-Theory and Methods*, 41:953–969.

DerSimonian, R. and Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, 28:105–114.

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

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

Duval, S. and Tweedie, R. (2000b). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56:455–463.

Egger, M., Smith, G., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315:629–634.

Fisher, S. (1932). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Fleiss, J. (1993). Review papers: The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2:121.

Gasparrini, A., Armstrong, B., and Kenward, M. G. (2012). Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine*, 31:3821–3839.

Gehr, B., Weiss, C., and Porzsolt, F. (2006). The fading of reported effectiveness. a meta-analysis of randomised controlled trials. *BMC Medical Research Methodology*, 6:25.

Givens, G., Smith, D., and Tweedie, R. (1997). Publication in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*, 12:221–250.

Goudie, A., Sutton, A., and Jones, D.R.and Donald, A. (2010). Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. *Journal of Clinical Epidemiology*, 63:983–991.

Harbord, R., Deeks, J., Egger, M., Whiting, P., and Sterne, J. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 8:239–251.

Hardy, R. and Thompson, S. (1996). A likelihood approach to meta-analysis. *Statistics in Medicine*, 15:619–629.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, 41:901–916.

Hartung, J. and Knapp, G. (2003). An alternative test procedure for meta-analysis. In Schulze, H., Holling, H., and Bøhning, D., editors, *Meta-analysis*. Hogrefe & Huber, Gøttingen.

Hartung, J. and Makambi, K. (2003). Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics: Simulation and Computation*, 32:1179–1190. DOI: 10.1081/SAC-120023884.

Hedges, L. and Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, Orlando.

Hedges, L. and Olkin, I. (1993). Regression models in research synthesis. *The American Statistician*, 37:137–140.

Hedges, L., Tipton, E., and Johnson, M. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1:39–65.

Henmi, M. and Copas, J. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, 29:2969–2983.

Higgins, J. and Green, S., editors (2011). *Cochrane handbook for systematic reviews of interventions: Version 5.1.0 [updated March 2011]*. Cochrane Collaboration.

Higgins, J. and Thompson, S. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine*, 21:1531–1558.

Higgins, J., Thompson, S., and Spiegelhalter, D. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A*, 172:137–159.

Higgins, J., Whitehead, A., and Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. *Statistics in Medicine*, 30:903–921.

Hoaglin, D., Hawkins, N., Jansen, J., and et al. (2011). Conducting indirect-treatment-comparison and network-meta-analysis studies: Report of the ISPOR task force on indirect treatment comparisons good research practices: Part 2. *Value Health*, 14:429–437.

Huizenga, H., Visser, I., and Dolan, C. (2011). Testing overall and moderator effects in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology*, 64:1–19.

Inc., S. (1989). *SAS/STAT User's Guide, Version 6*. SAS Institute Inc., Cary, North-Carolina.

Jackson, D. (2006). The power of the standard test for the presence of heterogeneity in meta-analysis. *Statistics in Medicine*, 25:2688–2699.

Jackson, D., Riley, R., and White, I. (2011). Multivariate meta-analysis: potential and promise. *Statistics in Medicine*, 30:2481–2498.

Jackson, D., White, I., and Thompson, S. (2010). Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine*, 29:1282–1297.

James, G. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38:324–329.

Jansen, J., Fleurence, R., Devine, B., and et al. (2011). Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: Report of the ispor task force on indirect treatment comparisons good research practices: Part 1. *Value Health*, 14:417–428.

Jones, A. P., Riley, R. D., Williamson, P. R., and Whitehead, A. (2009). Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical Trials*, 6:16–27.

Kampichler, C. and Bruckner, A. (2009). The role of microarthropods in terrestrial decomposition: a meta-analysis of 40 years of litterbag studies. *Biological Reviews*, 84:375–389.

Konnerup, M. and Kongsted, H. (2009). There is more to seeing than meets the eye: Observational studies, research synthesis, and the social sciences. Technical report, London School of Economics and Political Science, Centre for the Philosophy of Natural and Social Science: Contingency and Dissent in Science.

Koricheva, J., Gurevitch, J., and Mengeresen, K. (2013). *Handbook of meta-analysis in ecology and evolution*. Princeton University Press.

Koricheva, J. and J., J. M. L. (2011). Temporal changes in effect sizes: causes, detection and implications. In Koricheva, J., Gurevitch, J., and Mengersen, K., editors, *The Handbook of Meta-Analysis for Ecology and Evolution*. Princeton University Press, Princeton, USA.

Kulinskaya, E., Dollinger, M., and Bjørkestøl, K. (2011a). On the moments of cochran's Q statistic; with application to the meta-analysis of risk difference. *Research Synthesis Methods*, 2:254–270.

Kulinskaya, E., Dollinger, M., and Bjørkestøl, K. (2011b). Testing for homogeneity in meta-analysis I. the one parameter case: Standardized mean difference. *Biometrics*, 67:203–212.

Kulinskaya, E. and Koricheva, J. (2010). Use of quality control charts for detection of outliers and temporal trends in cumulative meta-analysis. *Research Synthesis Methods*, 1:297–307.

Kulinskaya, E., Morgenthaler, S., and Staudte, R. (2008). *Meta Analysis: a guide to calibrating and combining statistical evidence*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd.

Kulinskaya, E. and Olkin, I. (2013). An overdispersion model in meta analysis. *Statistical Modelling: An International Journal*, in print.

Kulinskaya, E., Staudte, R., and Gao, H. (2003). Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Communications in Statistics—Theory and Methods*, 32:2353–2371.

Lan, K. and DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663.

Lan, K. K. G., Hu, M., and Cappelleri, J. (2003). Applying the law of iterated logarithm to cumulative meta-analysis of a continuous endpoint. *Statistica Sinica*, 13:1135–1145.

Lau, J., Antman, E., Jimenez-Silva, J., Kupelnick, B. M. F., and Chalmers, T. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N. Engl. J. Med.*, 327:248–254.

Lehrer, J. (2010). The truth wears off. *The New Yorker*.

Li, Y., Shi, L., and Roth, H. (1994). The bias of the commonly-used estimate of variance in meta-analysis. *Commun. Statist.-Theory Meth.*, 23:1063–1085.

Light, R. and Pillemer, D. (1984). *Summing up: the science of reviewing research*. Harvard University Press.

Lin, D. and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97:321–332.

Malloy, M., Prendergast, L., and Staudte, R. (2011). Comparison of methods for fixed effect meta-regression of standardized differences of means. *Electronic Journal of Statistics*, 5:83–101.

Malloy, M., Prendergast, L., and Staudte, R. (2013). Transforming the model-T: Random effects meta-analysis with stable weights. *Statistics in Medicine*, 32:1842–1864. DOI: 10.1002/sim.5666.

Mandel, J. and Paule, R. (1970). Interlaboratory evaluation of a material with unequal number of replicates. *Analytical Chemistry*, 42:1194–1197.

Mathew, T., Nahtman, T., von Rosen, D., and Sinha, B. (2010). Nonnegative estimation of variance components in heteroscedastic one-way random effects ANOVA models. *Statistics: A Journal of Theoretical & Applied Statistics*, 44:557–569.

Mathew, T. and Nordstroöm, K. (1999). On the equivalence of meta-analysis using literature and using individual patient data. *Biometrics*, 55:1221–1223.

Mathew, T. and Nordstroöm, K. (2010). Comparison of one-step and two-step meta-analysis models using individual patient data. *Biometrical Journal*, 52:271–287.

Normand, S.-L. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18:321–359.

Paul, M., Riebler, A., Bachmann, L., Rue, H., and Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested laplace approximations. *Statistics in Medicine*, 29:1325–1339.

Pigott, T. D. (2012). *Advances in Meta-analysis*. Springer Science+ Business Media.

Pogue, J. and Yusuf, S. (1997). Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials*, 18:580–593.

Putter, H., Fiocco, M., and Stijnen, T. (2010). Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biometrical Journal*, 52:95–110.

Reeves, B. C. and Wells, G. A., editors (2013). *Special Issue of Research Synthesis: Inclusion of Non-Randomized Studies in Systematic Reviews*, volume 4.

Riley, R., Thompson, J., and Abrams, K. (2008). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, 9:172–186.

Ritz, J., Demidenko, E., and Spiegelman, D. (2008). Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects. *Journal of Statistical Planning and Inference*, 138:1919–1933.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86:638.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.

Rukhin, A. (2003). Two procedures of meta-analysis in clinical trials and interlaboratory studies. *Tatra Mountains Mathematical Publications*, 26:155–168.

Rukhin, A. (2007). Estimating common vector mean in interlaboratory studies. *Journal of Multivariate Analysis*, 98:435–454.

Rukhin, A. (2009). Weighted means statistics in interlaboratory studies. *Metrologia*, 46:323–331.

Rukhin, A. (2013). Estimating heterogeneity variance in meta-analysis. *Journal of Royal Statistical Society B*, 75.

Rukhin, A., Biggerstaff, B., and Vangel, M. (2000). Restricted maximum likelihood estimation of a common mean and the mandel-paule algorithm. *Journal of Statistical Planning and Inference*, 83:319–330.

Rukhin, A. and Vangel, M. (1998). Estimation of a common mean and weighted means statistics. *Journal of the American Statistical Association*, 93:303–308.

Salanti, G. and Ioannidis, J. (2009). Synthesis of observational studies should consider credibility ceilings. *Journal of Clinical Epidemiology*, 62:115–122.

Schulze, R., Holling, H., and Böhning, D. (2003). *Meta-analysis*. Hogrefe & Huber.

Sharma, G. and Mathew, T. (2011). Higher order inference for the consensus mean in inter-laboratory studies. *Biometrical Journal*, 53:128–136.

Shuster, J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine*, 29:1259–1265.

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

Sidik, K. and Jonkman, J. N. (2003). On constructing confidence intervals for a standardized mean difference in meta-analysis. *Communications in Statistics: Simulation and Computation*, 32:1191–1203. DOI: 10.1081/SAC-120023885.

Sidik, K. and Jonkman, J. N. (2006). Robust variance estimation for random effects meta-analysis. *Computational Statistics and Data Analysis*, 50:3681–3701. DOI: 10.1016/j.csda.2005.07.019.

Sidik, K. and Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, 50:1964–1981. DOI: 10.1002/sim.2688.

Simmonds, M. C., Higgins, J. P. T., Stewart, L. A., Tierney, J. F., Clarke, M. J., and Thompson, S. G. (2005). Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials*, 2:209–217.

Simpson, R. and Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *The British Medical Journal*, 2:1243–1246.

Sinha, B. (1985). Unbiased estimation of the variance of the Graybill-Deal estimator of the common mean of several normal populations. *The Canadian Journal of Statistics*, 13:243–247.

Sørensen, H. (2008). Small sample distribution of the likelihood ratio test in the random effects model. *Journal of Statistical Planning and Inference*, 138:1605–1614.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). *BUGS: Bayesian inference using Gibbs sampling (version 0.6.* MRC Biostatistics Unit, Institute of Public Health, Cambridge.

Stanley, T. D. and Doucouliagos, H. (2012). *Meta-Regression Analysis in Economics and Business*, volume 5. Routledge.

Staudte, R. (2010). Book review of 'Multivariate meta-analysis: Modeling the heterogeneity; mixing apples and oranges; dangerous or delicious?', by lidia r. arends, alblasserdam, 2006. *Research Synthesis Methods*, 1:316–318.

Sterne, J. A., editor (2009). *Meta-Analysis in Stata: An Updated Collection from the Stata Journal.* Stata Press.

Stroup, D., Berlin, J., Morton, S., Olkin, I., Williamson, G., Rennie, D., Moher, D., Becker, B., Sipe, A., and Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology. a proposal for reporting. *Journal of the American Medical Association*, 283:2008–2012.

Sutton, A., Abrams, K., Jones, D., Sheldon, T., and Song, F. (2000). *Methods for Meta-Analysis in Medical Research.* John Wiley & Sons.

Sutton, A. J. and Higgins, J. (2008). Recent developments in meta-analysis. *Statistics in Medicine*, 27:625–650.

Takkouche, B., Cadarso-Suarez, C., and Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*, 150:206–215.

Thompson, S., Ekelund, U., Jebb, S., Lindroos, A., Mander, A., Sharp, S., Turner, R., and Wilks, D. (2011). A proposed method of bias adjustment for meta-analyses of published observational studies. *International Journal of Epidemiology*, 40:765–777.

Tippett, L. et al. (1931). The methods of statistics. *The methods of statistics.*

Trikalinos, T. and Olkin, I. (2008). A method for the meta-analysis of mutually exclusive binary outcomes. *Statistics in Medicine*, 27:4279–4300.

van der Tweel, I. and Bollen, C. (2010). Sequential meta-analysis: an efficient decision-making tool. *Clinical Trials*, 7:136–146.

van Houwelingen, H., Arends, L., and Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21:589–624.

van Houwelingen, H., van Zwinderman, K., and Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine*, 12:2273–2284.

Vardi, Y. (1985). Empirical distribution in selection bias models. *The Annals of Statistics*, 13:178 – 203.

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 60:29–60.

Viechtbauer, W. (2009). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, 36:1–48.

Wallace, B. C., Schmid, C. H., Lau, J., and Trikalinos, T. A. (2009). Meta-analyst: software for meta-analysis of binary, continuous and diagnostic data. *BMC Medical Research Methodology*, 9:80.

Welch, B. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38:330–336.

Wetterslev, J., Thorlund, K., Brok, J., and Gluud, C. (2008). Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology*, 61:64–75.

White, I. R. (2009). Multivariate random-effects meta-analysis. *Stata Journal*, 9:40–56.

Whitehead, A. (1997). A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in Medicine*, 16:2901–2913.

Whitehead, A. (2002). *Meta-Analysis of controlled clinical trials*. Applied Statistics. Wiley, Chichester.

Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statistics in Medicine*, 18:2271–2286.

Wilson, E. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*, volume 59. SAGE Publications, Incorporated.

Wolpert, R. and Mengersen, K. (2004). Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science*, 19:450–471.

Xie, M., Singh, K., and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106(493).

Yates, F. and Cochran, W. (1938). The analysis of groups of experiments. *J. Agric. Sci*, 28:410–269.

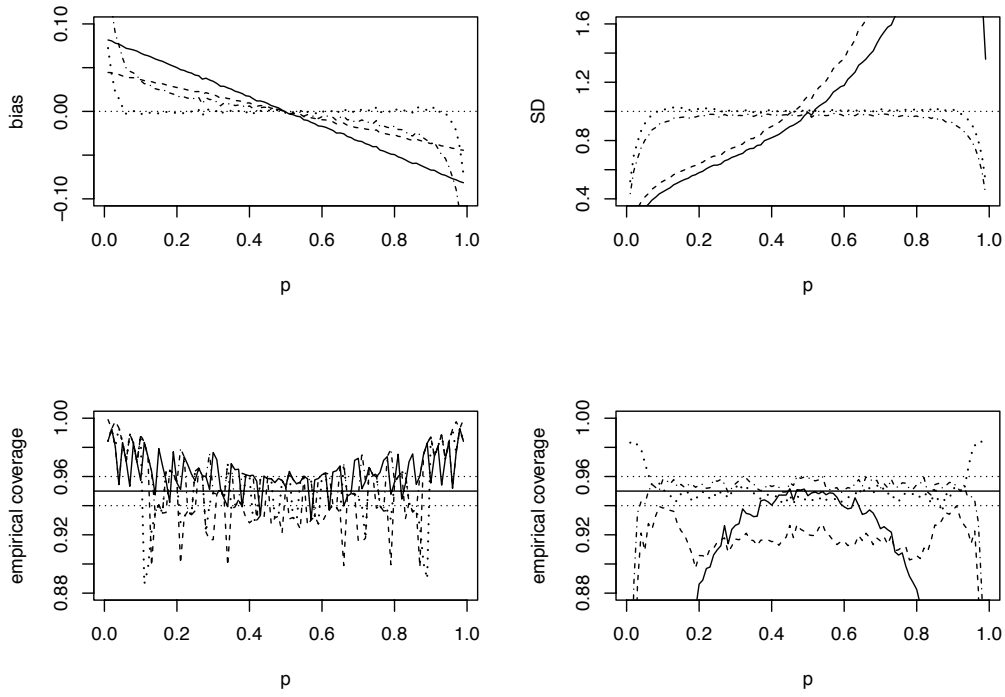

Figure 3: *Simulation study results showing bias and standard deviation of point estimates, together with coverage probabilities for the four intervals under comparison. The AC estimator properties are depicted by solid lines, the BV estimator by dashed lines, the AS (arcsine) estimator by dot-dashed lines and the ASBC (bias-corrected) estimator by dotted lines. The top left plot shows the biases in the centers of the intervals when* $n = 20$ *and* $K = 1$*. Continuing with these parameters, the top right plot depicts the standard deviations of the respective centre estimates. Continuing, the bottom left plot shows the empirical coverages of nominal 95% confidence intervals for* p*. The bottom right plot again shows empirical coverages of 95% confidence intervals, but this time based on a standard meta-analysis that combines 5 studies each of size* $n = 20$*. See text for more details and interpretation.*