Statistical Assessment and Adjustment of Publication Bias in the Cochrane Database of Systematic Reviews

---

Master Thesis in Biostatistics (STA495)

by

Giuachin Kreiliger
12123832

supervised by

Dr. Simon Schwab
Prof. Dr. Leonhard Held

Zurich, July 27, 2019

Statistical Assessment and Adjustment of Publication Bias in the Cochrane Database of Systematic Reviews

Giuachin Kreiliger

Version July 27, 2019

Contents

Preface

Giuachin Kreiliger
June 2019

Chapter 1

Introduction

Studies get more attention and are more likely to be published, read and cited if they contain significant effects. Studies with no evidence for an effect are less likely to get published. This generates a bias called "publication bias", a distorted view of the evidence for an effect. Recently, publication bias has been identified as one of the major concerns in irreproducible research (Bishop, 2019). The issue has been discussed in clinical science by many researchers (Dickersin *et al.* (1987), Sterne *et al.* (2001)). Consensus is that publication bias exists in clinical science.
In medical research, clinical trials study the efficacy of therapies and drugs. The gold standard in such intervention studies are randomized controlled trials (RCTs). Results from RCTs influence the treatment of patients in daily clinical practice. The results from multiple intervention studies can be summarized in a meta-analysis to estimate an overall treatment effect (access all studies). However, publication bias can bias the overall effect estimates from meta-analyses and eventually lead to ineffective treatments that could lead to patient harm, distortion and financial expenses. There is extensive literature on publication bias in meta-analyses, e.g. (Jones *et al.* (2013), Turner *et al.* (2008), Egger *et al.* (2003), McAuley *et al.* (2000)). The authors agree on that the exclusion of unpublished results in meta-analyses is responsible for overestimation of treatment effects. Although there are policies that make it mandatory to make all study results publicly accessible (Law, 2007), it is not clear if the situation has improved yet.
There are multiple ways to assess the amount of publication bias. For instance, it is possible to follow studies and assess if they are getting published depending on their findings (Decullier *et al.*, 2005). One often finds that positive findings (*i. e.* large effects) are reported and published more often (see also an example in social sciences: Franco *et al.* (2014)). Another way is to compare results in study registries with results published in journals (*e. g.* Jones *et al.* (2013)). Again one finds systematic differences between published and unpublished results.
A third way is to assess the so-called small study effect in a meta-analysis, that is, smaller studies sometimes showing different, often larger treatment effects than large ones (cite sb). Importantly, this method does not assess publication bias directly. But the estimation of small study effects is oftentimes the most efficient way to investigate publication bias in a large number of meta-analyses. However, evidence for small study effects can oftentimes be interpreted as evidence for publication bias (Egger *et al.*, 1997), as this is the most likely cause, even though there may also be other explanations. The underlying rationale is that journal editors are guided by two simple rules (Ioannidis and Trikalinos (2007b), Ioannidis and Thombs (2019)):

- Publish new and trend-setting findings that lead to many citations, increasing the journal impact factor (JIF).
- Publish large trials that are likely to set the standard for a given scientific question, again to increase the JIF.

This then leads to an asymmetry in published studies; small studies show larger effect sizes than larger studies do. A funnel plot(cite sb) allows visual inspection of small study effects by

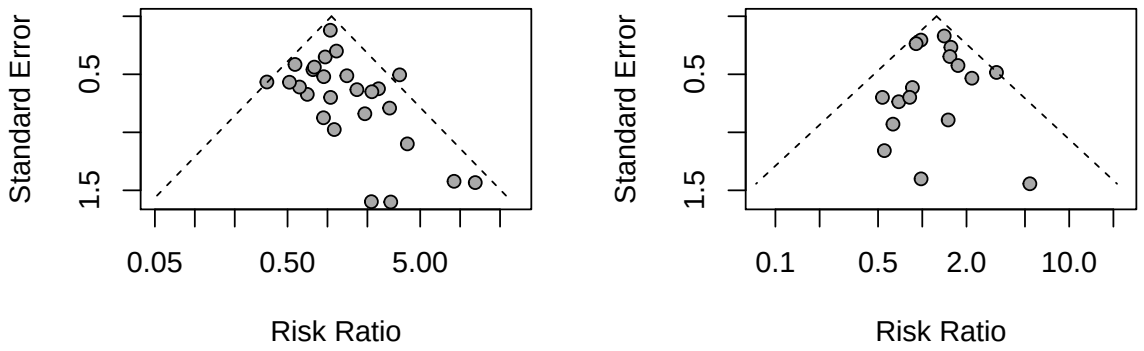

Figure 1.1: Funnel plots of two meta-analyses: On the left, the improvement in depression syndroms after application of tricyclic antidepressants is compared to placebo. The meta-analysis on the right measures the occurence of intracranial haemoerrhage by CT after application of any anti-thrombolytic agent. All studies are RCT's.

plotting the effects against their standard error. For illustration purposes, one meta-analysis with large funnel plot asymmetry and one with no asymmetry is shown in Figure **??**. By means of simple linear regression, one can investigate how much evidence there is for this asymmetry.

The purpose of a funnel plot is to visualise if an how the estimated treatment effects converge to a common "real" treatment effect value (here, the weighted mean treatment effect with weights = inverse variances). This is also shown by the triangle depicted by the dotted line (the spire is at the weighted mean).
We see that while the studies in the meta-analysis on the left are not symmetrically distributed, but rather accumulate at the left side, they seem to be more evenly distributed in the left triangle. From this, we would ultimately conclude that some sort of bias or heterogeneity is distorting the estimate of the overall treatment effect.
The Cochrane Organisation has specialized on systematic reviews of healthcare interventions. Researchers that write a systematic review collect data across studies, review them and try to provide up-to-date information about specific treatment efficacies (Higgins JPT, 2011). By extensive literature scrutinization, they try to circumvent the issue of publication bias. They do not apply funnel plot asymmetry tests to analyse publication bias. Earlier research however suggests that the efforts are only partially succesful, and that there still is publication bias within the reviews (Egger *et al.* (1997), Ioannidis and Trikalinos (2007a), Kicinski *et al.* (2015), van Aert *et al.* (2019)). In these publications, Cochrane systematic reviews is analysed with methods to detect publication bias, for example small study effect tests or bayesian hierarchical selection models (Kicinski *et al.*, 2015). They all find moderate to large evidence for publication bias in the Database. Their results will be compared to the results of this study in the last chapter.

1.1 Aim of the Study

None of the research so far has estimated the amount and impact of publication bias on meta-analytical findings thoroughly and with the most suitable methods. Also, the results are ironically often presented in the form of dichotomized hypothesis tests, a practice that is partly responsible for publication bias.

The aim of this thesis is to use prevailed methods to detect publication bias, and make use of the full amount of data that the Cochrane Organisation provides. At the end, an approximate, up-to-date estimate of the prevalence of publication bias in the data shall be given. To achieve this, methods to detect and adjust for publication bias in meta-analysis are applied on the data. It is also possible to adjust for publication bias with suitable methods. These methods are applied on the dataset as well, to achieve publication bias adjusted treatment effect estimates. It will be shown if and to what extent treatment effects are overestimated due to publication bias in the Cochrane systematic reviews.

Chapter 2

Methods

2.1 Introduction and Notation

The analysis has already been said to be restricted on clinical or health care interventions. The interventions are restricted to comparisons of two treatment groups by some measure of melioration or worsening of health. The difference in this measure between the groups is referred to as the treatment effect. Where it is not particularly mentioned, the term treatment effect refers to any effect measure such as log risk ratio, log hazard ratio, Cohen's d or standardized mean difference, Fisher's z score or Pearsons correlation coefficient.
Let us consider a meta-analysis with n study treatment effects ($n > 1$, but typically small). A study is indexed by i, and it's treatment effect by θ_i. The observed treatment effect is $\hat{\theta}_i$. The pooled treatment effect of a meta-analysis will be denoted as θ_M, and consequently, the observed pooled treatment effect as $\hat{\theta}_M$. Furthermore, each treatment effect is typically measured with some standard error se_i and an estimate of se_i is denoted as $\hat{\text{se}}_i$. The $\hat{}$ sign thus indicates if it is an estimate.
For continuous outcomes, let m_t be the mean of the treatment group, m_c the mean of the control group, and equivalently sd_t and sd_c the corresponding standard deviations. n_t and n_c are the total number of participants in the groups. In the case of binary outcomes, let e_t be the count of events in the treatment arm e_c the events in the control group. The observed counts in a study i are referred to as $e_{t,i}$ and analogously $e_{c,i}$.

2.2 Effect Measures and p-values

2.2.1 Continuous Outcomes

For given $(m_t, m_c), (\text{sd}_t, \text{sd}_c)$ and (n_t, n_c), one can compute mean difference as well as a standardized mean difference (here: Cohen's d) and a standard error thereof. The mean difference can θ and its standard error se can be obtained as

$$\theta = m_t - m_c \qquad \text{se} = \sqrt{\text{sd}_t^2/n_t + \text{sd}_c^2/n_t} \tag{2.1}$$

The standardized mean difference d and its standard error se can similarly be obtained by

$$\text{se} = \sqrt{\frac{(n_t-1)\text{sd}_t^2 + (n_c-1)\text{sd}_c^2}{n_t + n_c - 2}} \qquad g = \frac{m_t - m_c}{\text{se}} \tag{2.2}$$

Both estimators take into account that the two groups might have unequal variances. A p-value to test the null hypothesis that the mean between group is equal is commonly obtained with the Students t test. The t statistic is obtained, using se and d from (2.2), by

$$t = d/(\text{se}\sqrt{(1/n_t) + (1/n_c)})$$

and the p-value can be obtained with the cumulative Student's t-distribution F with $n_t + n_c - 2$ degrees of freedom:

$$p = 2(1 - F(|t|))$$

The t-test is known to be not very reliable if combined sample size is small ($n < 30$), see for example Kasuya (2001).

2.2.2 Binary Outcomes

Two commonly used effect measures for binary outcome data are risk ratios and odds ratios between treatment and control group. The methods presented here can also be found, for example, in (Borenstein *et al.*, 2011, 34). Let θ be the natural logarithm of the odds ratio. θ and its variance se^2 can be obtained by

$$\hat{\theta} = \log(\frac{e_t \cdot (n_c - e_c)}{e_c \cdot (n_t - e_t)})$$
$$\text{se}^2 = 1/e_t + 1/(n_t - e_t) + 1/e_c + 1/(n_c - e_c)$$

Plugging in the observed counts will give the corresponding estimates. The logarithm of the risk ratio θ and its variance se^2 is similarly defined as

$$\hat{\theta} = \log(\frac{e_t/n_t}{e_c/n_c}) \tag{2.3}$$
$$\text{se}^2 = 1/e_t - 1/n_t + 1/e_c - 1/e_t \tag{2.4}$$

Using likelihood theory, one could show that the estimators are maximum likelihood estimators and that one can use the asymptotic normal distribution of the maximum likelihood estimator to calculate a p-value, *e. g.* (Held and Sabanés Bové, 2014, 98).
Thus, with Φ as the cumulative standard normal distribution, we get

$$p = 2 \cdot (1 - \Phi(\left|\hat{\theta}/\hat{\text{se}}\right|),$$

a p-value for the corresponding estimate, which summarizes the evidence against $\hat{\theta}$ being zero (*i. e.* the true risk/odds ratio being 1).
Binary and continuous effect measures can be converted into each other, as described in Section 2.6.

2.2.3 Survival Outcomes

Time-to-event data with censoring has to be analyzed by special means. One frequently used method to take into account right-censoring is the Cox proportional hazards regression model (Cox, 1972). Because the method itself is not applied in this thesis, but only the resulting estimates of the parameters are used, the reader is referred to the extensive literature covering this topic (*e. g.* Cox and Oakes (1984)).
The so-called hazard ratio estimated by Cox regression is the ratio of the instantaneous risk of experiencing the event between two groups. Because it is a maximum likelihood estimator, one

can again use its Wald test statistic to test for equal hazards. Let $\hat{\theta}$ be an estimate of the log hazard ratio and $\hat{\text{se}}$ an estimate of the standard error of it. As before

$$p = 2 \cdot (1 - \phi(\left|\hat{\theta}/\hat{\text{se}}(\hat{\theta})\right|)$$

will give a p-value for the evidence against the null hypothesis.

2.3 Fixed and Random Effects Meta-Analysis

The fixed effects meta-analysis estimator of the pooled treatment effect is a mean of the single treatment effect estimators, weighted by their standard errors (Rosenthal and Rubin, 1982). Let $w_i = 1/\,\text{se}_i^2$ be the weights, and θ_M be the pooled estimator and s_M^2 its variance. Then

$$\theta_M = \frac{\sum_{i=1}^n w_i \theta_i}{\sum_{i=1}^n w_i} \qquad\qquad s_M^2 = \frac{1}{\sum_{i=1}^n w_i} \tag{2.5}$$

This estimator minimizes the variance between the effects. An estimate $\hat{\theta}_M$ can be obtained by plugging in the observed treatment effects and variances $\hat{\theta}_i$ and $\hat{\text{se}}_i^2$. The underlying idea is that we assume $\theta_i \sim N(\theta_M, \text{se}_i^2)$, θ_M being the true effect, all θ_i being distributed around an equal mean.
The random effects model (Whitehead and Whitehead, 1991) assumes instead that

$$\theta_i \sim N(\mu_i, \text{se}_i^2) \qquad\qquad \mu_i \sim N(\theta_M, \tau^2) \tag{2.6}$$

Marginally, we have θ_i being distributed around a common mean θ_M with additional variance τ^2:

$$\theta_i \,|\, \mu_i \sim N(\theta_M, \text{se}_i^2 + \tau^2)$$

τ^2 is often referred to as a population variance or between-study variance, whereas se_i^2 can be interpreted as sampling error. The pooled treatment effect estimate θ_M of the random effects model and its variance is obtained by replacing the weights w_i in equation (2.5) with $w_i = 1/(\text{se}_i^2 + \tau^2)$.
The model is superior to the fixed effects model whenever the standard errors of the treatment effects alone are unlikely to fully account for the entire variability observed between studies. Note that as τ^2 increases, each θ_i will eventually get equal weights, irrespective of it's sampling error se_i^2.
The estimation of τ^2 has been subject to some debate in the statistical literature. Oftentimes, the method of moment estimator of DerSimonian and Laird (1986) is used. We use the measure of heterogeneity, Q, and divide by C after having subtracted the degrees of freedom $n - 1$:

$$Q = \sum_{i=1}^n w_i (y_i - \theta_M)^2 \qquad\qquad C = \sum_{i=1}^n w_i - \frac{\sum_{i=1}^n w_i^2}{\sum_{i=1}^n w_i} \tag{2.7}$$

$$\tau^2 = \max\left(0, \frac{Q - (n-1)}{C}\right) \tag{2.8}$$

Again, $w_i, \theta_i ...$ have to be replaced by their estimates in order to get an estimate $\hat{\tau}^2$.
The Paule-Mandel estimator (Paule and Mandel, 1982) is considered to have most often better properties than the method of moments estimator (*e. g.* Veroniki *et al.* (2016)). Since we defined $w_i = 1/(\text{se}_i^2 + \tau^2)$, it also holds that

$$w_i \mathrm{Var}(\theta_i) = 1 \qquad\qquad \mathrm{Var}(\sqrt{w_i}\theta_i) = 1$$

For any w_i, the variance can be estimated and equated to its expected value:

$$\mathrm{se}^2(w_i\theta_i) = \frac{\sum_{i=1}^{n} w_i(\theta_i - \theta_M)^2}{n-1} \qquad\qquad \frac{\sum_{i=1}^{n} w_i(\theta_i - \theta_M)^2}{n-1} = 1 \tag{2.9}$$

θ_M can be estimated using equation (2.5), the only problem is to estimate τ^2. It can be obtained through an iterative process, using a newly defined function

$$F(\tau^2) = \sum_{i=1}^{n} w_i(\theta_i - \theta_M)^2 - (n-1)$$

In view of equation (2.9), τ^2 must be such that $F(\tau^2) = 0$. Then, we start with a arbitrary τ^2 and repeatedly add a term τ_0^2 to update τ^2 until $F(\tau^2 + \tau_0^2)$ is close to zero (using $\tau^2 + \tau_0^2$ for $\hat{w}_i, \hat{\theta}_M$). Using a truncated Taylor series expansion, one can obtain the partial derivative after τ^2, which is a reasonable choice for τ_0^2.
The estimation of τ^2 is accompanied by uncertainty. A common procedure is to test if it is there is significant heterogeneity between the studies (Borenstein *et al.*, 2011, 109). Compute Q, as given in (2.7), based on one of the estimators of τ^2. It is assumed that Q follows a central Chi-squared distribution with $n-1$ degrees of freedom under the null hypothesis of equally distributed effect sizes. Thus, the expected value of Q is $n-1$, and the excess dispersion is $Q - n + 1$. The p-value against the null hypothesis of equally distributed effect sizes is $1 - F(Q)$, using F as the cumulative distribution function of the Chi-squared distribution with d.f. $= n-1$.
An advantage of the τ^2 is that it is directly linked to the variability in the data. Additionally, one can use the I^2 statistic to see what portion the between study variance has of the overall variance. It is computed as

$$I^2 = \max\left(0, 1 - \frac{n-1}{Q}\right)$$

Importantly, all proposed methods above assume normally distributed effect sizes and proper estimates $\hat{\mathrm{se}}$ of the true standard error. This assumptions are not met for very small sample sizes and very few event counts. Alternatively, the Mantel-Haenszel method for risk and odds ratios (see *e. g.* Fleiss *et al.* (2013)) could be used in the latter case.

2.4 Small Study Effects Tests

The tests that will be presented on the following pages are a common way to detect publication bias. Importantly, they are however not interpretable directly as evidence for publication bias, but this is left for the discussion chapter. The tests are also often referred to as funnel plot asymmetry tests, because of the popularity of a recent test that has been used frequently to test and adjust for publication bias, which goes under the name of trim-and-fill (Duval and Tweedie, 2000). However, it is known for some time that the method has disadvantageous properties, therefore, it will not be discussed here, as well as the funnel plot (the radial plot will be used as an alternative).
A test for small study effects will test in some ways if the size of the estimated treatment effect of a study depends on some measure of its size. An association between study size and treatment effect size can be interpreted as an artifact of publication bias.

2.4.1 Continuous Outcome Tests

For a continuous outcomes that are normally distributed, the sample mean and variance are independent of each other (Schwarzer *et al.*, 2015, 120). Thus, the estimated standard errors of treatment effects can be used as a proxy for study size, as they should in principle not be tied to the effect size.

Begg and Mazumdar: Rank Correlation Test

Begg (1988) proposed a rank based test to test the null hypothesis of no correlation between effect size and variance. A standardized effect size $\theta_i^\star$ can be computed as in (2.10). $\text{se}_i^{2\star}$ is the variance of $\theta_i - \theta_M$ as defined in (2.11), θ_M being the fixed effects pooled treatment effect ((2.5).

$$\theta_i^\star = (\theta_i - \theta_M)/\,\text{se}_i^{2\star} \tag{2.10}$$

$$\text{se}_i^{2\star} = \text{se}_i^2 - 1/\sum_{i=1}^{n} \frac{1}{\text{se}_i^2} 1 \tag{2.11}$$

A rank correlation test based on Kendall's tau is then used. First, the pairs are ordered after their ranks based on $s^{2\star}$. Then, for each $s^{2\star}$ rank, the corresponding ranks based on $\theta^{2\star}$ that are larger are counted and summed up to u. The number of ranks based on $\theta^{2\star}$ that are in contrary, smaller, are counted and summed up to l. Then the normalized test statistic Z is given as

$$Z = (u - l)/\sqrt{n(n-1)(2n+5)/18}$$

Thus, large number of concordance between pairs will reflect in large $\hat{u}$ and small $\hat{l}$ and thus lead to a large $\hat{Z}$. The p-value is obtained using the standard normal distribution Φ:

$$p = 2 \cdot (1 - \Phi(|Z|))$$

The changes that have to be made int the case of ties are small and can be found in (Begg, 1988, 410).

Egger's Test: Weighted Linear Regression Test

First, the concept of simple linear regression is introduced. In short, the model assumes a dependent variable y to be a linear function of another explanatory variable x:

$$y = \beta_0 + \beta_1 x + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{2.12}$$

ϵ is the residual noise term that becomes necessary when n pairs (x_i, y_i) are given and there is no exact solution. Then it is common to look for the solution that minimizes the squared residuals, the least-squares solution. Formally,

$$\underset{\beta_0,\beta_1}{\text{argmin}} \left(\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i \right) \tag{2.13}$$

Let $\mathbf{X}$ be a matrix with the explanatory variables x and $\mathbf{y}$ a corresponding vector for all y:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} \\ 1 & x_{22} \\ 1 & x_{32} \\ \vdots & \vdots \\ 1 & x_{n2} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

Let $\beta = (\beta_0, \beta_1)^\top$. It can be shown that

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{2.14}$$

Is an estimator of β that minimizes the squared residuals. Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ and $\hat{\mathbf{r}} = \hat{\mathbf{y}} - \mathbf{y}$. The variance estimates $\hat{\sigma}^2$ and $\hat{\mathbf{s}}^2_{\hat{\beta}}$ are

$$\hat{\sigma}^2 = \frac{1}{n-2} \mathbf{r}^\top \hat{\mathbf{r}} \qquad\qquad \hat{\mathbf{s}}^2_{\hat{\beta}} = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} \tag{2.15}$$

The estimate $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ the slope of the regression line. Furthermore, in the simple linear regression setting, $\hat{\beta}_0$ can also be obtained by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\bar{x}, \bar{y}$ denoting the sample means of the corresponding values $x_1, .., x_n$ and $y_1, ..., y_n$. Thus, $\hat{\beta}_0$ is also called global mean. To test whether there is evidence for the intercept β_0 to be unequal to some value β_{H0}, a t-test can be used.

$$p = 2(1 - F(|(\beta_0 - \beta_{H0})/s_{\beta_0}|))$$

where F is the cumulative t distribution with $n - 2$ degrees of freedom. p will give the evidence against the null hypothesis $\beta_0 = \beta_{H0}$.
The concept is extendable to weighted linear regression. Weighted linear regression may be used if the residuals $\mathbf{r}$ have unequal variances, which is equivalent to ascribe different precision to the observed y. The least squares equation (2.13) is extended to

$$\underset{\beta_0, \beta_1}{\operatorname{argmin}} (\sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i))$$

with positive weights w_i that penalize large squared residuals for some i more if w_i is larger compared to other w_i.
Let $\mathbf{W}$ be a $n \times n$ matrix with $\mathbf{W}_{ii} = w_i$, the weights on the diagonal and zeros on the off-diagonals. The estimates in (2.14) and (2.15) can still be used if $\mathbf{X}$ is exchanged with $\mathbf{X}^\star = \mathbf{W}\mathbf{X}$ and $\mathbf{y}$ with $\mathbf{y}^\star = \mathbf{W}\mathbf{y}$.
Now it is shown how linear regression can be used to test dependency of effect sizes on study sizes. The simplest application was introduced by Egger *et al.* (1997). Let θ/s be the dependent variable y and $1/\operatorname{se}$ the explanatory variable x. If plotted, this corresponds to a radial or Galbraith plot (Galbraith, 1988). The linear regression equation as introduced before in (2.12) can be written in two ways:

$$\theta/s = \beta_0 + \beta_1/\operatorname{se} + \epsilon, \qquad\qquad \epsilon \sim N(0, \sigma^2) \tag{2.16}$$

Equation (2.16) is often provided due to the correspondence to the radial plot. However, it is equivalent to

$$\theta = \beta_0 + \beta_1 \operatorname{se} + \epsilon, \qquad\qquad \epsilon \sim N(0, w^{-2}\sigma^2) \tag{2.17}$$

with weights $w = 1/\operatorname{se}^2$. Thus testing β_0 of (2.16) or β_1 is equivalent. The corresponding p-value is then used as evidence for a small study effect. Plugging in $\theta_i/\operatorname{se}_i, 1/\operatorname{se}_i$ as y_i, x_i into equation (2.14) and (2.15) will give the estimates for $\hat{\beta}_0, \hat{\beta}_1, \hat{\operatorname{se}}_{\beta_0}$ and $\hat{\operatorname{se}}_{\beta_1}$.

Thompson and Sharp's Test: Weighted Linear Regression Random Effects Test

A method proposed in Thompson and Sharp (1999) allows for between study variance τ^2, as introduced before in section 2.3. It extends the previously seen linear regression approach with $x = 1/\operatorname{se}$ and $y = \theta/s$ by introducing weights. The effect size θ_i is assumed to be distributed as

$$\theta_i \sim N(\beta_0 + \beta_1 \operatorname{se}_i, \operatorname{se}_i^2 + \tau^2) \tag{2.18}$$

τ^2 is estimated as in equation (2.8) (method of moments). The weights are set as $w_i = 1/\sqrt{\operatorname{se}_i^2 + \tau^2})$. After adjusting for the weights as described in 2.4.1, we can proceed analogous to Egger's test. The p-value for $\beta_0 \neq 0$ reflects the evidence for a small study effect.

2.4.2 Generalized linear model

The introduction of group-sepific random effects allows to analyze grouped or repeated measurements by linear regression (McCullagh and Nelder, 1989). Let i be the group index, and

$$\mathbf{Y}_{ij}|U_i, \epsilon_{ij} = (1, {x_i}^\top)\beta + U_i + \epsilon_{ij} \tag{2.19}$$

the marginal distribution of $\mathbf{Y}_{ij}$ depending on U_i, ϵ_{ij}. The random effects $U_i \sim N(0, g)$ and the residual error term $\epsilon_{ij} \sim N(0, \tau^2)$ are considered independent. If there are more groups in the data we can extend (2.19) by adding a second random term Q with an index k:

$$\mathbf{Y}_{ij}|U_i, Q_k\epsilon_{ij} = (1, x_i^\top)\beta + U_i + Q_k + \epsilon_{ij} \tag{2.20}$$

Again, U, K, ϵ are distributed independently. The random intercepts model will account for a nested group structure by accordingly indexing the groups.
The conditional distribution in equations (2.19) and (2.20) can be used to construct a conditional likelihood; A bayesian posterior distribution is integrated with respect to the random effects, such that it can be used for restricted or unrestricted maximum likelihood estimation.

2.4.3 Dichotomous Outcomes Tests

The issue with dichotomous outcomes is that effect size and variance of effect size are correlated, which can readily be seen in (2.3) and (2.4). For example, a small number of event counts in one or group will inflate the variance and the effect size. Consequently, the tests above will tend to reject the null-hypothesis too often, *i. e.* report false positives. A number of solutions to this problem are provided in the literature.

Peters Test: Weighted Linear Regression Test

Instead of taking the standard error s as explanatory variable x as in Egger's Test, the inverse of the total sample size is used. Additionally, the variances se_i^2 are used as weights. Thus, the subsequent test procedure is identical to Egger's test. Peters test is a a small modification of Macaskill's test where the explanatory variable is the sample size instead of its inverse.

Harbord's Test: Score based Test

A rank based alternative to Peters test for binary outcomes is Harbord's test (Harbord *et al.*, 2006). It uses a different treatment effect and variance estimate: the score φ of the log-likelihood, evaluated at log odds ratio $\theta_0 = 0$ and its inverse Fisher information s^2. Formally,

$$\varphi = e_t - (e_t - e_c)(e_t + (n_t - e_t))/(n_t + n_c) \tag{2.21}$$

$$s^2 = \frac{(e_t + e_c)(e_t + (n_t - e_t))(e_c + (n_c - e_c))((n_t - e_t) + (n_c - e_c))}{(n_t + n_c)^2(n_t + n_c - 1)} \tag{2.22}$$

It can be shown that they are both good approximations of the log odds ratio and its variance if the real θ is not too far from zero. The standardized estimator r_i/v_i is also known as Peto odds ratio. The obtained scores and variances can be used in Egger's test as treatment effects and variances.

Schwarzer's Test: Rank Correlation Test

Schwarzer *et al.* (2007) developed a test for the correlation between $E_t - \mathbb{E}(E_t)$ and the variance of E_t, E_t being a random variable from the non-central hypergeometric distribution, assuming a fixed log odds ratio.
$\mathbb{E}(E_t)$ and variance of E_t are then estimated based on e_t. The standardized cell count deviation

$$(e_t - \mathbb{E}(E_t))/\sqrt{(\,\mathrm{se}_i^2)} \tag{2.23}$$

and the inverse of se_i^2 is then used as before in Begg and Mazumdar's test.

Rücker's Test: Using the Variance Stabilizing Transformation for Binomial Random Variables

The correlation between variance and effect size of dichotomous outcome measures can be abolished by the variance stabilizing transformation for binomial random variables. We use that the arcsine function is the variance stabilizing transformation for a proportion. Let

$$\theta_i = \arcsin e_t/n_t - \arcsin e_c/n_c \qquad\qquad \mathrm{se}_i^2 = 1/4n_t + /4n_c$$

Then one can optionally apply Begg and Mazumdar's rank correlation test or Thompson and Sharp's test using the newly obtained estimates.

2.4.4 Excess Significance Test

Ioannidis and Trikalinos (2007b) developed an exploratory test to detect if the proportion of significant findings is larger than expected. Note that significance is here defined as one-sided significance, with the direction of the test being the same for each study.
We assume that the effects are equally distributed around a true mean effect θ_M, which can be estimated by fixed effects Meta-Analysis (2.5). Let O be the number of significant study results out of n studies and α the significance threshold. Corresponding to the study effect θ_i, we can specify the power $1-\beta_i$, the probability to accepting a true result. Let $z_{\alpha,i}$ be the $1-\alpha$ quantile of a normal distribution with standard error $\hat{\mathrm{se}}_i$. The power of study i can be estimated as:

$$1 - \hat{\beta}_i = F(z_\alpha) \tag{2.24}$$

with F being the cumulative distribution of a normal with mean $\hat{\theta}_M$ and standard deviation $\hat{\mathrm{se}}_i$.
If we assume no bias in θ_i and θ_M, the expected number of significant study results is then just

$$E = \sum_{i=1}^{n} (1 - \beta_i)$$

E can then be compared to O by constructing a test statistic χ:

$$\chi = \left(\frac{(O-E)^2}{E} + \frac{(O-E)^2}{n-E}\right)$$

and consecutively, calculating a p-value for the evidence against the null-hypothesis of $O = E$ with a χ^2 distribution with one degree of freedom. Alternatively, one can also use a binomial test, which is encouraged when n and O is small. We will get a one sided p-value for excess significance, $\mathsf{Pr}(X \geq O)$, by

$$p = \sum_{i=O}^{n} \left(\binom{n}{i} p^i (1-p)^{n-i} \right) \tag{2.25}$$

with $p = E/n$ and X being a binomial random variable with probability p.

2.5 Small Study Effect and Publication Bias Adjustment

There are different approaches to correct for small study effects and publication bias. They can mainly be distinguished by their underlying methods: regression based approaches aim to regress the effect to a study with infinite precision (*i. e.* very small standard error) or to a summary effect, corrected for publication bias. Selection models aim to simulate different, hypothetical selection processes and attain a approximate treatment effect by sensitivity analysis:

2.5.1 Adjustment by Regression

Rücker *et al.* (2011) use a random effects model to obtain an unbiased estimate. Similarly to regression based tests for small study effects, we have

$$\theta_i = \beta_0 + \beta_1 \sqrt{\mathrm{se}_i^2 + \tau^2} + \epsilon_i \sqrt{v_i + \tau^2}, \epsilon_i \overset{\text{iid}}{\sim} N(0,1) \tag{2.26}$$

The only difference between Thompson and Sharp's variant is $x = \sqrt{\mathrm{se}^2 + \tau^2}$ instead of $x = \sqrt{\mathrm{se}^2}$. β_1 represents the bias introduced by small study effects, as illustrated in the following equations:

$$\mathbb{E}((\theta_i - \beta_0)/\sqrt{\mathrm{se}_i^2}) \to \beta_1 \text{ if } \mathrm{se}_i \to \infty$$
$$\mathbb{E}(theta_i) \to \beta_0 + \beta_1 \tau \text{if} \, \mathrm{se}_i \to 0$$

After estimating τ^2, one can estimate β_0 and β_1 as seen before in the simple linear regression framework. Now we have basically two possible estimates at hand:

- β_0 the treatment effect without any influence of study precision with standard error s_{β_0}
- $\beta_0 + \beta_1 \tau$ the treatment effect of a hypothetical study with infinite precision, corresponding standard error $\mathrm{se} = s_{\beta_0} + s_{\beta_1}$

Simulations in Rücker *et al.* (2011) suggested that the latter estimate is slightly superior with respect to coverage (and mean squared error).

2.5.2 Copas Selection Model

A method proposed in Copas and Shi (2001, 2000); Copas and Malley (2008) assumes that the given sample of treatment effects and standard errors is a selected part of a larger random sample. Selection of studies depends on their effect size and variance. Smaller variance is always accompanied by larger selection probability.
Let θ_i be the effect size estimate of study i. Then

$$\theta_i \sim N(\mu_i, \sigma_i^2)\mu_i \sim N(\theta, \tau^2) \tag{2.27}$$

which is similar to the random-effects meta-analysis setting. θ is the population mean effect, σ_i^2 the within study variance and τ^2 the between study variance. Equations in (2.27) are termed the *population model*.
The *selection model* is defined as follows. Suppose a selection of studies with reported ($\neq$ estimated) standard errors s (likely different from σ). Only a proportion of the selection will be published, with a defining the overall proportion of published studies and b (assumed to be positive) defining how fast this proportion increases with s becoming smaller. Formally,

$$P(\text{select} \mid s) = \Phi(a + b/s)$$

The equation can be rewritten as

$$z = a + b/s + \delta$$

with $\delta \sim N(0, 1)$. z is interpreted as the *propensity for selection*. It's sign must be positive in order for the study to be selected. Thus, the larger z, the more likely the study will be selected. Also, a is somewhat like to global retention or selection rate for each study, while b decides about the decline of selection probability with increasing s.
So far, we have, for a study i

$$\begin{aligned} \theta_i &= \mu_i + \sigma_i \epsilon_i \\ \mu_i &\sim N(\theta, \tau^2) \\ z_i &= a + b/\operatorname{se}_i + \delta_i \end{aligned}$$

where (ϵ_i, δ_i) are standard normal residuals. The two models are coupled by introducing a correlation $\rho = cor(\theta_i, z_i)$ by defining (ϵ_i, δ_i) as bivariate standard normals. It follows that, if ρ_i is large and positive and $z_i > 0$, then the estimate of a study i that is selected is likely to have positive ϵ_i and δ_i and thus, the true mean μ is likely to be overestimated.
Let $u_i = a + b/\operatorname{se}_i$, $\lambda(u_i)$ the Mill's ratio $\phi(u_i)/\Phi(u_i)$ (ϕ is the standard normal density function and Φ the cdf) and $\tilde{\rho}_i = \sigma/\sqrt{(\tau^2 + \sigma_i^2)}\rho_i$. The probability of a study being selected, given se_i and θ_i, is

$$P(\text{select} \mid \operatorname{se}_i, \theta_i) = \Phi\Big(\frac{u_i + \tilde{\rho}_i((\theta_i - \mu)/\sqrt{(\tau^2 + \sigma_i^2)})}{\sqrt{1 - \tilde{\rho}_i^2}}\Big)$$

Which again shows that larger se_i and θ_i lead to a larger selection probability. It can also be shown that the expected value

$$\mathbb{E}(\theta_i \mid \operatorname{se}_i, \text{select}) = \mu + \rho_i \sigma_i \lambda(u_i) \tag{2.28}$$

which shows that the expected value for a study is larger for larger σ.
A likelihood for θ_i, conditional on $z > 0$ can be formulated to estimate the parameters of the model. Regarding a and b, there is no way that they can be estimated because the number of missing studies and their effect sizes is not known. Instead, fixed values for a and b have to be chosen. The nuisance parameter σ_i can be replaced by an estimate if the sample size in the studies is large enough. Since

$$\mathrm{Var}(\theta_i \mid \mathrm{se}_i, z_i > 0) = \sigma_i^2(1 - c_i^2\rho_i^2)$$

with $c^2 = \lambda(u_i)(u_i + \lambda(u_i))$, we can replace σ_i^2 by $\hat{\sigma}_i^2 = \frac{1}{1-c_i^2\rho_i^2}$. Although one has to evaluate the likelihood for fixed pairs (a, b), one can compare the fit of the model to evaluate which one is more suitable: With equation (2.28), one can obtain fitted values of θ_i based on se_i. Also, for two different pairs (a, b), $(a^\star, b^\star)$,

$$\mathbb{E}(\theta_i \mid z_i > 0, a^\star, b^\star) - \mathbb{E}(\theta_i \mid z_i > 0, a, b) \approx c^\star + \rho(\lambda(a^\star) - \lambda(a))\,\mathrm{se}_i$$

and that local departures of θ_i can be approximated by adding a linear term in se_i to the expectation of θ_i. Thus, to test a pair (a, b), it is sufficient to test $\beta \neq 0$ in

$$\theta_i = \theta + \beta\,\mathrm{se}_i + \sigma_i\epsilon_i$$

with restriction that $\rho \geq 0$. To test some pair (a, b) against the scenario with no selection, we set $a^\star = \infty$ (or $\rho = 0$). A likelihood ratio test will give a test statistic to test against $H0 =$ no selection:

$$\chi^2 = 2 \cdot (\max_{\theta,\tau,\beta} \tilde{L}(\theta, \tau, \beta) - \max_{\theta,\tau} \tilde{L}(\theta, \tau, 0)) \tag{2.29}$$

with

$$\tilde{L}(\theta, \tau, \beta) = -\frac{1}{2}\sum_{i=1}^{n}[\log(\tau^2 + \sigma_i^2) + \frac{(\theta_i - \theta - \beta\,\mathrm{se}_i)^2}{(\tau^2 + \sigma_i^2)}]$$

χ^2 can be used with a χ^2 distribution with one degree of freedom to obtain a p-value. Note that the test is almost equivalent to Egger's small study effect test with $\tau^2 = 0$. Thus, although the copas selection model models publication bias, it is dependent on the small study effects to find the most suitable pair (a, b).
If one applies the model to a single meta-analysis, a sensitivity analysis is suggested. One can observe how θ and it's confidence intervals change dependent on the underlying selection process. Selection models are in general not recommended for inference (*e. g.* McShane *et al.* (2016)). Rücker *et al.* (2011) have shown how the method can be implemented in a simulation for inference purposes.
A range of values of (a, b) are applied, and the test for residual small study effect as described in equation (2.29) is applied. If all obtained p-values from the test are above a threshold 0.1, this is interpreted as no evidence, and no need for modelling, and the standard, classical random effects meta-analysis is retained. If none of the p-values is above the threshold, a wider range of values for (a, b) is used. When some p-values are above, and some below the threshold, the pair (a, b) with the smallest number of missing studies is retained (that is, the least intense underlying selection model is chosen).
Currently, there is no test to detect miss-specifications in the model itself, the authors themselves have argued that a non-parametric test of the residuals would lack power.

2.6 Transformation between Effect Measures

Assuming that binary outcomes result from a dichotomization of originally continuous random variables, in this case, the logistic distribution, a transformation from typical binary effect measures to Cohen's d can be achieved (Borenstein *et al.*, 2011, 47).
Let θ be a log odds ratio and s it's standard error. Cohen's d and it's variance s_d^2 is obtained by

$$d = \theta \frac{\sqrt{3}}{\pi} \qquad\qquad s_d^2 = \mathrm{se}^2 \frac{\sqrt{3}}{\pi}$$

The factor $\frac{\pi}{\sqrt{3}} = 1.81$ is the standard deviation of the logistic distribution $L(\mu, \eta)$ with scale parameter $\eta = 1$, so we just divide the log odds ratio and it's variance through the standard deviation. The approximation works only well if e_t and e_c are not very small, especially in the case of s_d^2. Plugging in the observed log odds ratio $\hat{\theta}$ and $\hat{\mathrm{se}}^2$ will give an estimate of Cohen's d.
Pearson's correlation can be attained by the formulas (Hedges and Olkin (1985), (Borenstein *et al.*, 2011, 48))

$$r = \frac{d}{\sqrt{d^2 + a}} \qquad\qquad a = (n_c + n_t)^2 / n_c n_t$$

where a is a correction factor if $n_t \neq n_c$. The variance of r, s_r^2 is computed by

$$s_r^2 = \frac{a^2 s_d^2}{(d^2 + a)^3}$$

Finally, we can get to a fisher's z-scaled correlation z and it's variance s_z^2 by using

$$z = 0.5 \ln\left(\frac{1 + r}{1 - r}\right) \qquad\qquad s_z^2 = \frac{1}{n - 3}$$

Chapter 3

The Cochrane Dataset

3.1 Cochrane Systematic Reviews

The Cochrane Group has specialized on systematic reviews in clinical science. Certain knowledge of standards and principles of the Cochrane Group may help to assess the quality and the properties of the dataset. The following information stems from the Cochrane Handbook for Systematic Reviews (Higgins JPT, 2011).
The definition of a systematic review is that it "attempts to collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question." Thus, the "key properties of a review are":

- "a clearly stated set of objectives with pre-defined eligibility criteria for studies"
- "an explicit, reproducible methodology"
- "a systematic search that attempts to identify all studies that would meet the eligibility criteria"
- "an assessment of the validity of the findings of the included studies, for example through the assessment of risk of bias"

At the end of a systematic review, "a systematic presentation, and synthesis, of the characteristics and findings of the included studies" is done.
53 Cochrane Review Groups prepare and maintain the reviews within specific areas of health care. A group consists of "researchers, healthcare professionals and people using healthcare services (consumers)".
The groups are supported by Method Groups, Centers and Fields. The Cochrane Method Groups aim to discuss and consult the groups in methodological questions concerning review preparation. The Centers play a main role in training and support of the Groups. The Fields are responsible for broad medical research areas and follow priorities in those areas by advice and control of the groups.
The first step in a review is writing a protocol, specifying the research question, the methods to be used in literature search and analysis and the eligibility criteria of the study. Changes in protocols are possible but have to be documented and the protocol is published in advance of the publication of the full review. The choices of methodology as well as the changes should not be made "on the basis of how they affect the outcome of the research study".
In order to avoid potential conflicts of interests, there is a code of conduct that all entities of the Cochrane Organization have to agree on: conflicts of interest must be disclosed and possibly be forwarded to the Cochrane Center, and participation of review authors in the studies used have to be acknowledged. Additionally, a Steering Group publishes a report of potential conflicts of interests based on information about external funding of Cochrane Groups.

In order for keeping the reviews up-to-date, they are revised in a two-year circle with exceptions. In addition to inclusion of new evidence in a field, the revision and maintenance process may as well includes change in analysis methods. This can reflect some advance in clinical science as for example new information about important subgroups, as well as new methods for conducting a Cochrane Review. However, there are no clear guidelines and the Cochrane Groups are free in the rate and extent of up-dating their reviews.

3.1.1 Methods for Cochrane Reviews

A research question defines the following points: "the types of population (participants), types of interventions (and comparisons), and the types of outcomes that are of interest". From the research question, usually the eligibility criteria follow. Usually, outcomes are not part of eligibility criteria, except for special cases such as adverse effect reviews.
The type of study is an important eligibility criterium. The Cochrane Collaboration focuses "primarily on randomized controlled trials", and also, the methods of study identification in literature search are focused on randomized trials. Furthermore, study characteristics such as blinding of study operators with respect to treatment and cluster-randomizing might be additional eligibility criteria which have to be chosen by the review authors.
After having specified the eligibility criteria, studies have to be collected. The central idea of systematic reviews, and also meta-analyses, is that the collected studies are a random sample of a population of studies, i.e. that they are representative and can be used to assess population properties. Therefore, the search process is crucial, as a selective search result may impose bias on the sample of studies available, making it a non-random sample. For this purpose, the Cochrane Groups are advised to go beyond MEDLINE !!cite!!, because a search restricted to it has been shown to deliver only 30% to 80% of available studies. "Time and budget restraints require the review author to balance the thoroughness of the search with efficiency in use of time and funds and the best way of achieving this balance is to be aware of, and try to minimize, the biases such as publication bias and language bias that can result from restricting searches in different ways." It is important to note that not only studies, but also study reports are occasionally used in the reviews, as they may provide useful information.
There are different sources that are being used to search for studies.

- The Cochrane Central Register of Controlled Trials is a source of reports of controlled trials. "As of January 2008 (Issue 1, 2008), CENTRAL contains nearly 530,000 citations to reports of trials and other studies potentially eligible for inclusion in Cochrane reviews, of which 310,000 trial reports are from MEDLINE, 50,000 additional trial reports are from EMBASE and the remaining 170,000 are from other sources such as other databases and handsearching." It includes citations published in many languages, citations only available in conference proceedings, citations from trials registers and trials results registers.

- MEDLINE. MEDLINE includes over 16 million references to journal articles. 5,200 journals publishing in 27 languages are indexed for MEDLINE. PubMed gives access to a free version of MEDLINE with up-to-date citations. NLM gateway such as the Health Services Research Project, Meeting Abstracts and TOXLINE Subset for toxicology citations allows for search in both databases together with additional data from the US National Library of Medicine.

- EMBASE. 4,800 Journals publishing in 30 languages are indexed to EMBASE, which includes more than 11 million records from 1974 onward. EMBASE.com also includes 7 million unique records from MEDLINE (1966 up to date) together with its own records. Additionally, EMBASE Classic allows access to digitized records from 1947 to 1973. EMBASE and MEDLINE each have around 1,800 journals not indexed in the other database.

- Regional or national and subject specific databases can additionally be consulted and

often provide important information. Financial considerations may limit the use of such databases.

- General search engines such as Google Scholar, Intute and Turning Research into Practice (TRIP) database can be used.
- Citation Indexes. The database lists articles published in around 6,000 Journals with articles in which they have been cited and is available online as SciSearch. This form of search is known as cited reference searching.
- Dissertation sources. Dissertations are often listed in MEDLINE or EMBASE but one is advised to also search in specific dissertation sources.
- Grey Literature Databases. Approximately 10% of the results in the Cochrane Database stems from conference abstracts and other grey literature. The Institute for Scientific and Technical Information in France provides access to entries of the previously closed System for Information on Grey Literature database of the European Association for Grey Literature Exploitation). Another source is the Healthcare Management Information Consortium (HMIC) database containing records from the Library and Information Services department of the Department of Health (DH) in England and the King's Fund Information and Library Service. The National Technical Information Service (NTIS) gives access to the results of US and non-US government-sponsored research, as well as technical report for most published results. References from newsletters, magazines and technical and annual reports in behavioral science, psychology and health are provided in the PsycEXTRA database which is linked to PsycINFO database.

3.1.2 Structure and Content

The dataset consists of 6354 systematic reviews from the Cochrane Library with 70662 studies and 744720 results. The results all stem from comparisons of a healthcare interventions or treatments to a reference. The reference may be a placebo control group or a other intervention or treatment. A result can not only be about efficacy of the treatments, but also about safety (adverse effects).
In Table 3.1, two results from a systematic review about barbiturates are shown as they are given in the dataset. As can be seen, further specifications are provided by the variables in the columns.
The `comparison.name` variable specifies *what kind* of treatments or interventions are compared, the `outcome.name` variable *how* it is compared, and the `subgroup.name` variable (not indicated in table) *if and to what experimental subgroup* the result belongs.
The result is of a binary type, and the counts of events in the treatment group are in `events1` and of the control group in `events2` and the total number of participants are given in columns `total1` and `total2`. As can be seen, events denote here "death at the end of follow-up".

study.name	comparison.name	outcome.name	events1	total1	events2	total2
Bohn 1989	Barbiturate vs no barbiturate	Death at the end of follow-up	11	41	11	41
Ward 1985	Barbiturate vs no barbiturate	Death at the end of follow-up	14	27	13	26

Table 3.1: Example of two results as given in the dataset. Events denotes the count of events in the treatment group while Events c the count of events in the group compared to. Further descriptive variables have been ommitted

A complete listing of the variables of a result is given in Table 3.2. Depending on the type of the data, *e. g.* if it is binary or continuous, some variables are missing for the specific results.

Results are part of studies that are again part of a (systematic) review. This structure of a review is shown in Figure **??**.

Variable	Description
`id`	An id of the review for identification purposes
`study.name`	Name of the study to which the result belongs
`study.year`	Year in which the study was published
`study.data.source`	Source of the study, either "Publ
`comparison.name/.nr`	Specification of the interventions compared in the study and a unique number for the comparison
`outcome.name/.nr`	Specification by which outcome the interventions are compared and a unique number for the outcome
`subgroup.name/.nr`	Potentially indication of affiliation to subgroups and a unique number for the subgroup
`outcome.measure`	Indication of the quantification method of the effect (of one intervention compared to the other).
`effect`	Measure of the effect given in the quantity denoted by `outcome measure`.
`se`	Standard error of the measure of the effect,
`events1/events2`	The counts of patients with an outcome *if* measurement/outcome is binary or dichotomous 2 (1 for treatment group and 2 for control group).
`total1/total2`	Number of patients in groups.
`mean1/mean2`	Mean of patient measurements *if* outcome is continuous.
`sd1/sd2`	Standard deviation of mean *if* outcome is continuous.

Table 3.2: Dataset variable names and descriptions

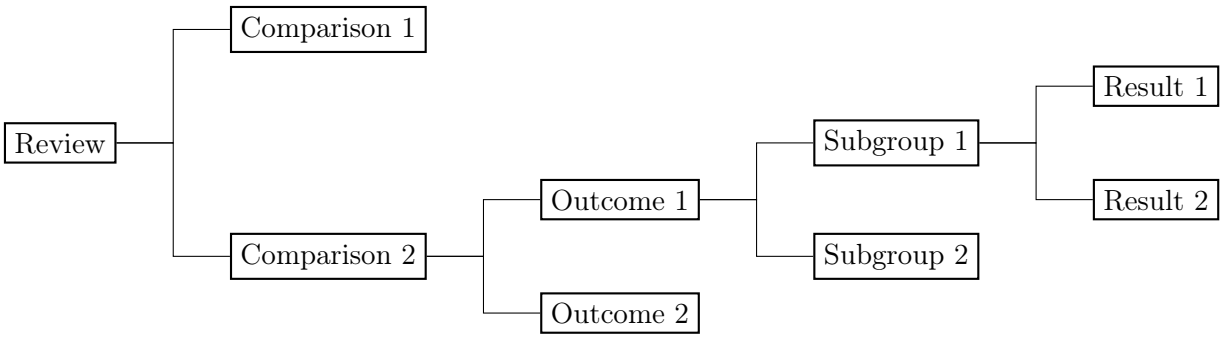

Figure 3.1: Structure of a hypothetical review with two different comparisons

The structure of a review will now be outlined based on an example of the dataset. Let us consider the previously mentioned barbiturate and head injury review. The aim was to "assess the effects of barbiturates in reducing mortality, disability and raised ICP (intra-cranial pressure) in people with acute traumatic brain injury" as well as to "quantify any side effects resulting from the use of barbiturates" The review comprises five studies in total. Three of them compared barbiturate to placebo, one compared barbiturate to Mannitol and one Pentobarbital to Thiopental. The studies have different outcomes, for example, death or death and severe disability at follow up, but also dropout counts or adverse effects (secondary outcomes). We have continuous (e.g. mean body temperature) and binary outcome data (e.g. death/no death). One study split up outcomes for patients with and without haematoma, which would be subgroups. Thus, it is important not to confuse results with studies. A study can contribute multiple results to a systematic review, for example, primary and secondary outcomes and adverse effects.

study.name	comparison.name	outcome.name
Bohn 1989	Barbiturate vs no barbiturate	Death at the end of follow-up
Bohn 1989	Barbiturate vs no barbiturate	Death or severe disability at the end of follow-up
Eisenberg 1988	Barbiturate vs no barbiturate	Death at the end of follow-up
Eisenberg 1988	Barbiturate vs no barbiturate	Uncontrolled ICP during treatment
Eisenberg 1988	Barbiturate vs no barbiturate	Hypotension during treatment
Perez-Barcena 2008	Pentobarbital vs Thiopental	Death at the end of follow-up (6 months)
Perez-Barcena 2008	Pentobarbital vs Thiopental	Death or severe disability at the end of follow-up (6 months)
Perez-Barcena 2008	Pentobarbital vs Thiopental	Uncontrolled ICP during treatment
Perez-Barcena 2008	Pentobarbital vs Thiopental	Hypotension during treatment
Schwartz 1984	Barbiturate vs Mannitol	Death at the end of follow-up (1 year)
Schwartz 1984	Barbiturate vs Mannitol	Death at the end of follow-up (1 year)
Schwartz 1984	Barbiturate vs Mannitol	Uncontrolled ICP during treatment
Ward 1985	Barbiturate vs no barbiturate	Death at the end of follow-up
Ward 1985	Barbiturate vs no barbiturate	Death or severe disability at the end of follow-up
Ward 1985	Barbiturate vs no barbiturate	Mean ICP during treatment
Ward 1985	Barbiturate vs no barbiturate	Mean arterial pressure during treatment
Ward 1985	Barbiturate vs no barbiturate	Hypotension during treatment
Ward 1985	Barbiturate vs no barbiturate	Mean body temperature during treatment

Table 3.3: Barbiturate and head injury review. In the columns, study names, comparison and outcome measure of the results are given

Information about missing values in the dataset is given in Table 3.4. The relative amount of missing values is low, except for study years. For continuous outcomes, the cases are counted were neither a mean difference nor means are available. Similarly, the counts of cases where neither standard errors nor standard deviations are available are provided. Study years before 1920 and after 2019 are declared as missing, as well as sample sizes equal to zero.

Neither means nor standard deviations (CONT)	775
Zero participants in one group	15162
Missing study publication year	7942

Table 3.4: Number of missing variables and measurements in the dataset

The studies that are included in the reviews and have been published are most often from the years after 1980 (5% quantile = 1982, 95% quantile = 2014,). The median of the publication years is 2003, the mean 2000.97 and the quartiles are 1996 and 2008. 1075 studies have been published in 2018, none in 2019.
The top treatment effect measure (risk ratio, mean difference, hazard ratio etc.) abundances are summarized in Table 3.5. One can conclude of the table that roughly 31 % of outcomes in the dataset are continuous and the rest being some sort of discrete or binary outcomes, most often binary (more than 65%).

The sample sizes among results vary to some extent. There are 5% of treatment group sample sizes that are smaller than 9, 95% smaller than 510. The first quartile is 23, the median 48, the mean 256.71 and the third quartile 116. The large difference between median and mean is caused by very large groups with over 2,000,000 participants. Analogously, the quantiles of the total sample size are: 5% quantile = 17, first quartile = 44, median = 94, third quartile = 223 and 95% = 983. The mean is 623.15.
There are 519 reviews with five or fewer results. The 5% and 95% quantiles are 4 and 447. The mean and median number of results per review are 13.6 and 8, and the quartiles are 16 and 109. Similarly, the number of reviews with a maximum of two studies included is 1040, the mean study number is 13.6, the median 8 and the interquartile range 4 and 16 and the 95% quantile 45. The discrepancy between mean and median is due to large reviews with a high number of studies and results, most extreme in which is a systematic review about antibiotic prophylaxis for preventing infection after cesarean section, with 95 studies and 1,497 results in total.

Outcome measure	n	Percentage
RR	361902	48.6%
MD	164923	22.1%
OR	76067	10.2%
SMD	70717	9.5%
PETO_OR	39710	5.3%
RD	11068	1.5%
Hazard Ratio	8054	1.1%
Rate Ratio	3724	0.5%
other	8555	1.1%

Table 3.5: Frequencies of outcome measures among results. n denotes the total number of results with the outcome measure and percentage the percentage of the outcome measure,

For results to be suitable for usage in meta-analysis, they have to be identical with respect to comparison and outcome. The studies in the dataset that have the same comparison, outcome and subgroup can be pooled in a meta-analysis. This distinction is also used by the Cochrane Organization itself, *i. e.* the meta-analyses are identical to the meta-analyses done in the systematic reviews.
The size of a meta-analysis denotes how many results are included in a group. Table 3.6 shows the number of meta-analysis with size $\geq n$ results.

```
## Warning:  Setting row names on a tibble is deprecated.
```

n	Number of groups	Cumulative sum of groups
1	143378	268906
2	45459	125528
3	23232	80069
4	14184	56837
5	9493	42653
6	6449	33160
7	4583	26711
8	3412	22128
9	2585	18716
10	2046	16131
11	1524	14085
12	1197	12561
13	1022	11364
14	785	10342
15	9557	9557

Table 3.6: Number and cumulative number of groups with meta-analysis size n.

3.2 Data Tidying and Processing

The dataset was scraped from the Cochrane Database of Systematic Reviews at the ... . A specialized code was used to convert the xtml. files from the webpage into the data at hand cite... .

3.2.1 Newly Introduced Variables

Some new variables are added to the obtained dataset:

- `outcome.measure.merged`: Outcome measure specifications given in `outcome.measure` were standardized whenever they were supposed to denote the same outcome measure.

An example would be the formally different notations for odds ratios: "Odds Ratio", "odds ratio", "OR", which were all denoted as "Odds Ratio".

- `lrr` and `var.lrr`: log risk ratio and variance of the log risk ratio for outcome.flag "bin".
- `smd` and `var.smd`: Hedges g and the variance of Hedges g for outcome.flag "cont".
- `smd.ordl` and `var.smd.ordl`: Cohen's d and its variance as obtained by transformation of a log odds ratio for outcome.flag "bin".
- `cor.Pearson` and `var.cor.Pearson`: Pearson correlation coefficient and variance as obtained from the d (for outcome.flag "bin") or g (for outcome.flag "cont") to r transformation.
- `z` and `var.z`: Fisher's z score and it's variance obtained from the Pearson correlation r to z transformation.
- `pval.single`: p-value against the null hypothesis of no treatment effect, derived by a t-test for outcome.flag "cont" or Wald test for outcome.flag "bin".
- `events1c` and `events2c`: Correction of `events1` and `events2` zero event counts or event counts = patient number. When no events occurred, 0.5 was added, and when all patients experienced the event, 0.5 was subtracted. When one of `events` had zero counts while the other had maximum counts, no adjustment occurred.
- `meta.id`: Meta-analysis ID variable to uniquely identify any potential meta-analysis in the dataset. Consistent to what has been discussed before, all results that share a common comparison, outcome and subgroup (optional, subgroups not given in any case) may be combined in a meta-analysis.
- `se.new`: Depending on `outcome.flag`, `se.new` is equal to the square-root of `var.lrr` (`outcome.flag` = `DICH`), square-root of `var.smd` (`CONT`) or `se` (`IV`).

3.2.2 Eligibility criteria for Publication Bias Test and Adjustment

Initially, the analysis is restricted to results with `outcome.flag` `DICH`, `CONT` and `IV`. binary, continuous and survival outcomes. This corresponds to 734954 out of a total of 744720 results. Ioannidis and Trikalinos (2007a) outlined criteria for application of small study effect tests:

- **Sample size**: A meta-analysis is comprised of at least ten studies (n = 9916 remaining).
- **Study size**: The ratio between largest variance of an estimate and smallest variance of an estimate is larger than four (n = 9334 remaining).
- **Significance**: At least one treatment effect has a p-value below the significance threshold 0.05 (n = 7509 remaining)
- **Heterogeneity**: The I^2 statistic of a given meta-analysis is smaller than 0.5, thus, the proportion of between study variance of the overall variance is smaller than 0.5 (n = 1403 remaining).

Additionally, the following criteria have been applied:

- **Sensitivity Analyses**: When the same results are used multiple times for different meta-analyses, only one is retained. More precisely, if a study hat the same `study.name` and same `effect`, it was considered a duplicate, and the smaller meta-analysis of the two was excluded. The intention is to exclude sensitivity analyses which are operated on subsets of the available results.

- **Zero events**: In the case of binary outcomes, meta-analyses with zero events in any study and any group are excluded ($n = 20$ out of meta-analyses with at least ten studies).

The results of this reduction of the dataset are shown in the flow-chart in Figure 3.2. By applying the criteria, we are only able to analyse a subset of the original data.
Only the data that had accesible all results data available was used for adjustment. Thus, all meta-analyses with `outcome.flag` == `IV` are ommitted in a second step of the analysis (Analysis dataset (1) and (2) in Figure 3.2).

Exclusions due to Computational Errors

There has been cases where meta-analyses could not be analysed with Copas selection model because the likelihood function could not be optimized based on the settings. It was ommitted in this case.

- Meta-analysis with `meta.id` = 93868, `id` = "CD004072", `subgroup.id` = "CMP-006.08.01". One small meta-analysis (`se.new` = 1.02) with a risk ratio 0.072, while the remaining risk ratios were all larger than 0.45. (and the fixed effects estimate close to 1). It failed when analysed based on log risk ratios.

- Meta-analysis with `meta.id` = 144568, `id` = "CD006109", `outcome.id` = "CMP-001.06" and `outcome.measure` "MD". z-score based analysis failed, although funnel plot looks more or less symmetric. Small studies with very similar `var.z` with z-scores 0.3 and -0.1.

- Meta-analysis with `meta.id` = 35673, `id` = "CD001790", `outcome.id` = "CMP-001.03" and `outcome.measure` = "PETO.OR". Analysis failed when using Hedges g transformed effects. Large studies have $g \sim 0$ while one smaller study (se = 0.3) has $g = 1$ (and another with a little larger se has $g \sim 0$).

Because the number of errors is low, the meta-analyses have only been ommitted for the specific outcome measures, but are kept in the others.

The Analysis Dataset

The dataset that was ultimately tested and adjusted for publication bias comprises 1 meta-analyses and 23243 results. The mean number of participants in the treatment group is 251.9 vs 256.7 in the unrestricted dataset and the mean total number of participants 584.7 vs 623.2. The mean publication year is 2000.3028869 vs 2000.9711014 in the unrestricted dataset.
From the meta-analyses with incomplete data (`outcome.type` == `IV`), there are 34 with `outcome.measure.merged` = "Risk Ratio", 27 "Std. Mean Difference", 26 "Hazard Ratio" and 13 "Odds Ratio" (and 25 additional, different outcome measures).

3.2.3 Analysis Procedure

All compuations were performed in the `R` computing environment (R Core Team, 2018). The `R` packages `meta` (Schwarzer, 2007) and `metafor` (Viechtbauer, 2010) were used for meta-analysis, small study effect tests and adjustments. The excess significance test was adapted from van Aert *et al.* (2019). For data manipulation procedures, the `tidyverse` packages were used (Wickham, 2017), and for plotting the `ggplot2` package (Wickham, 2016).
The methods described in the methods chapter 2 most often apply directly to the algorithms used in `meta` and `metafor`. For binary outcomes, log risk ratios were used as treatment effect estimates in the meta-analysis, and for continuous outcomes, Hedge's g. For the other outcomes, the `effect` and `se` as provided in the dataset was passed to `meta::metagen`.

The effect sizes were transformed for adjustment in the way described in section 2.6. The p-values for adjusted treatment effect estimates were calculated by the Wald method.
In the case of the one-sided test procedure (small study effect and excess significance tests), the effect side in which bias was expected had to be prespecified. This was solved by comparing the number of significant findings on each side (original effect scale, i.e. for binary outcomes log risk ratios, etc.); the side with more significant findings (two-sided p-value < 0.05) was considered the side of potential bias. If numbers were equal, the side of the fixed effects treatment was used. Details to Copas selection model algorithm and its application can be found in Rücker *et al.* (2011). In short, two values for a and b in section 2.5.2 were used; a limited range with a between -1.7 and 2, and b between 0.16 and 0.32 was applied first (analog to a most extreme selection process of P(select|small trial with sd = 0.4) = 0.1 and P(select|large trial w. sd = 0.05) = 0.9). If the most extreme selection process is unable to pass the significance test of no small study effect (p-value > 0.1), then a wider range was applied (a between -5.4 and 2 and b between 0 and 0.32). If there is still no non-significant small study effect, the result was `NA`.

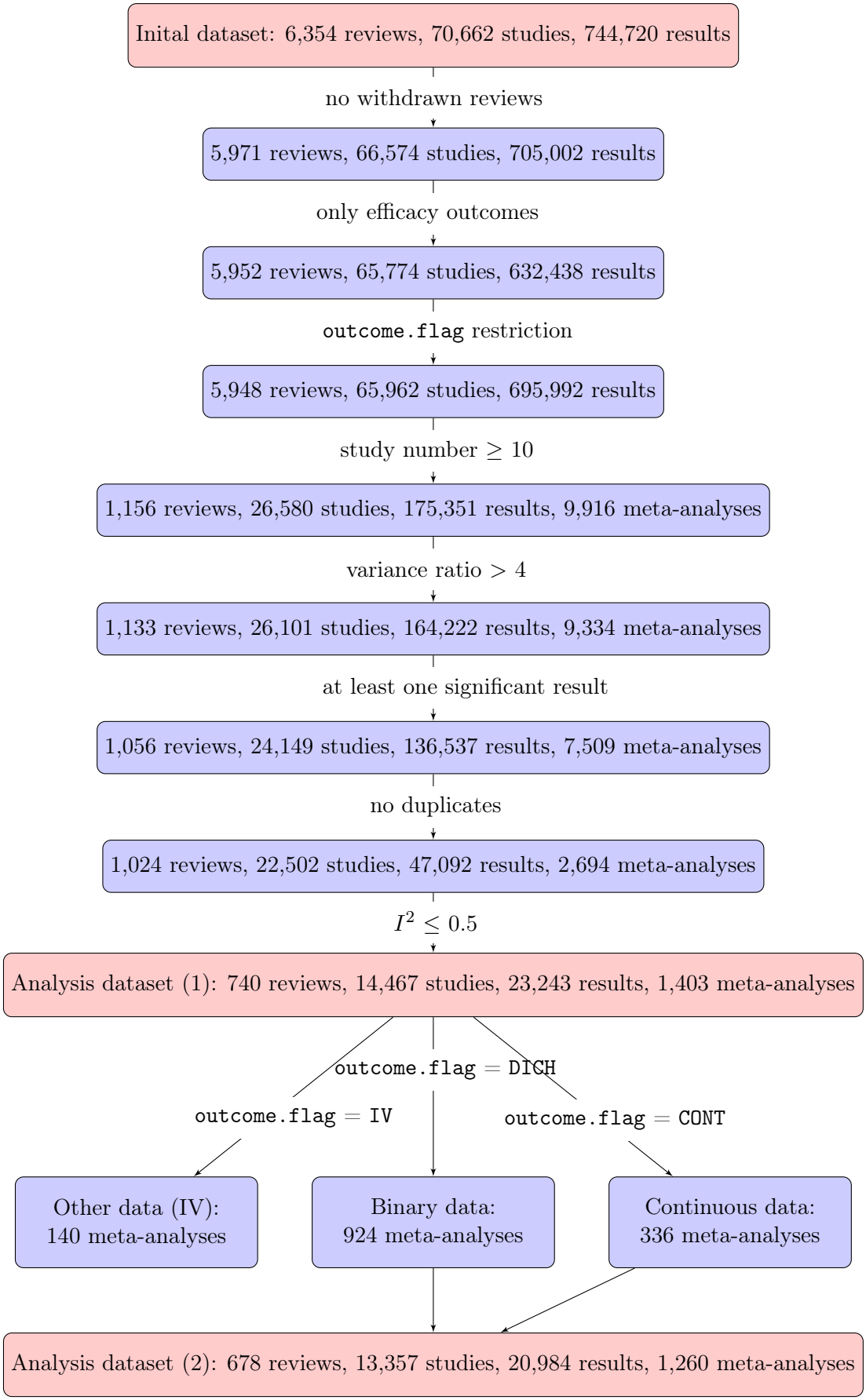

Figure 3.2: Flow-chart of the exclusion of meta-analyses for the final analysis. The exclusion criteria are given at the right of the arrows.

Chapter 4

Results

```
## Scale for 'x' is already present.  Adding another scale for 'x', which
## will replace the existing scale.
## Scale for 'x' is already present.  Adding another scale for 'x', which
## will replace the existing scale.
```

Using Fisher's z transformation, we can transform all effect sizes of the dataset on the same scale such that they are comparable to each other. Thus, as a first step, an exploratory plot of the median effect size and it's dependence on the study sample size can be shown.
The absolute value of the median z-score for a given sample size of a trial is shown in Figure 4.1. It is clearly visible that the absolute value decreases with increasing sample size, i.e. that the effect size is becoming smaller. The trend flattens off after $\sim$ sample size $= 400$ (not shown).
The pattern is similar in Figure 4.2, where the original effect size measures "Odds Ratio", "Risk Ratio", "Mean Difference" and "Std. Mean Difference" are used (the most common measurs in the dataset). Here, the effect sizes are scaled (mean subtraction and divison through standard error) with respect to all other effects of the same measure.

4.1 Publication Bias Test Results

The meta-analyses fulfilling the criteria from chapter 3, section 3.2, are analysed with one-sided small publication bias tests and excess significance tests. The direction in which bias is expected is the one on which more significant results are (p-value < 0.05, two-sided). The tests are applied on the original effect size measures, since the journal editors and the researchers also base their decisions on them. Different tests are applied depending on the outcome being binary or continuous.
Multiple tests are applied on the same data to compare their results. A histogram of p-values for each test and all meta-analyses will summarize the overall evidence against the null-hypothesis of no publication bias, as displayed in Figure 4.3. Skewedness indicates evidence for publication bias.
The abbreviations in Figure 4.3 are shortly explained with references to chapter 2:
"Excess significance" denotes the excess of significant p-values testing method from Ioannidis and Trikalinos (2007b), see 2.4.4. For continuous and `IV` outcomes, the names refer to:

- Egger's test, weighted linear regression test described in section 2.4.1
- Thompson and Sharp's test, weighted linear regression test adjusted for between-study heterogeneity, section 2.4.1
- Begg and Mazumdar's test, rank test described in section 2.4.1

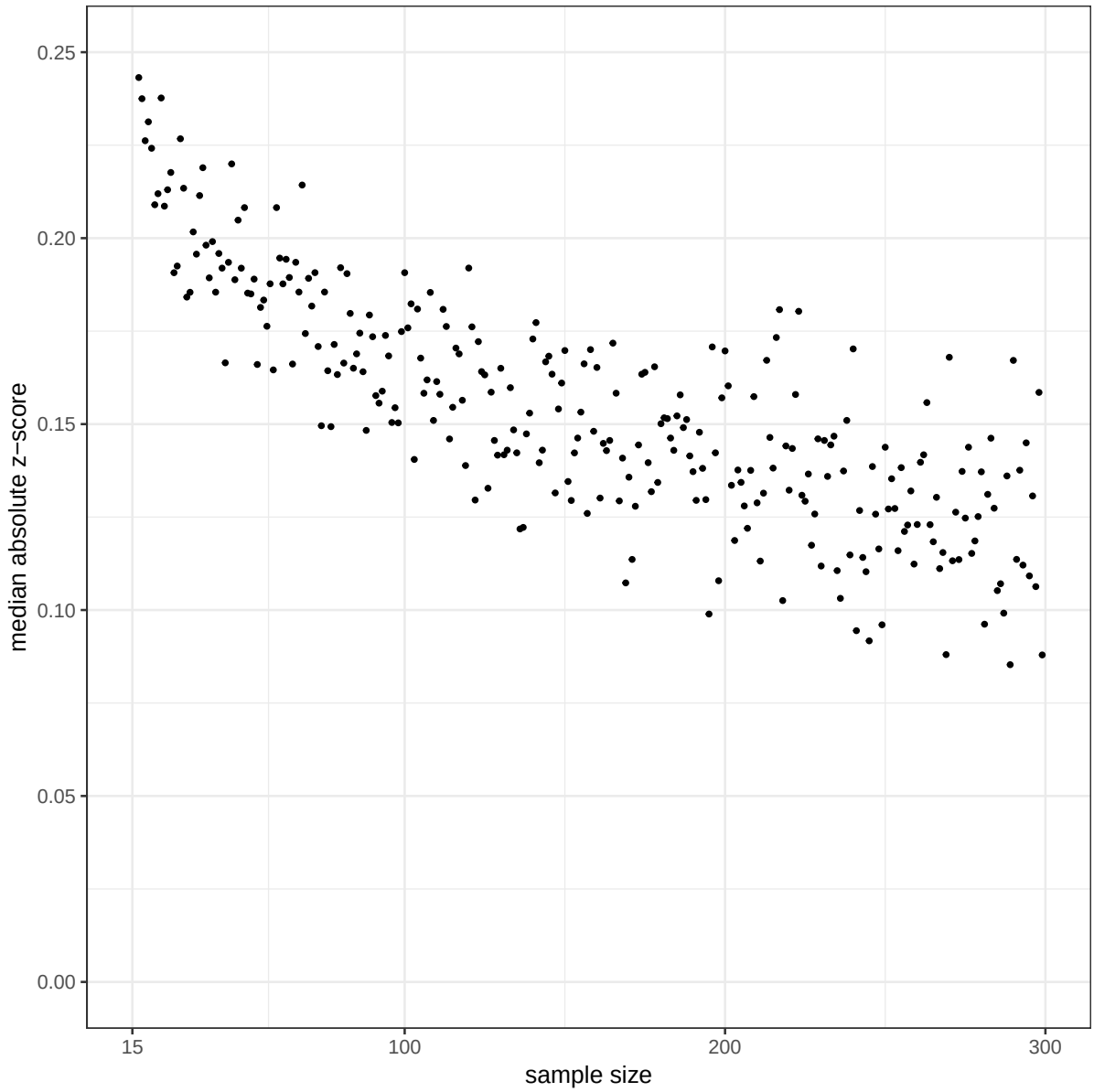

Figure 4.1: Median of the absolute z-score plotted against the total sample size.

For binary outcomes, the names refer to:

- Harbord's test, likelihood score based test (section 2.4.3)
- Peter's test, weighted linear regression with inverse sample size as explanatory variable described in section 2.4.3
- Rücker's test, test based on the arcsine transformation of proportions, in combination with Thompson and Sharp's regression test (section 2.4.3)
- Schwarzer's test, rank based test using the expected event counts computed with the hypergeometric distribution (section 2.4.3)

The histograms in Figure 4.3 show that the tests are not alway in agreement if publication bias is present in the data. The p-values of excess significance and Schwarzer's test and the p-values for `IV` outcomes are rather uniformly distributed, or even skewed to the right. Notably, excess significance test, Schwarzer's test and rank tests have been shown to lack power. The

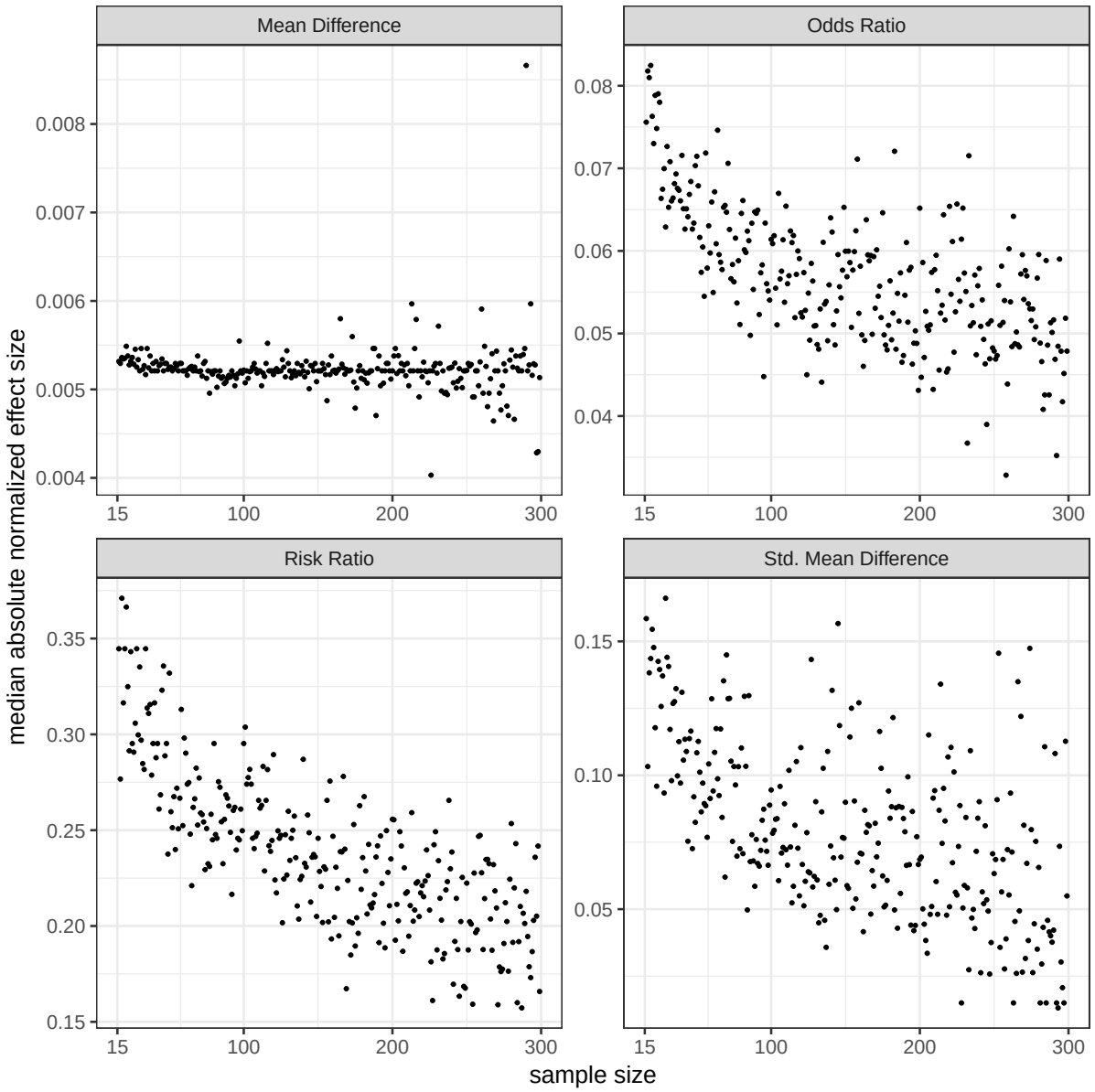

Figure 4.2: Median of the absolute value of the normalized, original effect size plotted against the total sample size.

tests that are more suitable (regression based tests in general) all have proportions of significant p-values ($p > 0.1$) clearly above 10 % which would be the expected false positive rate. The conclusion is that there is publication bias, however the extent can be debated. The meta-analyses with `outcome.flag` = `IV` seem to differ from the other meta-analyses.
In Figure 4.3, the meta-analyses with an estimated I^2 of zero are depicted, because some methods are known to only be suitable when no heterogeneity is present (excess significance test and also Egger's test). Others are specially constructed to adjust for between study heterogeneity (Thompson and Sharp's test and Rücker's test). In continuous outcomes, when Thompson and Sharp's test is applied, the evidence for publication bias decreases, as can be seen in the smaller proportion of blue on the left corner in Figure 4.3. However, this is also due to application to meta-analyses with I^2 equal zero, in which case the method is known to lack power. Similarly, the evidence decreases somewhat when Rücker's test is extended by Thompson and Sharp's method to account for heterogeneity. The moderate decrease indicates however that the previous restriction to meta-analyses with $I^2 < 0.5$ is sufficient to remove meta-analyses with large

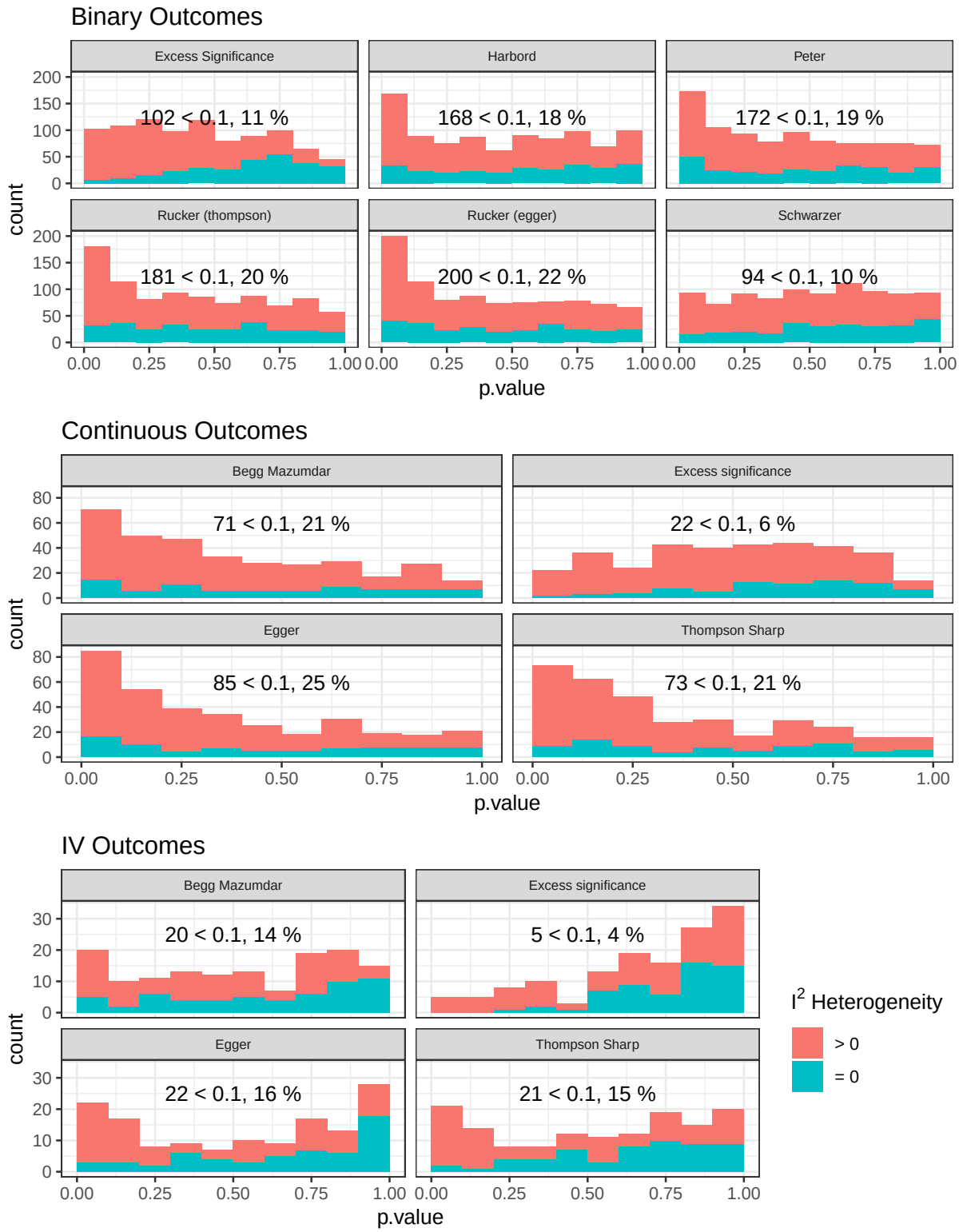

Figure 4.3: Histogram of one-sided p-values for small study effect in direction of larger effect sizes. The testing method is indicated in the header, bin width is equal to 0.1. The proportion of meta-analyses with signifiant publication bias based on the threshold of 0.1 is displayed inside the figures.

heterogeneity and that the test results are not heavily influenced by unaccounted between study heterogeneity.

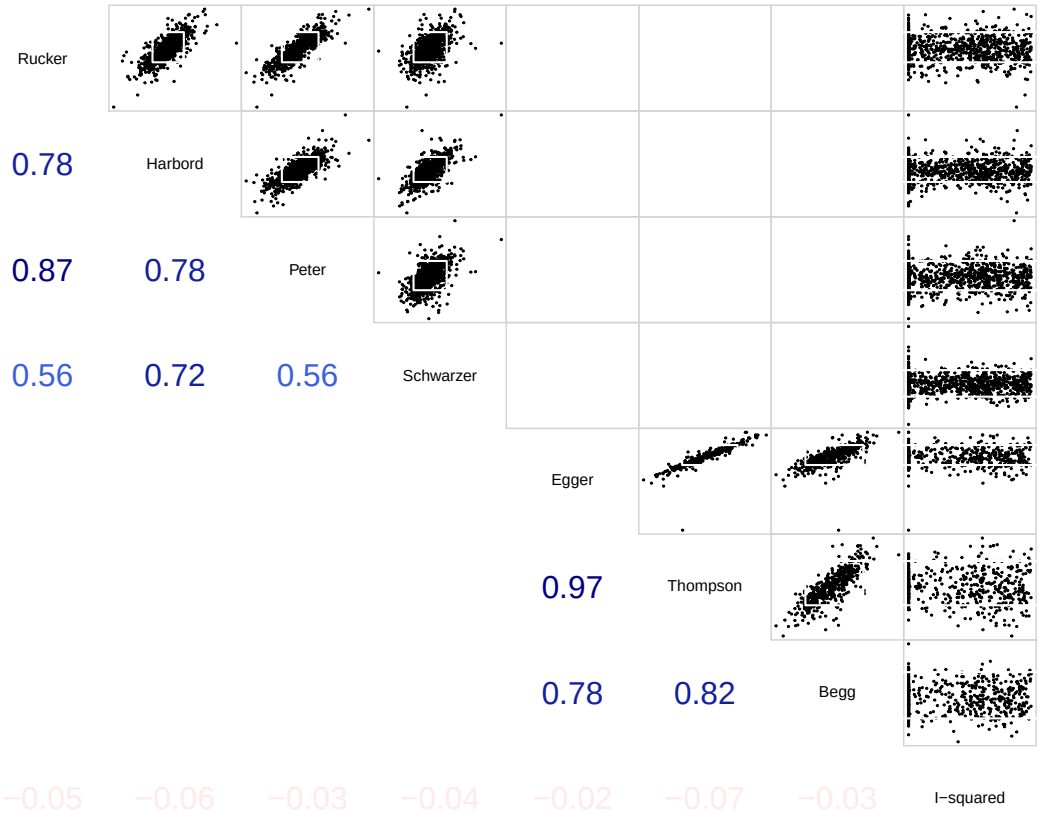

Figure 4.4: Pairs-plot for test statistics of small study effect and excess significance. The lower panel gives the Spearman correlations for the different test statistics, and the upper panel displays a scatterplot. The colors indicate magnitude and direction of the correlation coefficients. The rectangle with white borders displays the area within which both tests have absolute value < 1.64 (dots inside are statistically not significant by 0.1 p-value threshold).

The p-values of tests can be summarized by computing their harmonic mean. In the case of binary tests, the p-values of Rücker's, Peters, Harbord's, Schwarzer's and excess significance tests are used, in the case of continuous and `outcome.flag` = `IV` outcomes, Egger's, Thompson and Sharp's, Begg and Mazumdar's and excess significance tests are used. This lead's to an overall of 19.1% significant results ($p_{\text{harmonic}} < 0.1$, 19.7% `outcome.flag` = `DICH`, 20.1% `outcome.flag` = `CONT`, 12.9% `outcome.flag` = `IV`).

4.1.1 Publication Bias Test Consistency

It cannot be seen in Figure 4.3 if the tests used are finding small study effects for the same meta-analyses. A simple method to check the consistency of test results is to compare scatterplots and empirical Spearman correlations between the test statistics. This is done in Figure 4.4. Here, there is no separation between `IV` and continuous outcomes. The upper left rectangle is displaying binary outcome results and the lower right continuous and `IV` outcomes results. Also, the I^2 statistic is included. Since no approximately normally distributed test statistic is used for excess significance tests, it is not shown here

The observed patterns on the scatterplots differ, and some small study effect test statistics do align better than others. Regression based tests as Egger and Thompsons test which are methodically almost identical are closely aligned, which is reflected in large correlation coefficients. Continuous and`IV` outcome type tests align more closely than binary outcome tests. While correlation coefficients between binary outcome tests vary between each other, Harbord's test statistic has similar correlation coefficients with the other small study effect test statistics. Because scatterplots and correlation coefficients can be misleading, also a Tukey mean-difference or Bland-Altman of transformed p-values plot is shown for four scenarios in Figure 4.5:

- For Egger's and Thompson's tests, which is supposedly the most similar test and should show the least deviations and systematic errors
- For Egger's and excess significance tests
- For Harbord's and Rücker's tests
- For Harbord's and excess significance tests

This can be justified since all tests are supposed to measure the evidence for publication bias. For the plots, the p-values of the tests are transformed on the entire continuous scale by a logit transformation $f(x) = \log(\frac{p}{1-p})$. The mean p-value $((f(p\text{-value no. } 1) + f(p\text{-value no. } 2))/2)$ is then displayed against the difference between the $f(p\text{-value})$. If no systematic errors and biases exist between the measurement methods, then

- the mean of the differences should be around zero (no systematic error)
- the points should scatter independently on the y-axis and no general increase or decrease with the mean of the transformed p-values should be visible (and the linear regression fit is flat)

There are likely systematic errors and bias between the tests, although the extent seems to vary. Most error seems to be between small study effect tests and excess significance tests, because the slope of the linear regression fit is likely positive. This means that the excess significance test finds less evidence in cases when both p-values are small and more evidence when both p-values are large, on average.
The confidence intervals from Figure 4.5 can be used as limits of agreement, which gives, after back-transformation, around 0.9 for Egger's and Thompson's test, and around 0.99 for Harbord's and Rücker's tests. This suggests that additionally to the bias correspondence between the tests is not very good in general.

The previous results suggest that the results will also differ substantially after applying the common dichotomization of p-values. Some proportion of the meta-analyses will only be significant using a single test, while being non-significant otherwise. This can be seen in Table 4.1. It displays the percentage of meta-analyses with a certain number of significant test results. For both outcome types, there is only marginal agreement, very few cases are significant in any test. Around 67% of the dataset is not significant, no matter which test is used. When only using small study effect tests, 29.5% of binary outcome tests and 29% of `IV` and continuous outcome tests had at least one significant result. After applying the Bonferroni correction for multiple testing, this shrinks to 12.3% for binary and 12% for continuous and `IV` outcomes. To compare significant findings for small study effect tests and excess significance tests, Harbord's or Egger's test results are compared with excess significane tests. 24.1% of binary outcome analyses had at least one of the two test p-values being significant, and equivalently, 24.4% for continuous and `IV` outcomes. The numbers change to 13.9% for binary outcomes and 14.7% for continuous and `IV` outcomes after applying the Bonferroni correction. 5.1% have significant Harbord's test result and significant excess significance test result (2.3% with Bonferroni). Of the continuous

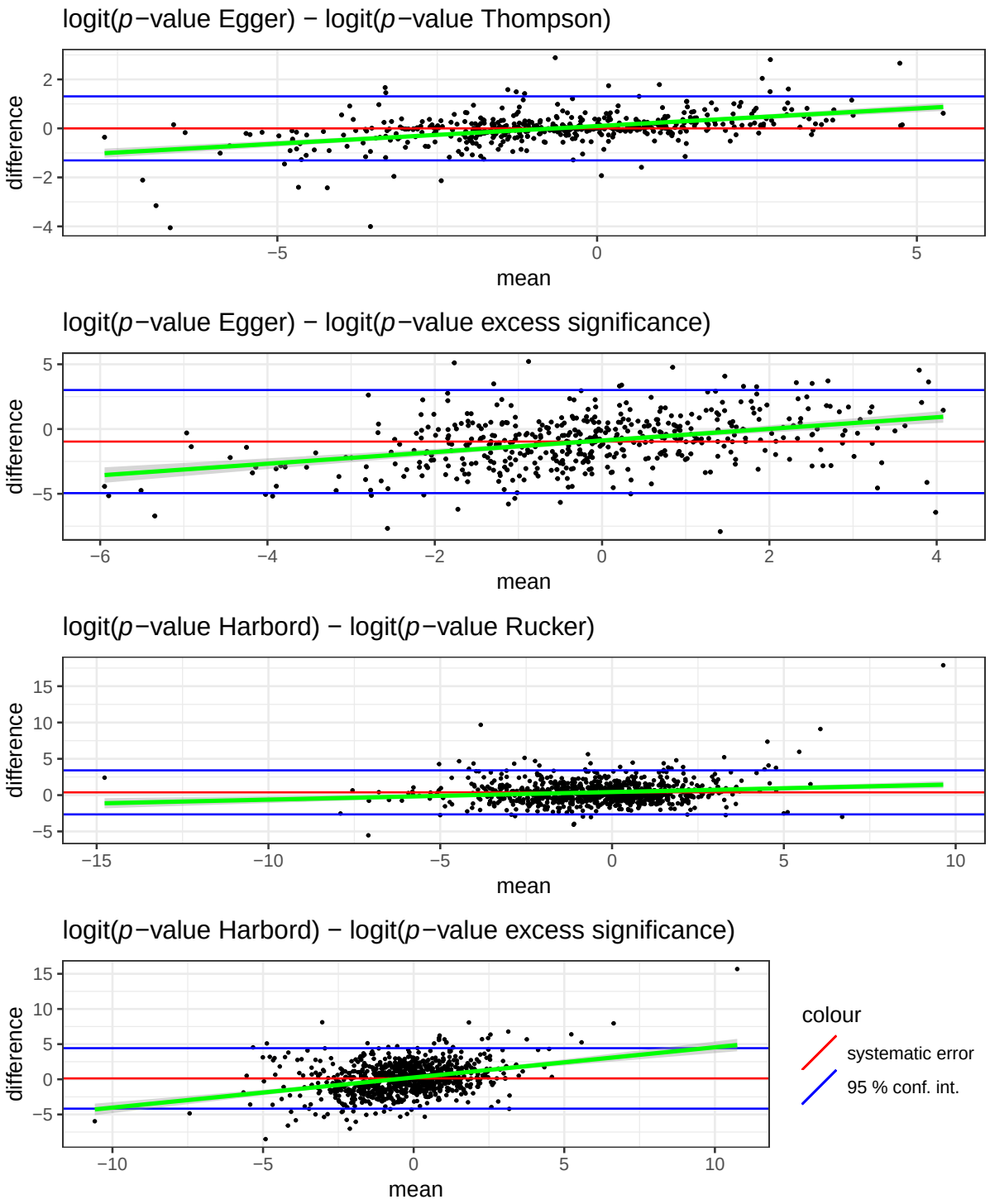

Figure 4.5: Mean - difference plots for logit transformed p-values. The mean of logit transformed p-values is displayed on the x-axis and the difference on the y-axis. Blue and red lines display the systematic error and the confidence intervals of the systematic error (limits of agreement). In green, a linear regression fit is shown with 95% CI bands.

outcomes, we have 3.3% with Egger's test and 1.4% with Bonferroni correction.

The precise proportions of agreement are provided in table 4.2. We see that only in few cases when excess significance test indicates statistical significance, small study effect tests do

Count	Binary Outcomes	Continuous and IV
0	66.1 %	68.9 %
1	12.3 %	11.2 %
2	7.7 %	6.6 %
3	6.8 %	11.4 %
4	5.7 %	1.9 %
5	1.3 %	-

Table 4.1: Counts Number of significant test results per meta-analysis, separated for outcome types. Last entry for continuous and IV outcomes is empty since one test less was applied

so. Correspondence between excess significance tests and small study effects thus has to be considered rather random. Linear regression based tests agree more often with other linear regression based tests, and agreement between small study effect tests is in general well above 60 % (at best, 95 % test agreement in significance for `IV outcome.flag`).

	Agreement (overall)	Agreement (significance)
Excess significance, Schwarzer	0.85	0.27
Excess significance, Peter	0.77	0.32
Excess significance, Rucker	0.79	0.43
Excess significance, Harbord	0.81	0.46
Peter, Schwarzer	0.84	0.63
Schwarzer, Rucker	0.85	0.72
Schwarzer, Harbord	0.87	0.78
Rucker, Peter	0.88	0.67
Harbord, Peter	0.87	0.63
Excess significance, Egger	0.77	0.64
Excess significance, Thompson	0.80	0.59
Excess significance, Begg	0.78	0.36
Thompson, Egger	0.93	0.92
Thompson, Begg	0.87	0.70
Egger, Begg	0.84	0.70
Excess significance, Egger (IV)	0.84	0.09
Excess significance, Thompson (IV)	0.84	0.10
Excess significance, Begg (IV)	0.85	0.10
Thompson, Egger (IV)	0.98	0.95
Thompson, Begg (IV)	0.86	0.55
Egger, Begg (IV)	0.87	0.60

Table 4.2: Overall proportion of agreement if significant or unsignificant, and for significance only. Horizontal lines separate binary, continuous and IV outcomes (order as in table). The reference significance test is the one with more significant results.

4.2 Small Study Effects Adjustment

4.2.1 Change in Effect Size after Adjustment

There are methods that can take into account the presence of publication bias in meta-analyses when estimating the overall treatment effect. The methods work in a semi-automatic manner; they will not only adjust for publication bias if smaller studies show larger effects, but also in the opposite case. The latter results in the adjusted overall treatment effect being *larger* than the unadjusted, overall treatment effect.

To compare the effects of adjustment between meta-analyses of different outcomes, the outcome measures are transformed to standardized mean differences and Fisher's z-scores (see section 2.6 for details). When comparing to unadjusted effects, fixed or random effects meta-analysis estimates are used as references

Figure 4.6 displays the difference δ between the estimated meta-analysis treatment effect and the regression adjusted treatment effect 2.5.1, $\hat{\theta}_M - \hat{\theta}_{\text{Adj.}}$. The absolute value $|\hat{\theta}_M|$ is taken and

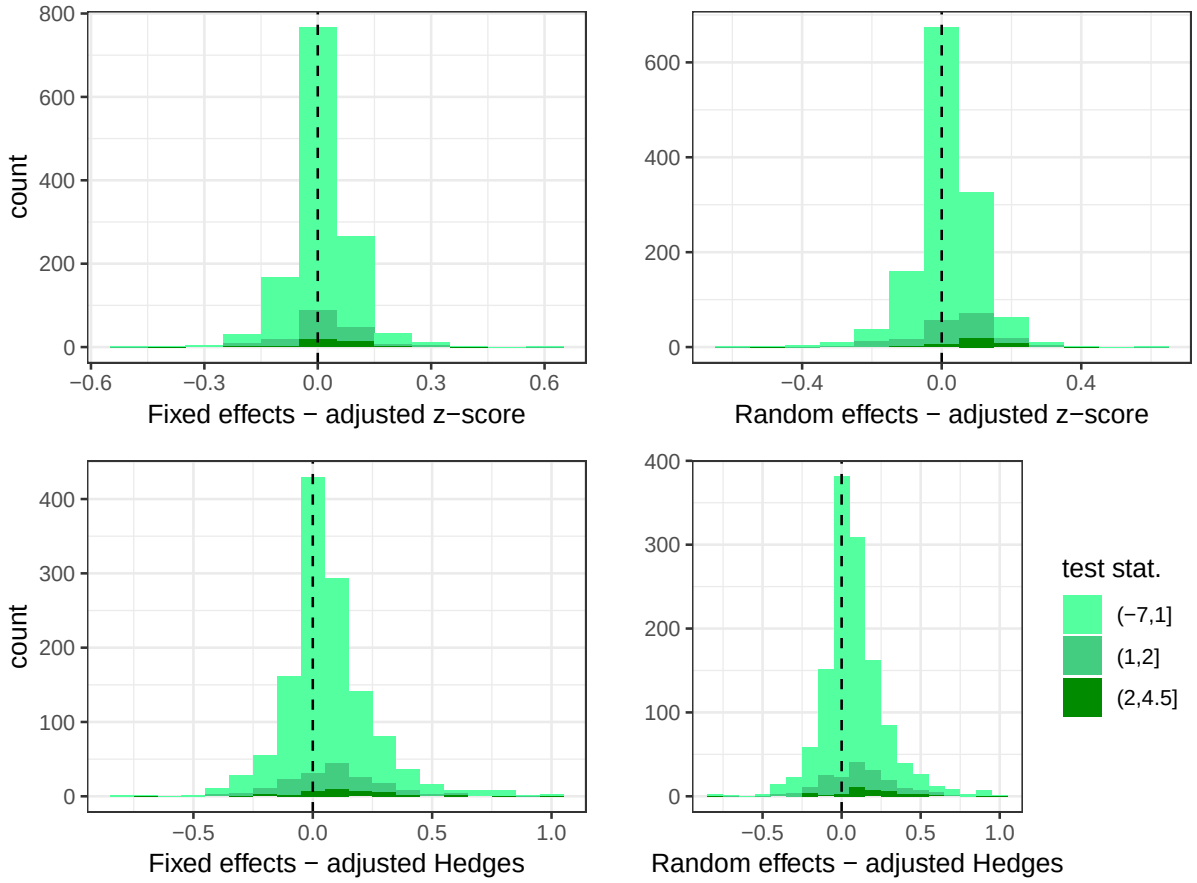

Figure 4.6: Histogram of the treatment effect differences between meta-analysis and regression adjusted meta-analysis. Negative differences indicate greater adjusted effect sizes than meta-analysis effect sizes. The bins are centered at zero and binwidth is equal to 0.1. Deeper green color indicates more evidence for small study effects.

$\hat{\theta}_{\text{Adj.}}$ is negative if the signs change in the original θ. Thus, a positive difference indicates a reduction of the original effect size, and the magnitude of the difference indicates the extent of the adjustment.
Additionally, the test statistic of heterogeneity adjusted publication bias tests (Rücker's and Thompson's test) are displayed with green color. Test statitics smaller < 1 (light green colored) are equivalent to no evidence for publication test, test statistics between one and two to weak evidence, and above two they indicate evidence for publication bias (dark green). An adjusted effect with evidence for publication bias can be regarded as a more realistic estimate of treatment efficacy. Some very large and very small differences have been omitted in the z-score and std. mean difference histograms; they are shown in Table 4.5.

Most often, adjustment leads to a reduction of overall treatment effect estimates, because the bins on the positive side of the histograms are larger. Adjustment is stronger when random effects meta-analysis is used as reference, because it gives larger weights to small studies.
Contrary to naive expectation, we see cases with large adjustment, but no evidence for small study effects. This is because the linear regression parameter estimates are large, but estimated with high uncertainity, such that there will be few evidence for small study effects (publication bias), but nonetheless, the adjustment will be large. The authors that developped the methods also recommend to use the methods in cases where there is clear evidence for publication bias. The color legend in Figure 4.6 thus gives a sense of confidence that there is in the credibility of

the adjustment.
Additionally, some meta-analyses with positively adjusted, larger overall treatment effects after adjustment (left side of the histogram) have at the same time evidence for publication bias. This is no wrong result, although the p-value of the test is one-sided, *i. e.* only allowing for publication bias for large effects (small studies with large effects). But the effects and their variances have been transformed, and it is possible that the shape of the funnel plot changes upon transformation; Figure 4.7 shows this for illustratory purposes. From the left to the right, the funnel plot is shown for mean differences (the original measure), standardized mean differences and Fisher's z scores. This is quite the most extreme case, with three different effects of adjustment, depending on outcome measure; there is no change with mean differences, reduction of effect sizes with std. mean differences and increase with Fisher's z transformed correlation. Note that while the rank of the effect sizes is usually preserved after transformation, the relative size and especially the variance may vary. One effect of the Fisher's z-transformation is that the effect sizes are bounded on $[-1, 1]$, and thus, very large effect sizes will influence the fit of the linear regression less than for example in std. mean differences which are not bounded. In contrast, the variance of the correlation is directly tied to the sample size, which makes it a suitable proxy for study size (variance of the mean difference is in contrast strongly influenced by the standard deviations). Thus, cases where the direction of adjustment changes upon transformation are not very rare.

Figure 4.8 shows the same for effect sizes adjusted by Copas; Copas selection model substitutes its estimates with random effect estimates when it finds no evidence for small study effects. Therefore, the effect of adjustment by Copas can be seen when comparing adjusted with random effects meta analysis estimates. Again, we clearly see that more effect sizes are adjusted downwards. Additionally, there is more coincidence between the publication bias test statistics and adjustment, which is as expected.

Table 4.3 shows quantiles and means for the various differences and the overall proportion of downward adjusted effect sizes. When Hedge's g is used as an effect measure, there are (substantially) more reduced effect sizes. The means in Table 4.3 suggest that the average reduction is small. To recall some other findings out of Table 4.3: 5% or 70 cases have their z-score reduced by more than 0.15 by regression adjustment (and 5% or 70 increased by -0.13, fixed effects reference). Also, std. mean difference is reduced by 0.39 compared to fixed effects estimates in 5% or 70 meta-analyses (or increased by 0.24).

	5%	25%	50%	75%	95%	mean	= 0 (%)	>= 0 (%)	> 0 (%)	No adj. est. (%)
z: Fixed - Copas	-0.04	-0.01	0.00	0.01	0.04	-0.02	8.03	49.40	41.36	64.61
z: Random - Copas	-0.00	0.00	0.00	0.00	0.04	-0.01	66.24	84.01	17.77	64.61
z: Fixed - Regression	-0.11	-0.03	0.01	0.05	0.13	0.01	0.00	51.31	51.31	0.00
z: Random - Regression	-0.13	-0.02	0.02	0.06	0.16	0.02	0.00	55.79	55.79	0.00
g: Fixed - Copas	-0.05	-0.01	0.00	0.01	0.11	-0.00	19.12	55.58	36.46	56.15
g: Random - Copas	-0.00	0.00	0.00	0.00	0.16	0.01	59.28	83.80	24.52	56.15
g: Fixed - Regression	-0.21	-0.03	0.04	0.14	0.40	0.06	0.00	59.35	59.35	0.00
g: Random - Regression	-0.20	-0.02	0.05	0.17	0.44	0.08	0.00	61.83	61.83	0.00
IV: Fixed - Copas	-0.05	-0.00	0.00	0.01	0.07	0.16	30.71	41.43	72.14	57.86
IV: Random - Copas	-0.01	0.00	0.00	0.00	0.10	0.21	62.86	25.71	88.57	57.86
IV: Fixed - Regression	-0.52	-0.09	-0.01	0.05	0.69	0.34	0.00	44.29	44.29	0.00
IV: Random - Regression	-0.52	-0.09	-0.01	0.07	0.76	0.39	0.00	45.00	45.00	0.00

Table 4.3: Quantiles and Means of the differences between meta-analysis pooled treatment effects and small study adjusted treatment effects. The column with the names "> 0" give the percentages of estimates larger than zero or larger or equal zero. The column "No adj. est." gives the percentage of missing estimates due to non-significant publication bias test (for Copas) and computational errors. The row names indicate which outcome measure, meta-analysis method and adjustment method is used. Abbreviations are used for z-score (= z) and Hedges g (= g).

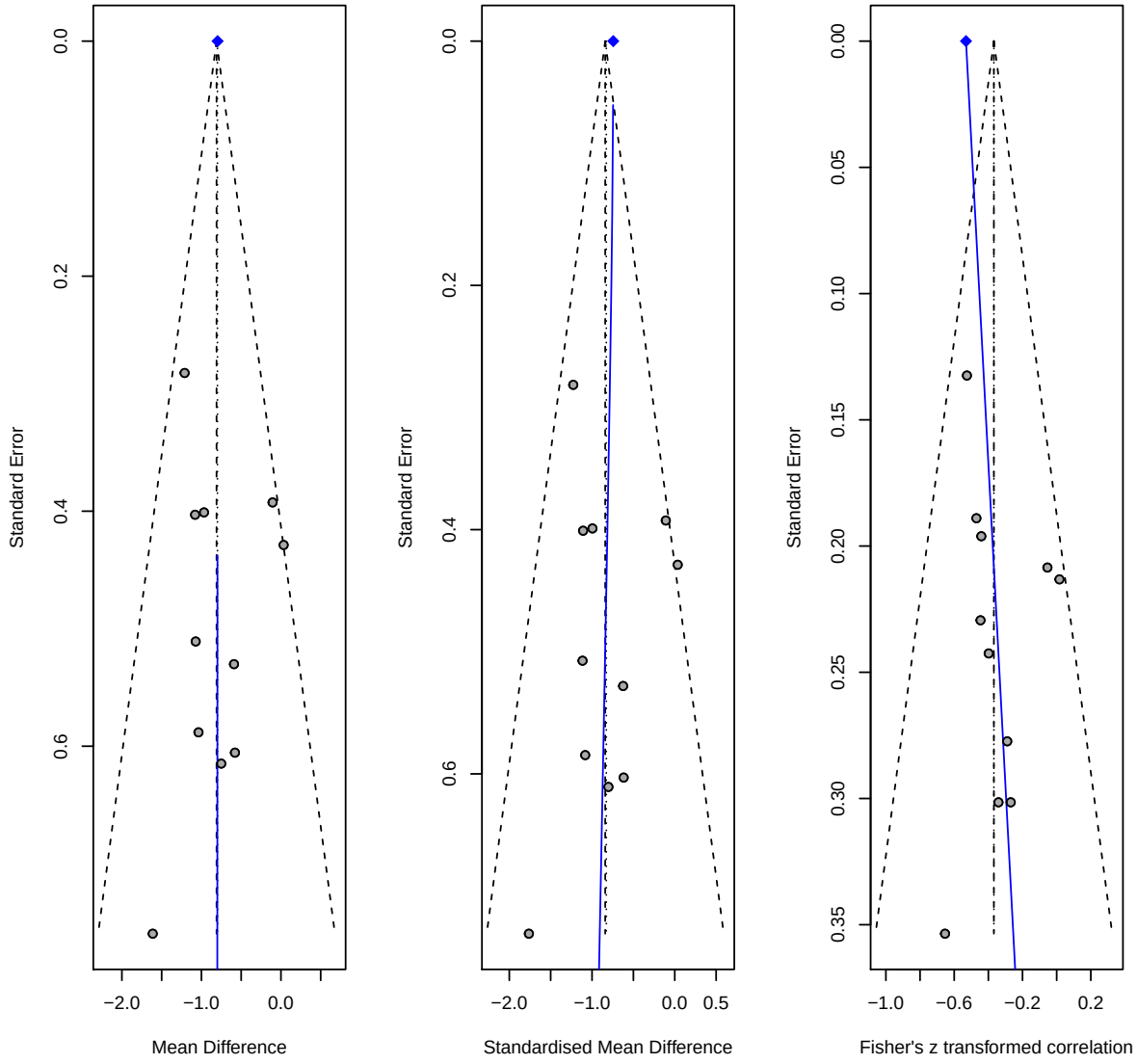

Figure 4.7: Funnel plots for a meta-analysis based on three different effect size measures: Mean differences, std. mean differences and Fisher's z transformed correlations and corresponding standard errors. Vertical dashed lines indicate meta-analysis estimates, the rhombus with the curved blue line the adjusted treatment effect.

4.2.2 Change in Evidence for Treatment Effects

Adjustment for small study effects in meta-analysis will not only provide new effect sizes, but also standard errors thereof. Thus, also the evidence for efficacy of a treatment can be obtained, which is usually summarized in a suitable test statistic or p-value. It is of interest if and how the evidence for treatment effects changes if adjusted for publication bias. There is an important difference between Copas selection model and regression adjustment. The regression adjustment method incorporates an additional parameter with corresponding uncertainity in the treatment effect estimate, while the application of the Copas algorithm does not so. Instead, the copas selection model derives the uncertainity of the estimate from the inverse of the fisher information matrix of the likelihood. Additionally, the Copas selection model algorithm provides only an estimate when it can detect a small study effect with reasonable precision at the first place (see

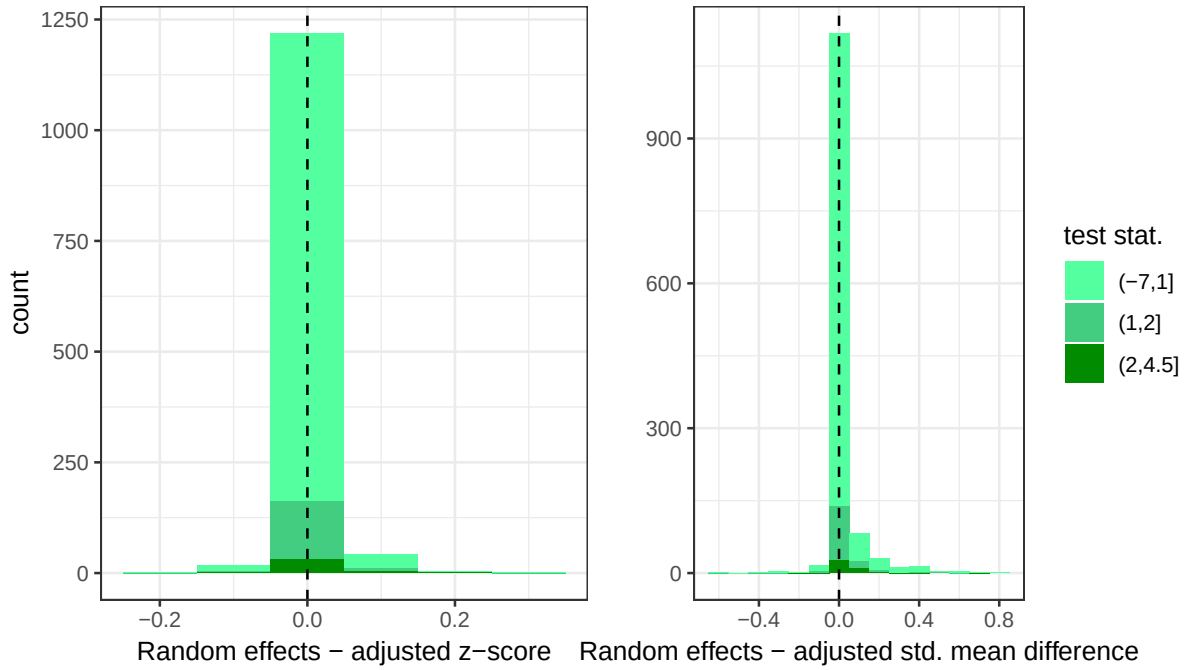

Figure 4.8: Histogram of the treatment effect differences between meta-analysis and Copas adjusted meta-analysis. Negative differences indicate greater adjusted effect sizes than meta-analysis effect sizes. The bins are centered at zero and binwidth is equal to 0.1. Deeper green color indicates more evidence for small study effects.

2.5.2).
The Wald test statistics p-value for fixed and random effects meta-analyses and Copas and regression adjusted treatment effects have been calculated. They are shown in Figure 4.9 for meta-analysis based on z-score, std. mean difference and other effect measures (`outcome.flag` = `IV`).
It can be seen that evidence for a treatment effect decreases after adjusting for publication bias. It does so more for Hedge's g compared to z-scores, and for regression adjustment compared to Copas selection model. In the case of other effect measures (`outcome.flag` = `IV`) adjustment, the adjustment has only negligible impact on the p-values.
Also, Copas selection method has a very small impact on the evidence compared to random effects meta-analysis, at least using z-scores (again, we should rely on comparisons to random effects meta-analysis to evaluate Copas method).

The Copas selection model also gives an estimate of the number of missing studies. It finds that 0 are missing, which corresponds to 11.3% from all 2.3274×10^4 analysed studies. Figure 4.10 shows a histogram of the overall fraction of missing studies. Note that we have excluded 0 out of 1407, for which no Copas selection model estimate, but a random effects estimate was retained because the algorithm initially found no evidence for small study effects.

We can see that in some occasions, the method finds more than half of all studies are missing. In most occasions, the estimate of missing studies is zero, as can be seen in Table 4.4. The discrepancy between mean and median may indicate that the estimate of 0% missing studies depends somewhat on these extreme cases. As can be written of from Table 4.4, 5%, *i. e.* 0 meta-analyses have 17.6 or more studies missing: in fact, these 5% most extreme make up for 0, more than 30% of missing studies.

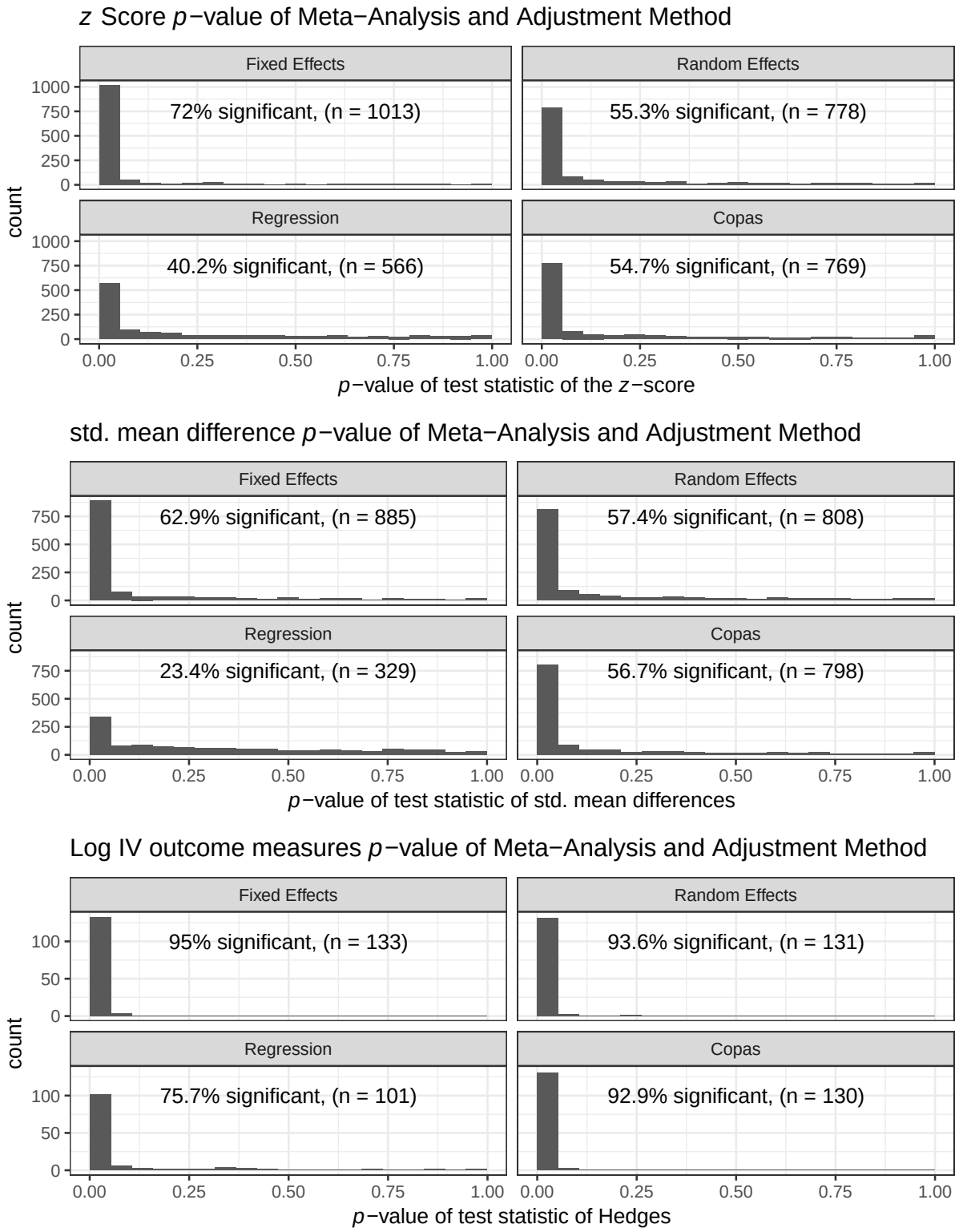

Figure 4.9: Histogram of the Wald test-statistic p-value of meta-analysis and adjusted pooled treatment effect, based on different treatment effect measures. The method is indicated in the header, bin width is set to 0.05. The significant proportion based on the threshold of 0.05 is displayed inside the figures.

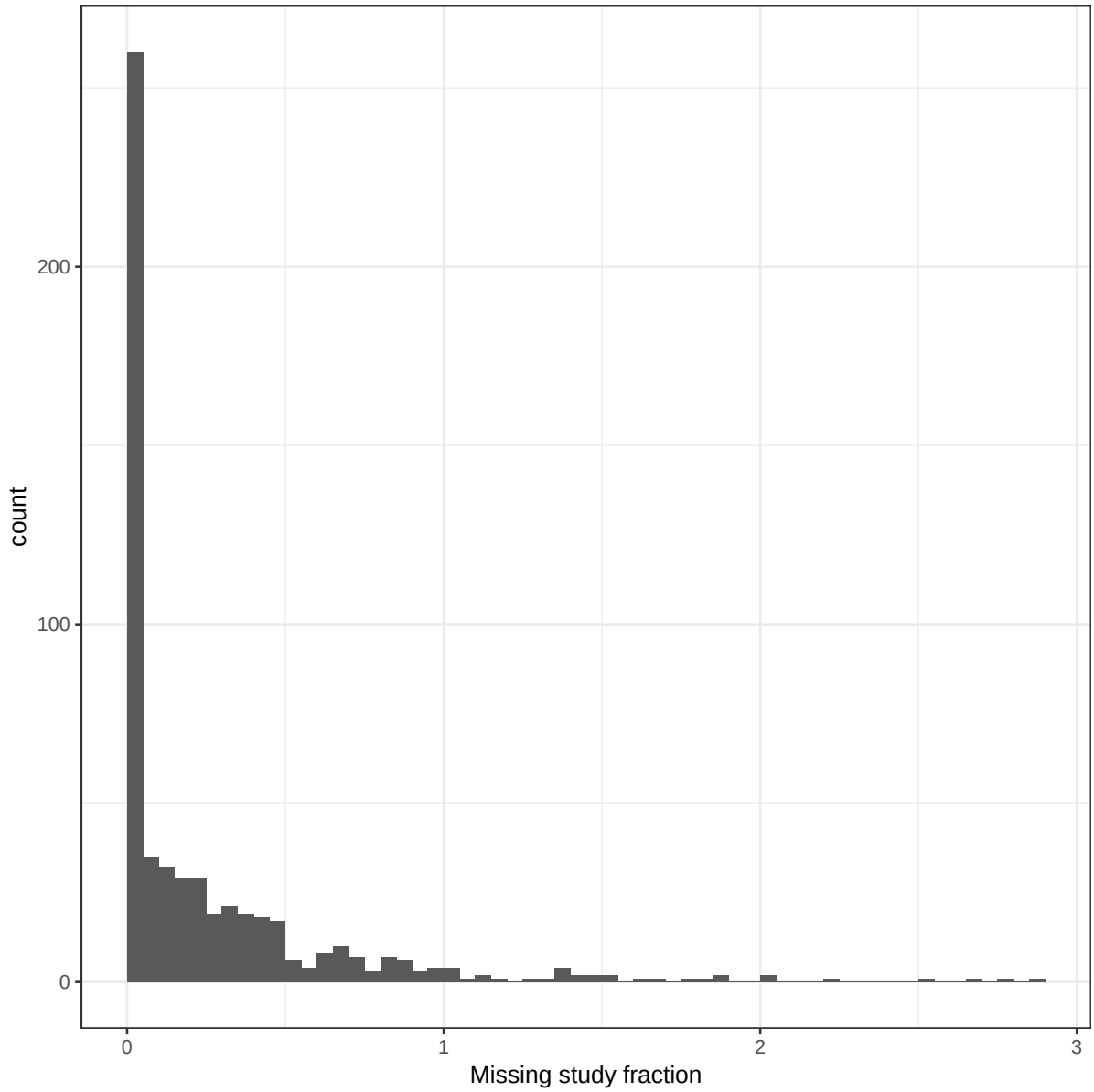

Figure 4.10: Histogram of the fraction of missing studies from the total number of studies in a meta-analyses (only data shown where Copas estimate was obtained, thus $n = 0$)

	$= 0$	5%	25%	50%	75%	95%	mean
Missing fraction	240	0	0	0.1	0.4	1.1	0.3
Missing study number	240	0	0	1.3	5.4	21.0	4.6

Table 4.4: Fraction of missing studies and estimates of missing studies with their zero counts ("= 0"), quantiles and means.

4.2.3 Comparison of adjustment methods

Regression adjusted estimates are compared to the estimates of Copas selection model if these are not equal to random effects meta-analysis. A Tukey mean difference plot can serve to reveal systematic differences and biases between the two measurement methods. We will compare estimates based on all three outcome measures, i.e. original measure, z-score and std. mean difference in Figure 4.11. Note that the 140 `IV` outcome data is not included when using z-score

meta.id	id	comparison.nr	subgroup.nr	z fixed	z random	z Copas	z regression	g fixed	g random	g Copas	g regression
9957	CD000370	8	2	0.66	0.66	0.66	0.82	1.59	1.59	1.40	0.28
22324	CD001183	7	0	-0.48	-0.48	-0.48	-0.21	-1.10	-1.12	-0.50	-0.11
49559	CD002307	2	1	-0.10	-0.08	-0.10	0.03	-0.47	-0.42	-0.41	-3.00
169178	CD007076	2	2	0.37	0.34	21.92	0.14	0.74	0.74	0.70	0.23
197675	CD008625	2	2	-0.60	-0.60	-0.60	-0.39	-1.72	-2.02	-1.01	-0.69
217035	CD009676	8	1	0.36	0.35	0.36	0.08	0.69	0.71	20.28	0.12
222767	CD010060	1	0	0.31	0.32	0.31	0.23	0.50	0.52	0.20	-0.49

Table 4.5: Missing meta-analysis pooled treatment effect and adjusted treatment effects. Abbreviations are used for z-score (= z) and Hedges g (= g).

and std. mean difference.

No formal tests are provided, but the at least there seems to be no clear bias or systematic error. The limits of agreement in Figure 4.11 are large. We conclude thus that the impact of regression adjustment on the effect sizes is in general not substantially larger than the impact of Copas selection model in the subset of data where the estimate of the Copas selection model is not equal to a random effects estimate. There is however a small difference, with regression estimates to have a little bit a larger absolute value. There might be some bias between adjusted z-scores, where regression estimates seem to be somewhat smaller when the mean is a little above zero, and somewhat larger when the mean is a little below zero.

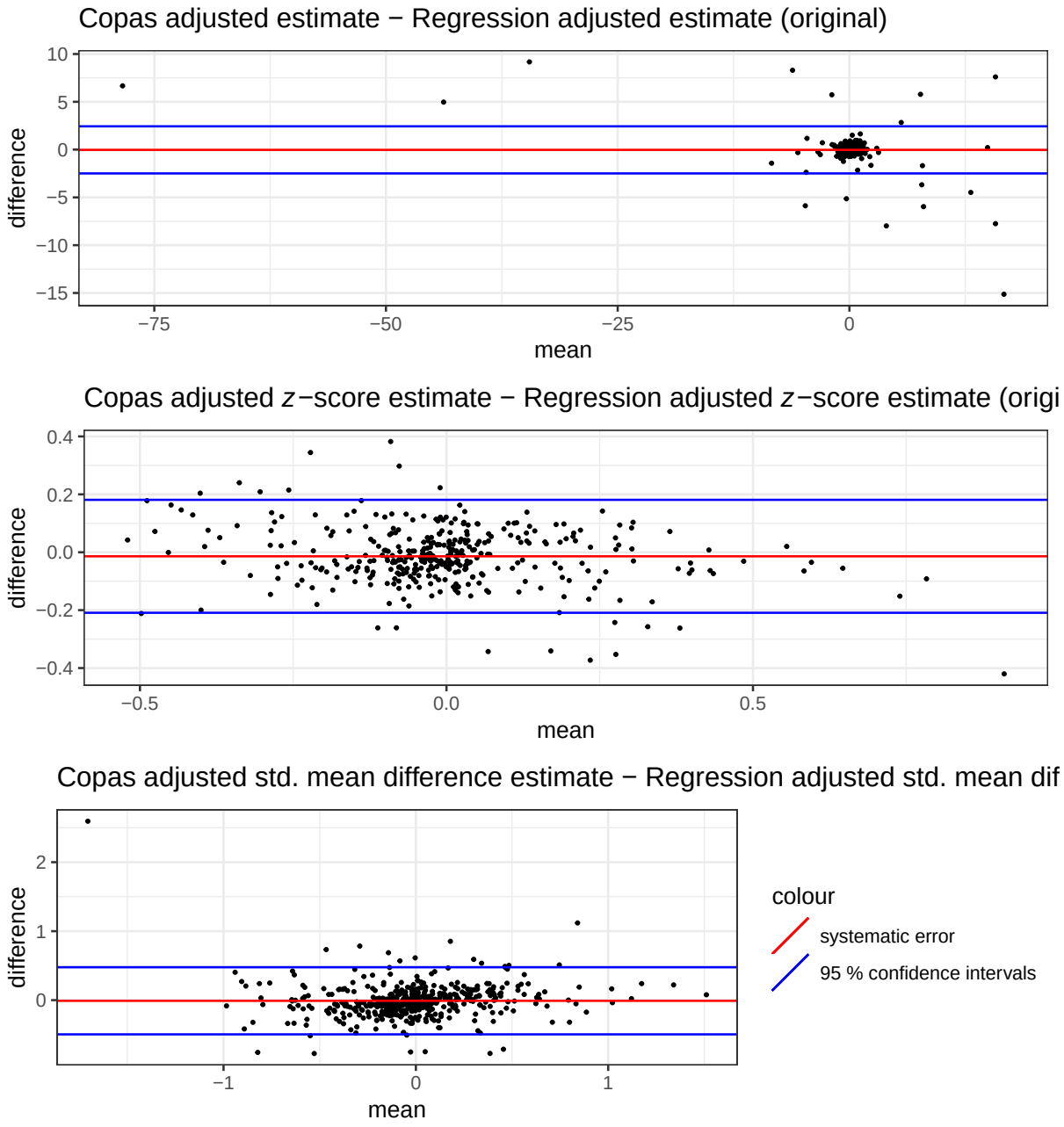

Figure 4.11: Mean - difference plots for publication bias adjustment methods. The mean of the adjusted treatment effects is displayed on the x-axis and the difference on the y-axis. Blue and red lines display the systematic error and the confidence intervals of the systematic error (limits of agreement). Two values have been omitted in the middle plot for std. mean difference and one for z-score (see Table 4.5).

Appendix A

Appendix

Maybe some R code here, probably a `sessionInfo()`

Bibliography

Begg, C. B. (1988). Statistical methods in medical research p. armitage and g. berry, blackwell scientific publications, oxford, u.k., 1987. no. of pages: 559. price Â£22.50. *Statistics in Medicine*, **7**, 817–818. 9

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, **568**, 435. 1

Borenstein, M., Hedges, L. V., Higgins, J. P. T., Rothstein, H. R., and Higgins, D. J. P. T. (2011). *Introduction to Meta-Analysis*. John Wiley & Sons, Incorporated, Hoboken, UNKNOWN. 6, 8, 16

Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis . *Biostatistics*, **1**, 247–262. 14

Copas, J. B. and Malley, P. F. (2008). A robust p-value for treatment effect in meta-analysis with publication bias. *Statistics in Medicine*, **27**, 4267–4278. 14

Copas, J. B. and Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research*, **10**, 251–265. 14

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall. 6

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187–202. 6

Decullier, E., Lhéritier, V., and Chapuis, F. (2005). Fate of biomedical research protocols and publication bias in france: retrospective cohort study. *BMJ (Clinical research ed.)*, **331**, 19–19. 1

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

Dickersin, K., Chan, S., Chalmersx, T., Sacks, H., and Smith, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials*, **8**, 343 – 353. 1

Duval, S. and Tweedie, R. (2000). Trim and fill: A simple funnel-plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**, 455–463. 8

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

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, **315**, 629–634. 1, 2, 10

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons. 8

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, **345**, 1502–1505. 1

Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, **7**, 889–894. 10

Harbord, R. M., Egger, M., and Sterne, J. A. C. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, **25**, 3443–3457. 11

Hedges, L. V. and Olkin, I. (1985). Chapter 11 - combining estimates of correlation coefficients. In Hedges, L. V. and Olkin, I., editors, *Statistical Methods for Meta-Analysis*, 223 – 246. Academic Press, San Diego. 16

Held, L. and Sabanés Bové, D. (2014). Applied statistical inference. *Springer, Berlin Heidelberg, doi*, **10**, 978–3. 6

Higgins JPT, G. S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration. 2, 17

Ioannidis, J. P. and Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, **58**, 543 – 549.

Ioannidis, J. P. and Trikalinos, T. A. (2007a). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *CMAJ*, **176**, 1091–1096. 2, 23

Ioannidis, J. P. and Trikalinos, T. A. (2007b). An exploratory test for an excess of significant findings. *Clinical Trials*, **4**, 245–253. 1, 12, 27

Ioannidis, J. P. A. and Thombs, B. D. (2019). A user's guide to inflated and manipulated impact factors. *European Journal of Clinical Investigation*, **0**, e13151. 1

Jones, C. W., Handler, L., Crowell, K. E., Keil, L. G., Weaver, M. A., and Platts-Mills, T. F. (2013). Non-publication of large randomized clinical trials: cross sectional analysis. *BMJ (Clinical research ed.)*, **347**, f6104–f6104. 1

Kasuya, E. (2001). Mann-whitney u test when variances are unequal. *Animal Behaviour*, **6**, 1247–1249. 6

Kicinski, M., Springate, D., and Kontopantelis, E. (2015). Publication bias in meta-analyses from the cochrane database of systematic reviews. *Statistics in medicine*, **34**, 2781–2793. 2

Law, U. P. (2007). *United States Code 110–85*. Food and Drug Administration Amendments Act. 1

McAuley, L., Pham, B., Tugwell, P., and Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *The Lancet*, **356**, 1228 – 1231. 1

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall. 11

McShane, B. B., Böckenholt, U., and Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, **11**, 730–749. 15

Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, **87**, 377–385. 7

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 24

Rosenthal, R. and Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500–504. 7

Rücker, G., Carpenter, J. R., and Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal*, **53**, 351–368. 13, 15, 25

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News*, **7**, 40–45. 24

Schwarzer, G., Antes, G., and Schumacher, M. (2007). A test for publication bias in meta-analysis with sparse binary data. *Statistics in Medicine*, **26**, 721–733. 12

Schwarzer, G., Carpenter, J. R., and Rücker, G. (2015). *Meta-analysis with R*, volume 4724. Springer. 9

Sterne, J. A. C., Egger, M., and Smith, G. D. (2001). Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, **323**, 101–105. 1

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, **18**, 2693–2708. 11

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., and Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, **358**, 252–260. 1

van Aert, R. C. M., Wicherts, J. M., and van Assen, M. A. L. M. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLOS ONE*, **14**, 1–32. 2, 24

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, **7**, 55–79. 7

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, **36**, 1–48. 24

Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, **10**, 1665–1677. 7

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. 24

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1. 24