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, August 13, 2019

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

Giuachin Kreiliger

Version August 13, 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. 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.* (2001a), Dwan *et al.* (2013)). 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) (Cochran, 1954). 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 The authors agree on that the exclusion of unpublished results in meta-analyses can lead to overestimation of treatment effects (*e. g.* Egger *et al.* (1997)). Although there are policies that make it mandatory to make all study results publicly accessible (US Public Law, 2007), it is not clear if the situation has improved yet. There are also notable efforts from both journals (Abbasi, 2004) and researchers (?) in the same direction.

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 (Dwan *et al.* (2013), Decullier *et al.* (2005), Lee *et al.* (2008)). One often finds that positive findings (*i. e.* large effects) are reported and published more often (see also an example in the 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. The rationale is that studies with larger standard errors have to have larger effects in order to be significant. The estimation of small study effects is an efficient way to investigate publication bias in a large number of meta-analyses. Although there are other reasons for small study effects as well, evidence for small study effects can oftentimes be interpreted as evidence for publication bias (Egger *et al.*, 1997), as it is the most likely cause.

A funnel plot (Egger and Smith, 1995) allows visual inspection of small study effects by 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

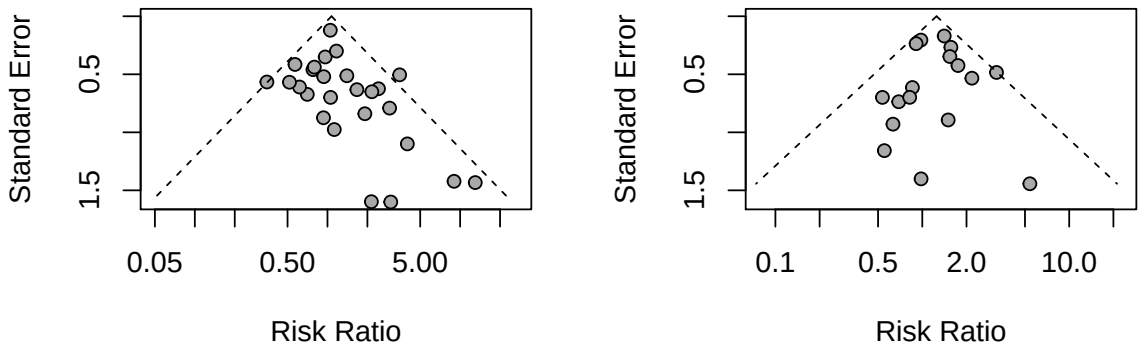

Figure 1.1: Funnel plots of two meta-analyses: On the left, the improvement in depression syndroms after application of tricyclic antidepressants is compared to placebo. The meta-analysis on the right measures the occurence of intracranial haemoerrhage by CT after application of any anti-thrombolytic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

linear regression, one can investigate how much evidence there is for this asymmetry.

The purpose of a funnel plot is to visualise if the findings of the studies disperse on both sides of the combined "true" effect or if they accumulate with increasing standard error on one particular side.

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. 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 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, log rate ratio, Cohen's d or standardized mean difference, Fisher's z transformed score.
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. 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. n_t and n_c are the total number of participants in the groups (c for control and t for treatment).

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 θ and its standard error se can be obtained as

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

Cohen's d and its standard error se can similarly be obtained by

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

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 ($n_t + n_c < 30$).

2.2.2 Binary Outcomes

Two commonly used effect measures for binary outcome data are risk ratios and odds ratios between treatment and control groups. 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. $\hat{\theta}$ and its variance $\hat{\text{se}}^2$ can be obtained by computing

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

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

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

Asuming binomial distribution of the events and 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). The approximation is only good if there are enough events and sample size is large enough.
Thus, with Φ as the cumulative standard normal distribution, we get

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

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).
Odds ratios can be transformed to std. mean differences, which will be described in Section 2.7.

2.2.3 Time-to-Event Outcomes

Usually, time-to-event data of two experimental groups can be compared by rate ratios. The normal approximations of the maximum likelihood estimators also works here when using the log rate ratio. 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

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

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 se_M^2 its variance. Then

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

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

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

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. The method assigns larger weights to studies with larger standard errors.
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 $n - 1$:

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

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

The estimators 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 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

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

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

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

After estimating θ_M with equation (2.5), the only problem remaining is to estimate τ^2. $\hat{\tau}^2$ 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. Using $\tau^2 + \tau_0^2$ for $\hat{w}_i, \hat{theta}$, we can update $F(\tau^2)$ and check convergence to zero.
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). For this, Q has to be computed as given in (2.7). 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. $= n-1$.
τ^2 is is directly linked to the variability in the data. The I^2 statistic of excess/total dispersion can be used alternatively to assess the extent of additional variance to the variances of the primary study estimates. It is computed as

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

The statistic takes value between zero and one, and is easily interpretable. 0 is equal to 0% excess dispersion and *e. g.* 0.5 equal to 50% additional between-study variance of the total variance of the estimates.
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 Linear, Weighted and Linear Mixed Regression Models

First, the concept of simple linear regression is introduced (Fahrmeir *et al.*, 2007). In short, the model assumes a dependent variable y to be a linear function of another explanatory variable x, with the residuals being distributed independently and identically and following a normal distribution:

$$y = \beta_0 + \beta_1 x + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{2.10}$$

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

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \Big(\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i \Big) \tag{2.11}$$

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

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{2.12}$$

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

$$\hat{\sigma}^2 = \frac{1}{n-2} \mathbf{r}^\top \hat{\mathbf{r}} \qquad\qquad \hat{\mathbf{s}}^2_{\hat{\beta}} = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} \tag{2.13}$$

If one plots the values of y and x, 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:

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

$\bar{x}$ and $\bar{y}$ denoting the sample means of the corresponding values $x_1, .., x_n$ and $y_1, ..., y_n$. Thus, $\hat{\beta}_0$ is also called the 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. The p-value 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 (heteroscedasticity). The least squares equation (2.11) is extended as follows:

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

with the positive weights w_i penalizing large squared residuals for some i more if w_i is larger. 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 (2.12) and (2.13) can again be used if $\mathbf{X}$ is exchanged with $\mathbf{X}^\star = \mathbf{W}\mathbf{X}$

and $\mathbf{y}$ with $\mathbf{y}^\star = \mathbf{W}\mathbf{y}$.
The introduction of group-specific random effects allows to analyze grouped or repeated measurements by linear regression. Let j be the group index, i the index of the single observation from the group j and x_i the i^{th} row of $\mathbf{X}$, $(1, \mathbf{X}_{i2})$. Then the equation

$$\mathbf{y}_i | U_j, \epsilon_i = x_i\beta + U_j + \epsilon_i \tag{2.14}$$

gives the marginal distribution of $\mathbf{y}_i$ depending on U_j and ϵ_i. The random effects $U_j \sim N(0, \mathbf{G})$ and the residual error term $\epsilon_i \sim N(0, \tau^2)$ are independent from each other. Let $\mathbf{y}_j$ be a vector of all observations of group j. The expectation $\mathbb{E}(\mathbf{y}_i)$ is still $x_i\beta$, but the covariance between the observations within a group j, is modeled as

$$\text{Cov}(\mathbf{y}_j) = \mathbf{D}_j\mathbf{G}\mathbf{D}_j^\top + \tau^2\mathbf{I}_{n_j} \tag{2.15}$$

where $\mathbf{D}_j$ is in the case of the random intercept model a $n_j \times 1$ matrix of 1's, $\mathbf{G}$ is a scalar to be estimated and $\mathbf{I}_{n_j}$ is an $n_j \times n_j$ identity matrix. Thus, if $\mathbf{G} \neq 0$, the observations within a group will be uniformly correlated (uniform correlation between each observation).
Nested groups can be modelled by extending the design matrix $\mathbf{X}$ with an additional column of 1's and using according indices that specify the nesting structure.
An extension of the random intercepts model is the random slopes model, which allows for additional, group-specific slopes with respect to a explanatory variable x. It can be implemented by modifying (2.14). Let $\mathbf{U}_j$ be a two-dimensional random vector with a 2×2 covariance matrix $\mathbf{G}$. Again, we can specify the marginal distribution of a observation i within a group j:

$$\mathbf{y}_i | U_j, \epsilon_i = x_i\beta + x_i\mathbf{U}_j + \epsilon_i \tag{2.16}$$

$\mathbf{D}_j$ is a $n_i \times 2$ equal to all rows of observations of j in $\mathbf{X}$. The covariance matrix $\mathbf{V_j}$ for group j is defined as in equation (2.15), using the new $\mathbf{D}_j$. Defining $\mathbf{x}_j$ as the matrix with all observations from group j of $\mathbf{X}$, and equivalently the vector $\mathbf{y}_j$, $\hat{\beta}$ is obtained by computing

$$\hat{\beta} = \Big(\sum_{j=1}^{m} \mathbf{x}_j^\top \hat{\mathbf{V}}_j^{-1}\mathbf{x}_j\Big)^{-1} \sum_{j=1}^{m} \mathbf{x}_j^\top \hat{\mathbf{V}}_j^{-1}\mathbf{y}_j \tag{2.17}$$

with m being the number of groups and $\hat{\mathbf{V}}_j$ being the estimated covariance, obtained by maximizing the log likelihood of the normal distribution, as $\mathbf{y}_j$ is distributed as

$$\mathbf{y}_j \sim N(\mathbf{x}_j\beta, \mathbf{V}_j) \tag{2.18}$$

The matrices $\sum_{j=1}^{m} \mathbf{x}_j^\top \hat{\mathbf{V}}_j^{-1}\mathbf{x}_j$ and $\hat{\mathbf{V}}_j$ are assumed to be positive definite . The approximate covariance matrix of $\hat{\beta}$ is given by

$$\hat{\text{Cov}}(\hat{\beta}) = \Big(\sum_{j=1}^{m} \mathbf{x}_j^\top \hat{\mathbf{V}}_j^{-1}\mathbf{x}_j\Big)^{-1} \tag{2.19}$$

Weights can be introduced by replacing $\mathbf{I}_{n_j}$ in (2.15) with the previously introduced weight matrix $\mathbf{W}_j$. Hypothesis tests for $\beta = \beta_{\text{H0}}$ can be made using the Wald method.

2.5 Publication Bias Tests

The tests that will be presented on the following pages are a common way to detect publication bias. A frequently to test and adjust for publication bias, which goes under the name of trim-and-fill (Duval and Tweedie, 2000) is not discussed because of its disadvantageous properties, therefore, it will not be discussed here (see *e. g.* Moreno *et al.* (2009)).

2.5.1 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 (2.20). $\mathrm{se}_i^{2\star}$ is the variance of $\theta_i - \theta_M$ as defined in (2.21) and θ_M is the fixed effects pooled treatment effect ((2.5).

$$\theta_i^\star = (\theta_i - \theta_M)/\,\mathrm{se}_i^{2\star} \tag{2.20}$$

$$\mathrm{se}_i^{2\star} = \mathrm{se}_i^2 - 1/\sum_{i=1}^{n} \frac{1}{\mathrm{se}_i^2} 1 \tag{2.21}$$

A rank correlation test based on Kendall's tau is then used. First, the pairs are ordered after their ranks based on $\mathrm{se}^{2\star}$. Then, for each $\mathrm{se}^{2\star}$ 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 concordant pairs will reflect in large $\hat{u}$ and small $\hat{l}$ and thus lead to a large $\hat{Z}$. A two-sided p-value is obtained using the standard normal distribution Φ:

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

A one-sided test for positive correlation is obtained by computing $1 - \Phi(Z)$ instead. The changes that have to be made int the case of ties are small and can be found in (Begg, 1988, 410).

2.5.2 Egger's Test: Weighted Linear Regression Test

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 $\theta/\,\mathrm{se}$ be the dependent variable y and $1/\,\mathrm{se}$ the explanatory variable x. If plotted, this corresponds to a radial or Galbraith plot (Galbraith, 1988). The linear regression equation as introduced before in (2.10) can be written in two ways:

$$\theta/s = \beta_0 + \beta_1/\,\mathrm{se} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{2.22}$$

Equation (2.22) is often provided due to the correspondence to the radial plot. However, it is equivalent to

$$\theta = \beta_0 + \beta_1\,\mathrm{se} + \epsilon, \qquad \epsilon \sim N(0, w^{-2}\sigma^2) \tag{2.23}$$

with weights $w = 1/\,\mathrm{se}^2$. Thus testing β_0 of (2.22) or β_1 of (2.23) is equivalent. The corresponding p-value is then used as evidence for a small study effect. Plugging in $\theta_i/\,\mathrm{se}_i, 1/\,\mathrm{se}_i$ as y_i, x_i into equations (2.12) and (2.13) will give the estimates for $\hat{\beta}_0, \hat{\beta}_1, \hat{\mathrm{se}}_{\beta_0}$ and $\hat{\mathrm{se}}_{\beta_1}$.

2.5.3 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 $y = \theta/\operatorname{se}$ by introducing new weights. The effect size θ_i is assumed to be distributed as

$$\theta_i \sim N(\beta_0 + \beta_1 \operatorname{se}_i, \operatorname{se}_i^2 + \tau^2) \tag{2.24}$$

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, we can proceed analogous to Egger's test. The p-value for $\beta_0 \neq 0$ reflects the evidence for a small study effect.

2.5.4 Peters Test: Weighted Linear Regression Test

When the outcome is dichotomous, effect sizes and variances of effect size are correlated, which can readily be seen in (2.3) and (2.4) (see also (Schwarzer *et al.*, 2015, 120)). 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.
Instead of taking the standard error se 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.
The method will give less false positives than Egger's test, but will be more imprecise, because total sample size is not a very good approximation for statistical power (the overall rate of events in both groups plays an important role as well).

2.5.5 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 $\theta_{\text{H0}} = 0$ and its inverse Fisher information s^2. Formally,

$$\varphi = e_t - (e_t - e_c)(e_t + (n_t - e_t))/(n_t + n_c) \tag{2.25}$$

$$s^2 = \frac{(e_t + e_c)(e_t + (n_t - e_t))(e_c + (n_c - e_c))((n_t - e_t) + (n_c - e_c))}{(n_t + n_c)^2(n_t + n_c - 1)} \tag{2.26}$$

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 $\varphi_i/\operatorname{se}_i^2$ is also known as Peto odds ratio. The obtained scores and variances can be used in Egger's test as treatment effects and variances.

2.5.6 Schwarzer's Test: Rank Correlation Test

Schwarzer *et al.* (2007) developed a test for the correlation between the event counts in the treatment group and the expected event counts $E_t - \mathbb{E}(E_t)$. When the marginals in a two-by-two table and the log odds ratio is fixed, it can be shown that E_t follows a non-central hypergeometric distribution. Using the Mantel-Haenszel log odds ratio and the marginal total parameters, the variance and the expectation of E_t is calculated. The inverse of the variance $\operatorname{se}(E_t)^2$ and the standardized event count

$$(e_t - \mathbb{E}(E_t))/\sqrt{(}\,\mathrm{se}_i^2) \tag{2.27}$$

are then used as before in Begg and Mazumdar's test.

2.5.7 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

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

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

2.5.8 Excess Significance Test

Publication bias does not need to be accompained by small study effects. In the absence of any treatment effect, significant effects could be included in a meta-analysis on both directions of treatment effects (*i. e.* large *and* small effects) are published. Also, the afore-mentioned methods do not use statistical significance directly to investigate publication bias. Thus, a different test is introduced.
Ioannidis and Trikalinos (2007b) developed an exploratory test to detect if the proportion of significant findings is larger than expected. 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 be 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:

$$1 - \hat{\beta}_i = F(z_\alpha) \tag{2.28}$$

with F being the cumulative normal distribution 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 computing

$$p = \sum_{i=O}^{n}\left(\binom{n}{i}p^i(1-p)^{n-i}\right) \tag{2.29}$$

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

2.6 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 are used for a sensitivity analysis, where the selection parameters are assumed to be fixed.

2.6.1 Adjustment by Regression

In Rücker *et al.* (2010) and Rücker *et al.* (2011), a random effects model is proposed to obtain unbiased treatment effect estimates. Similarly to regression based tests for small study effects, we have

$$\theta_i = \beta_0 + \beta_1 \sqrt{\mathrm{se}_i^2 + \tau^2} + \epsilon_i \sqrt{v_i + \tau^2}, \tag{2.30}$$

$$\epsilon_i \overset{\text{iid}}{\sim} N(0,1) \tag{2.31}$$

The only difference between Thompson and Sharp's variant and this method is that $x = \sqrt{\mathrm{se}^2 + \tau^2}$ is used instead of $x = \sqrt{\mathrm{se}^2}$. β_1 is the bias parameter and can be interpreted as 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.There are two possible estimates at hand:

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

Simulations in Rücker *et al.* (2011) suggested that the latter estimate is slightly superior to the former, because it showed a smaller mean-squared error in simulation studies Rücker *et al.* (2010). From the formulas above, it becomes clear that is has a larger standard error. The results of the simulation furthermore emphasize that adjustment is more reliable when effects of publication bias are strong within a meta-analysis. When no publication bias affects the meta-analysis, and event rates are low, the method provided biased estimates and the mean squared error was large. The coverage of the confidence intervals was always larger or equal to meta-analysis. It outperforms or is equal to classical meta-analysis estimates with respect to coverage, mean squared error and bias if there is publication bias, especially if event rates in the control group are small.

2.6.2 Copas Selection Model

A method proposed in Copas and Shi (2001, 2000); Copas and Malley (2008) assumes that study results are selected based on the specific properties of their effect sizes and variances.
Let θ_i be the effect size estimate of study i. Then

$$\theta_i \sim N(\mu_i, \sigma_i^2) \qquad \mu_i \sim N(\theta, \tau^2) \tag{2.32}$$

which is identical 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. This is termed the *population model*.
The *selection model* is defined as follows. Suppose a selection of studies with reported standard errors se (possibly different from σ). Only a proportion of the selection will be published, with the parameter a defining the overall proportion of published studies and b (assumed to be positive) defining how fast this proportion increases with se becoming smaller. Formally, the probability of selection given a reported standard error se is defined as

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

The equation can be rewritten as

$$z = a + b/\,\text{se} + \delta$$

with $\delta \sim N(0, 1)$. z is interpreted as the *propensity for selection*. It is defined that the sign of z must be positive in order for the study to be selected. a is some kind of global selection rate for each study and b decides about the decline of selection probability with increasing se.
So far, we have, for a study i

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

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 unequal to zero and positive and $z_i > 0$, then the estimate of a study i that is selected is likely to have positive δ_i and thus positive ϵ_i, such that the true mean μ is likely to be overestimated. Let $u_i = a + b/\,\text{se}_i$, $\lambda(u_i)$ 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

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

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

$$\mathbb{E}(\theta_i \mid \text{se}_i, \text{select}) = \mu + \rho_i \sigma_i \lambda(u_i) \tag{2.33}$$

increases for larger σ.
A likelihood for θ_i, conditional on $z > 0$ can be formulated to estimate the parameters of the model. a and b are not estimated because the number of missing studies and their effect sizes is not known. Instead, fixed values for a and b have to be imputed. The nuisance parameter σ_i can be estimated, as

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

with $c^2 = \lambda(u_i)(u_i + \lambda(u_i))$. Thus we can replace σ_i^2 by $\hat{\sigma}_i^2 = \frac{1}{1-c_i^2\rho_i^2}$.
With equation (2.33), one can obtain fitted values of θ_i based on se_i and fixed a and b. For two different pairs (a, b), $(a^\star, b^\star)$,

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

Local departures of two fitted values of θ_i can be approximated by adding a linear term in se_i to the expectation of θ_i. Thus, to test a single pair (a, b) (chosen such that $\rho \geq 0$), it is sufficient to test $\beta \neq 0$ in

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

If $\beta \neq 0$, there is still bias in the fitted values. To test a pair (a, b) against the scenario with no selection, we set $a^\star = \infty$ (or $\rho = 0$) and $b^\star = 0$. A likelihood ratio test will give a test statistic to test against $H0 =$ no selection ($\beta \neq 0$):

$$\chi^2 = 2 \cdot (\max_{\theta,\tau,\beta} \tilde{L}(\theta, \tau, \beta) - \max_{\theta,\tau} \tilde{L}(\theta, \tau, 0)) \tag{2.34}$$

with

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

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 very similar to Egger's small study effect test when $\tau^2 = 0$.
In practice ne can observe how θ and it's confidence intervals change dependent on the underlying selection process, and how the choice of the parameters affect the evidence for remaining publication bias (selection). Rücker *et al.* (2011) used the method in a simulation for inference purposes, and have implemented it in the Schwarzer (2007). However, in the simulations, the method was outperformed by the regression adjustment method when publication bias was present. Especially when event rates in the control group were small and publication bias was strong, bias, mean squared error and coverage was substantially worse. Other authors argue that selection models should in general not be used for inference (*e. g.* McShane *et al.* (2016)).
The procedure in Rücker *et al.* (2011) is the following: A range of values of (a, b) are applied, and the test for residual small study effect as described in equation (2.34) 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 adjustment, 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 pair of a and b that implies the fewest publication bias is chosen).
Currently, there is no test to detect miss-specifications in the model itself and the authors themselves have argued that a non-parametric test of the residuals would lack power.

2.7 Transformation between Effect Measures

Assuming that binary outcomes result from a dichotomization of originally continuous random variables, binary outcome measure can be transformed into continuous outcome measures. Here, the logistic distribution is used to achieve the transformation from a typical binary effect measures to a std. mean difference (Borenstein *et al.*, 2011, 47).

Let θ be a log odds ratio and se it's standard error. The std. mean difference d and it's variance se_d^2 can be obtained as

$$d = \theta \frac{\sqrt{3}}{\pi} \qquad \text{se}_d^2 = \text{se}^2 \frac{\sqrt{3}}{\pi}$$

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.
The Pearson's correlation coefficient can be attained by the formulas (Hedges and Olkin (1985), (Borenstein *et al.*, 2011, 48))

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

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

$$z = \arctan(r) \qquad s_z^2 = \frac{1}{n - 3}$$

Chapter 3

The Cochrane Dataset

3.1 Cochrane Systematic Reviews

Cochrane has specialized on systematic reviews in clinical science. Certain knowledge of standards and principles of the organization 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 Cochrane 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. Cochrane 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 Library 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. A result is a outcome of a study, thus studies can contribute multiple results. The studies too a very large extent randomized and investigate the effects of healthcare and medical interventions and treatments. The reference to the treatment may be a placebo control or a different intervention or treatment. A result can not only be about efficacy of the treatments, but also about safety (adverse effects).
It will be continued with an example form the dataset. 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

Results are part of studies that are again part of a (systematic) review. This structure of a review is shown in Figure 3.1.

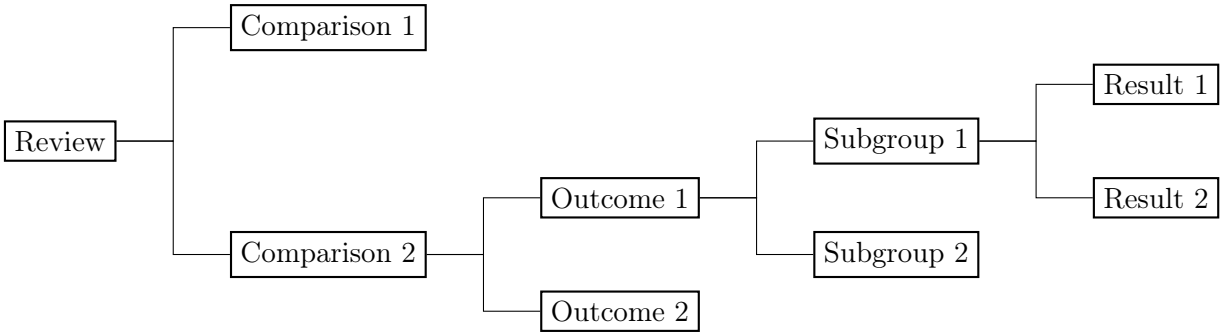

Figure 3.1: Structure of a hypothetical review with two different comparisons

A listing of important 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.

The structure of a review will now be outlined based on an example of the dataset. The previously mentioned barbiturate and head injury review will be outlined. 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.

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.

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.
For results to be suitable for usage in meta-analysis, they have to be identical with respect to

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.
`comparison.name/.nr`	Specification of the interventions compared in the study and a unique number for the comparison.
`comparison.id`	Specification of the comparison by a string of the type "CMP-xxx" for the xxx. comparison within the review.
`outcome.name/.nr`	Specification by which outcome the interventions are compared and a unique number for the outcome.
`outcome.id`	Specification of the outcome by a string of the type "CMP-xxx.xx" for the xx. outcome of the xxx. comparison within the review.
`subgroup.name/.nr`	Potentially indication of affiliation to subgroups and a unique number for the subgroup.
`subgroup.id`	Specification of the comparison by a string of the type "CMP-xxx.yy.xx" for the xx. subgroup for the yy. outcome of the xxx. comparison within the review.
`outcome.measure`	Indication of the quantification method of the effect (of one intervention compared to the other).
`outcome.measure.merged`	Indication of the quantification method of the effect, merged such that each method is uniquely classified.
`outcome.flag`	A outcome flag to simplify programming; `DICH` for binary outcomes with fully available information on event counts, as given in a two-by-two table. `CONT` for continuous outcomes with available means and standard deviations. `IV` for all results with effects and standard errors but without the data necessary for their computation. `IPD` for individual patient data.
`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 (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

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 Cochrane, *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.
```

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

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

3.2 Data Tidying and Processing

3.2.1 Newly Introduced Variables

Some new variables are added to the obtained dataset:

- `lrr` and `var.lrr`: log risk ratio and variance of the log risk ratio for `outcome.flag DICH`.
- `cohensd` and `var.cohensd`: Cohen's d and the variance 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 DICH`.
- `cor.Pearson` and `var.cor.Pearson`: Pearson correlation coefficient and variance as obtained from the d (for `outcome.flag DICH`) or d (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 DICH`.
- `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

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,

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.

comparison, outcome and subgroup (optional, subgroups not given in any case) may be combined in a meta-analysis.

- `smd.pool` and `se.smd.pool`: Depending on `outcome.flag`, `smd.pool` is equal to `smd.ordl` (`outcome.flag` = `DICH`), `cohensd` (`CONT`), or `effect` (`IV` and `outcome.measure.merged` = `SMD`). 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`. If `IV`, the only meta-analyses with `outcome.measure.merged` = `OR` / `RR` / `SMD` / `MD` / `Hazard Ratio` / `Rate Ratio` are used. 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 = 9772 remaining).
- **Study size**: The ratio between largest variance of an estimate and smallest variance of an estimate is larger than four (n = 9473 remaining).
- **Significance**: At least one treatment effect has a p-value below the significance threshold 0.05 (n = 7452 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 = 1388 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).
- **No withdrawn reviews**: Reviews that have been withdrawn are not included.
- **No adverse effects**: No results of adverse effects of treatment are used for meta-analyses. Exclusion is incomplete because only results with "adverse" in the comparison or outcome name could be removed, but not all adverse effect results are classified as such.

The results of this reduction of the dataset are shown in the flow-chart in Figure 3.2. Only the data that had accessible all results data available was used for adjustment. Thus, all meta-analyses with `outcome.flag` == `IV` and `outcome.measure.merged` not equal to `SMD` are omitted in a second step of the analysis (Analysis dataset (1) and (2) in Figure 3.2).

Exclusions for other Reasons

In some more cases, meta-analyses were excluded because their results seemed to be erroneous (1) or because participant numbers were missing (5), necessary to compute the variance of Fisher's z transformed correlation coefficient. The latter meta-analyses (all `outcome.flag` = `IV`) were only omitted when doing the adjustment based on the z-scores, but kept when adjusting based on std. mean differences. The meta-analysis that is considered erroneous is from a review with the title "School-based programs for preventing smoking". The reason why it is suspected to be erroneous is because there are results from one study (Severson, 1991) with rather large effects (13, 0.4, 6.3, 2.4) compared to a mean std. mean difference in 12 other results of 0.22 (max.: 0.91) and very large standard errors (64, 60, 46, 70) compared to a mean `se` of 0.6 (max.: 3.8). It is not included in the analysis.

The Analysis Dataset

The dataset that was ultimately tested and adjusted for publication bias comprises 1388 meta-analyses and 22,937 results. The mean number of participants in the treatment group is 253.5 vs 256.7 in the unrestricted dataset and the mean total number of participants 589.3 vs 623.2. The mean publication year is 2000 vs 2001 in the unrestricted dataset.
From the meta-analyses with incomplete data (`outcome.flag` == `IV`), there are 36 with `outcome.measure.merged` = `RR`, 26 `SMD`, 25 `Hazard Ratio`, 15 `OR`, 11 `MD` and 8 `Rate Ratio`.

3.2.3 Analysis Procedure

Meta-analysis groups within a review were obtained by grouping by comparison, outcome and subgroup. The last step is optional, and the Cochrane Groups do also meta-analysis without taking into account subgroups. Subgroups are often specifications of treatment, *e. g.* the form of the medication or the protocol. Thus, using the subgroups should increase the within-study homogeneity. However, the Cochrane disencourages meta-analyses based on subgroups only, because subgroups are often set up after the collection of the data analysis. Since the main interest of this study is not to precisely assess the efficacy of treatments, but the interest is merely on systematic differences between single studies investigating the same scientific question and not

in the effect per se, it has been decided that the increase in precision is worthwhile the loss of a number of studies. While it is possible that the use of subgroups introduces bias in the meta-analysis, it is also possible that ignoring subgroups will do (which is one of the reasons why subgroups are analyses as well). When one subgroup has substantially different real treatment effects, and the specific properties influence study size, publication bias assessment fails. Subgroups are not indicated in 50.8% of the meta-analyses in the analysis dataset 1, thus, there the analyses are identical to the overall analyses that the Cochrane Review Groups did. The methods described in the methods chapter 2 most often apply directly to the algorithms used in `meta` and `metafor`.

For applying publication bias tests and adjustment methods, the analysis was applied such that the effects used are similar or identical to the effects on which journal editors decided upon publication bias. For `outcome.flag` = `DICH`, log risk ratios were used as treatment effect estimates in the meta-analysis, for `outcome.flag` = `CONT`, mean difference or standardized mean difference, depending on `outcome.measure.merged`. For `outcome.flag` = `IV`, the `effect` and `se` as provided in the dataset was passed to `meta::metagen`. Since the intuition behind small study effect tests is that effects with larger standard errors have to be larger to reach significance, the publication bias tests and adjustments are most meaningful if the effects are used in their original scale. Transformation will change the relative size of uncertainty estimates and effects. To be able to compare the effects of adjustment among meta-analyses, the effect sizes were transformed on a common scale, as described in section 2.7. As shown in the flow-chart in Figure 3.2, this leads to a reduced dataset (compare analysis dataset (1) and (2)). 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 pre-specified. 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 2.6.2 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`.

All computations 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).

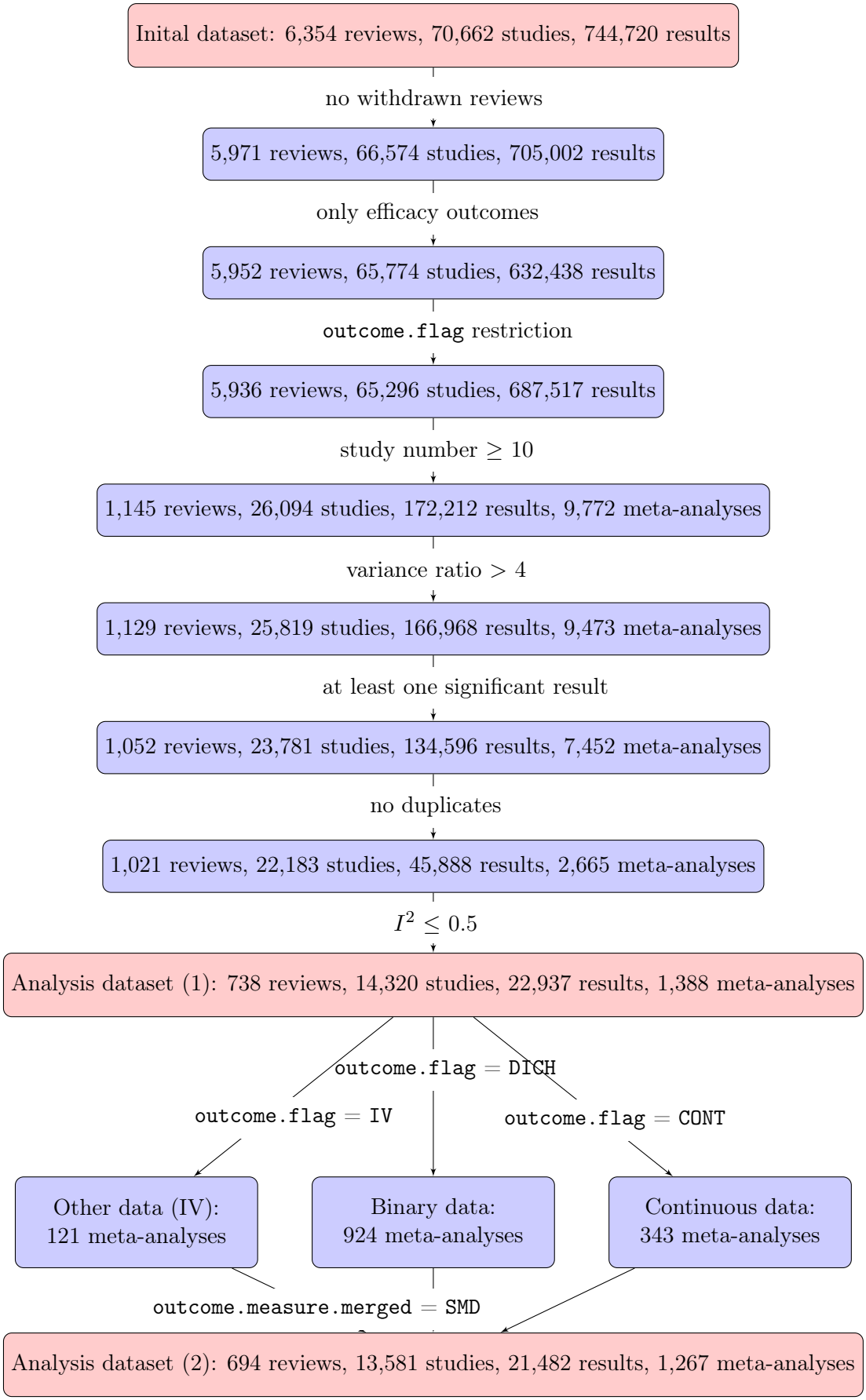

Figure 3.2: Flow-chart of the exclusion of meta-analyses for the final analysis. The exclusion criteria are given at the right of the arrows.

Chapter 4

Results

Meta analyis are based on results of primary studies. Therefore, in 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. The medians are calculated separately for efficacy and safety outcomes. It is clearly visible that the absolute value of the medians for efficacy decreases with increasing sample size. The sample size is much smaller for safety outcomes, such that it is not clear if the median effect sizes of safety outcomes do not decrease with increasing sample size, or the variation between medians is just too large to detect a decrease.
The pattern for efficacy outcomes is the same for all common outcome measures as shown in Figure 4.2, where the original effect size measures "log Odds Ratio", "log Risk Ratio", "Mean Difference" and "Std. Mean Difference" are used (the most common measures in the dataset, 98%).

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 of primary studies are (two-sided p-value < 0.05). 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 or if the data is only partially available (`outcome.flag` = `IV`).
Multiple tests are applied in order to compare their results. A histogram of p-values for each test will summarize the overall evidence against the null-hypothesis of no publication bias, as

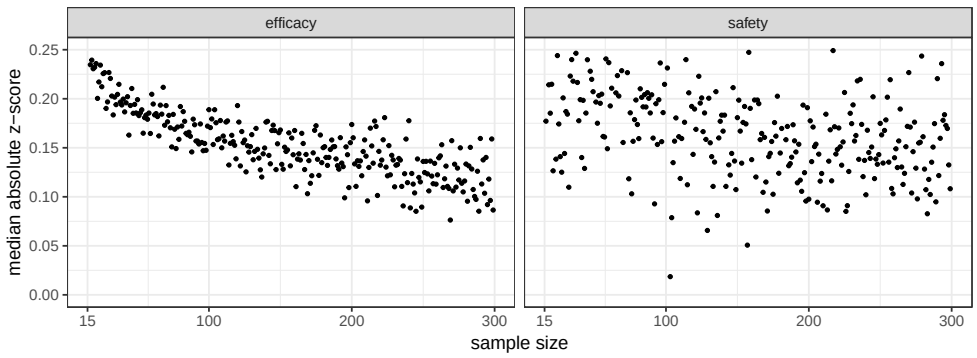

Figure 4.1: Median of the absolute z-score across sample size plotted against the total sample size.

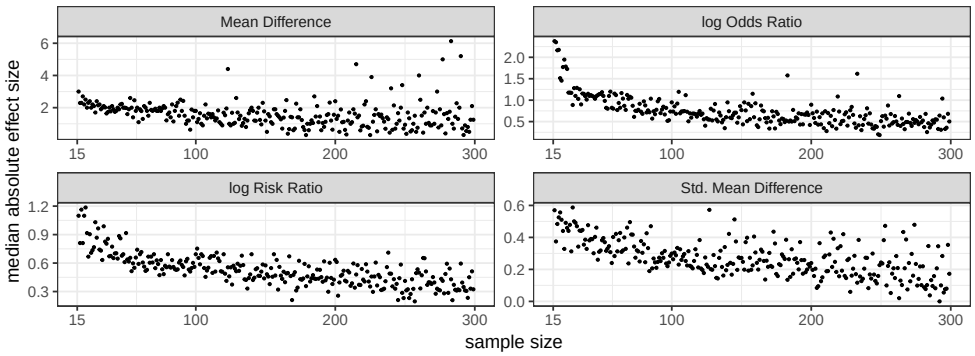

Figure 4.2: Median of the absolute value of the original effect size across sample size plotted against the total sample size.

displayed in Figure 4.3.
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 2.5.8. For continuous and `IV` outcomes, the names refer to:

- Egger's test, weighted linear regression test described in Section 2.5.2
- Thompson and Sharp's test, weighted linear regression test adjusted for between-study heterogeneity, Section 2.5.3
- Begg and Mazumdar's test, rank test described in Section 2.5.1

For binary outcomes, the names refer to:

- Harbord's test, likelihood score based test (Section 2.5.5)
- Peter's test, weighted linear regression with inverse sample size as explanatory variable described in Section 2.5.4
- Rücker's test, test based on the arcsine transformation of proportions, in combination with Thompson and Sharp's regression test (Section 2.5.7)
- Schwarzer's test, rank based test using the expected event counts computed with the hypergeometric distribution (Section 2.5.6)

The histograms in Figure 4.3 show that the tests mostly find evidence for publication bias in the dataset. The p-values of excess significance and Schwarzer's test are rather uniformly distributed, but notably, excess significance test, Schwarzer's test and rank tests have been shown to lack statistical power. The 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.
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). Other tests are specially constructed to adjust for between study heterogeneity (Thompson and Sharp's test and Rücker's test). These tests find a smaller proportion of significant results in Figure 4.3. However, this is also due to application to meta-analyses with no between-study heterogeneity, where the methods lack statistical 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 heterogeneity

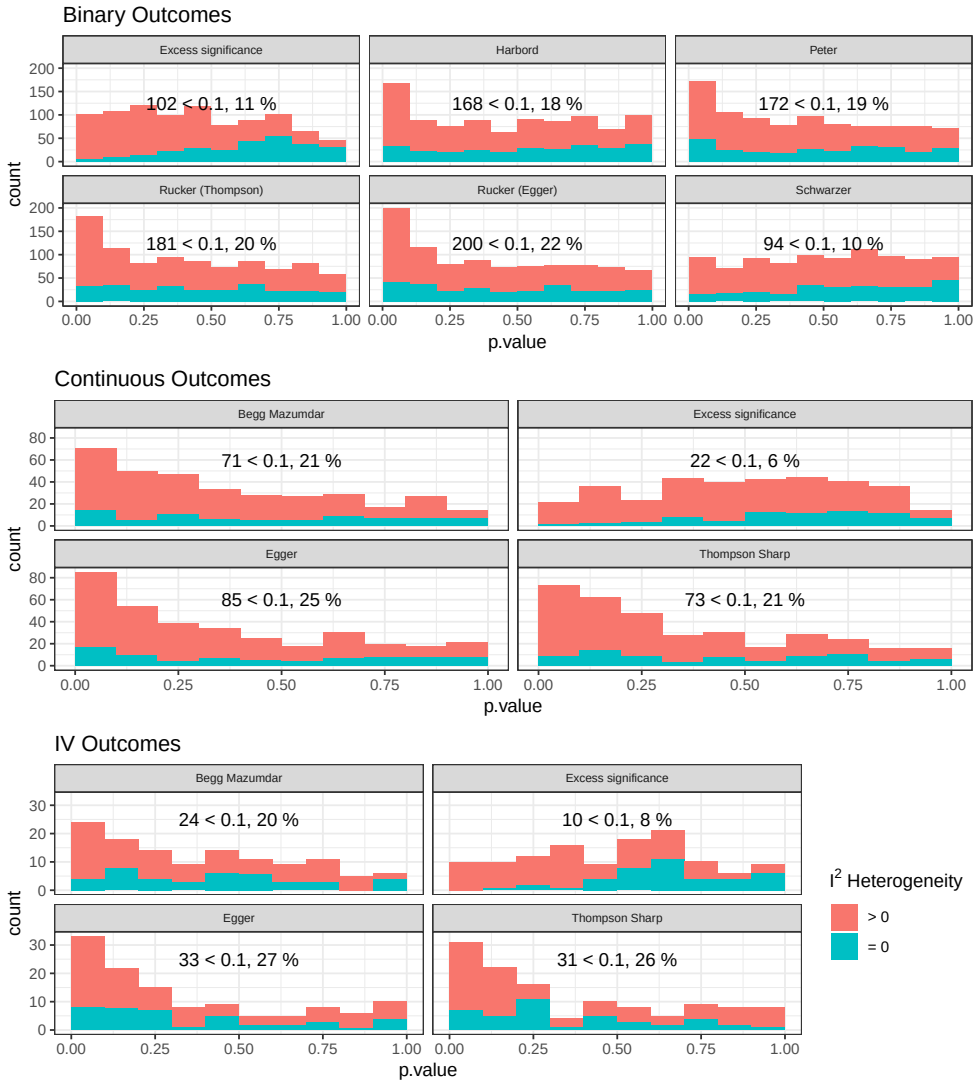

Figure 4.3: Histogram of one-sided p-values for small study effect in direction of larger effect sizes. The testing method is indicated in the header, bin width is equal to 0.1. The proportion of meta-analyses with significant publication bias based on the threshold of 0.1 is displayed inside the figures.

and that the test results are not heavily influenced by unaccounted between-study heterogeneity. The p-values of tests can be summarized by computing their harmonic mean (Good, 1958). 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 20.3% significant results ($p_{\text{harmonic}} < 0.1$, 19.7% `outcome.flag` = `DICH`, 20.1% `outcome.flag` = `CONT`, 25.6% `outcome.flag` = `IV`).

4.1.1 Publication Bias Test Consistency

In a next step, the consistency between the tests is examined, *e. g.* if the tests identify significant publication bias in the same meta-analyses. There is, so far, no agreement in the literature about which of the methods are to prefer, even non regression based tests have generally low power. The following analysis will reveal if there are *e. g.* testing methods that are coming to identical results and are thus compatible/interchangeable.

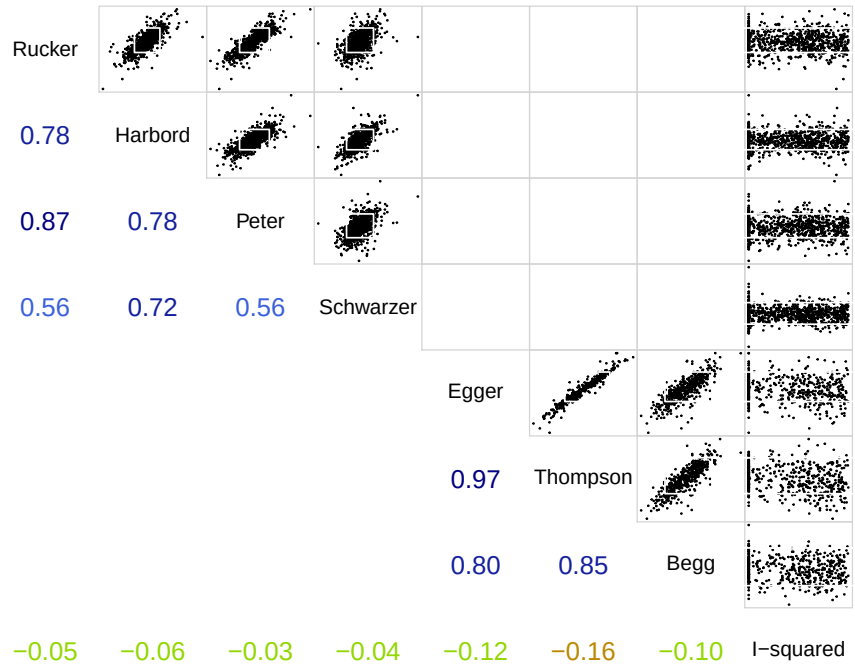

Figure 4.4: Pairs-plot for test statistics of small study effects. The lower panel gives the Spearman correlations for the different test statistics, and the upper panel displays a scatterplot. The colors indicate magnitude and direction of the correlation coefficients. The rectangle with white borders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

A simple method to check the consistency of test results is to compare scatterplots and Spearman correlations between the test statistics. This is done in Figure 4.4. Here, there is no separation between `IV` and continuous outcomes because the same publication bias tests have been used. 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 normally distributed test statistic under the null hypothesis is used for excess significance tests, it is not shown here.

The observed patterns on the scatterplots differ, and some methods 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.
However, the confidence intervals from Figure 4.5 are very large. 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. Very few meta-analyses are give a significant result, independently of the test applied. Around 67% of the dataset is not significant, no matter which test is used.

Count	Binary Outcomes	Continuous and IV
0	66.1 %	65.3 %
1	12.3 %	11.9 %
2	7.7 %	6.9 %
3	6.8 %	14.2 %
4	5.7 %	1.7 %
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

When leaving away the excess significance test, 29.5% of binary outcome tests and 31.9% of `IV` and continuous outcome tests had at least one significant result. To compare significant findings for small study effect tests and excess significance tests, Harbord's or Egger's test results are compared with excess significance tests. 24.1% of binary outcome analyses had at least one of the two test p-values being significant, and equivalently, 28.4% for continuous and `IV` outcomes. The numbers change to 13.9% for binary outcomes and 18.1% 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 outcomes, we have 3.9% with Egger's test and 1.3% with Bonferroni correction.
The precise proportions of agreement in significance/non-significance are provided in Table 4.2. A separate column provides the proportion of significant results of the test with fewer significant results that are also significant using the test with more significant results. 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`). The agreement on statistical significance between small study effect

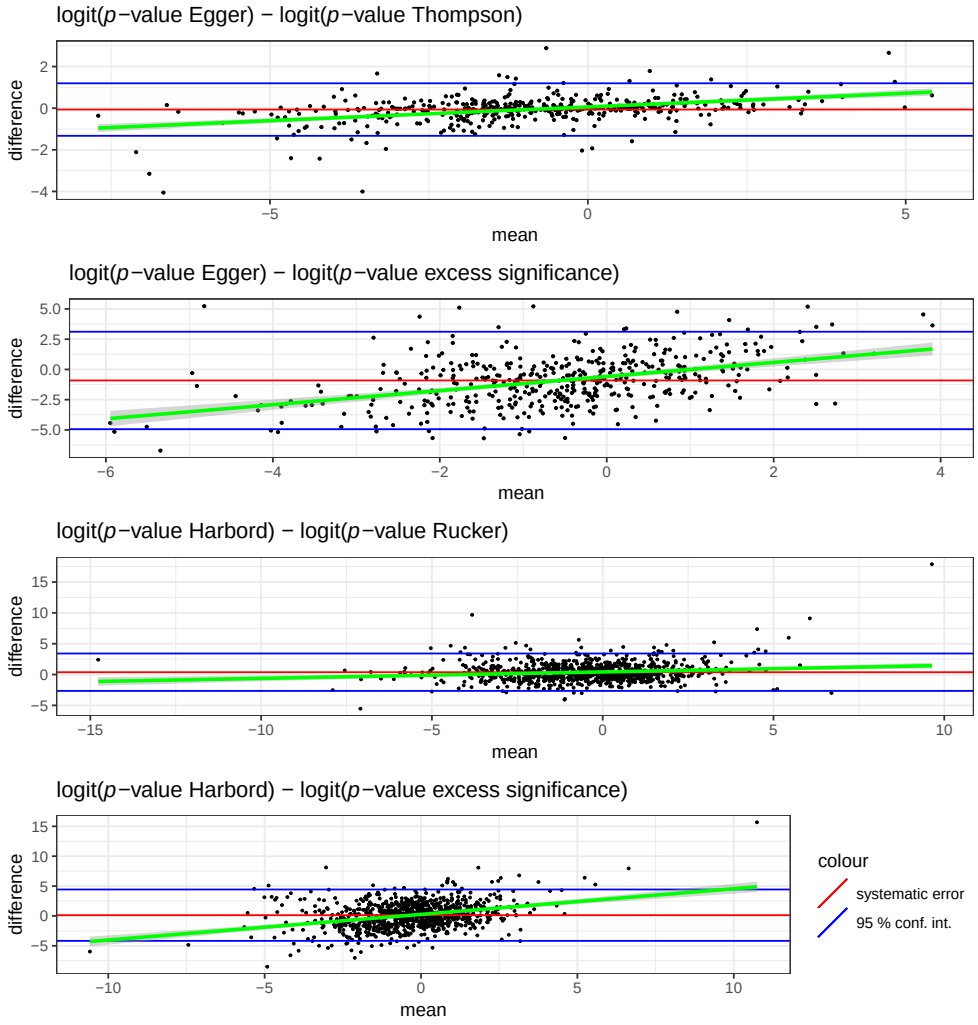

Figure 4.5: Mean - difference plots for logit transformed p-values. The mean of logit transformed p-values is displayed on the x-axis and the difference on the y-axis. Blue and red lines display the systematic error and the confidence intervals of the systematic error (limits of agreement). In green,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

tests and excess significance tests ranges from 64% to 4% (for `IV` and rank tests). Note that these numbers are difficult to interpret because there are substantially less significant results of the excess significance tests, and thus, a agreement of 100% still does not indicate perfect compatibility/replaceability of the tests.

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 2.7 for details). When comparing to unadjusted effects, fixed or random effects meta-analysis

	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)	0.71	0.12
Excess significance, Thompson (IV)	0.71	0.10
Excess significance, Begg (IV)	0.74	0.04
Thompson, Egger (IV)	0.95	0.94
Thompson, Begg (IV)	0.86	0.79
Egger, Begg (IV)	0.86	0.83

Table 4.2: Overall proportion of agreement if publication bias is significant or non-significant, and if significant only. When comparing agreement if significant only, the proportion of the test with fewer significant results that is significant with another test as well is shown.

estimates are used as references
Figure 4.6 displays the difference between the estimated meta-analysis treatment effect and the regression adjusted treatment effect 2.6.1, $\hat{\theta}_M - \hat{\theta}_{\text{Adj.}}$. The absolute value $|\hat{\theta}_M|$ is taken and $\hat{\theta}_{\text{Adj.}}$ is negative if it's sign is different from the sign of the original $\hat{\theta}_M$. 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 statistics of heterogeneity adjusted publication bias tests (Rücker's and Thompson's test) are displayed with green color. Test statitics smaller $t < 1$ (light green colored) are equivalent to no evidence for publication bias, test statistics t 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 the treatment effect. Some very large and very small differences have been omitted in the z-score and std. mean difference histograms; they are shown in Table 4.4.
Most often, adjustment leads to a reduction of overall treatment effect estimates, which can be seen by the size of the bins on the positive side of the histograms. 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 negative adjustment, but no evidence for small study effects. This is because the linear regression parameter estimates are large, but estimated with high uncertainty, such that there will be few evidence for small study effects (publication bias), but nonetheless, the adjustment will be large. It is recommended to use the methods only in cases where there is clear evidence for publication bias. The color legend in Figure 4.6 thus gives a sense of the confidence that is put into the adjusted effects.

Additionally, some meta-analyses with positively adjusted, larger overall treatment effects after adjustment (left side of the histogram) have also evidence for publication bias (defined as the tendency of small studies to show *larger* results because only significant results are published). However, this is not wrong, because 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 illustrative purposes. From the left to the right, the funnel plot is shown for mean

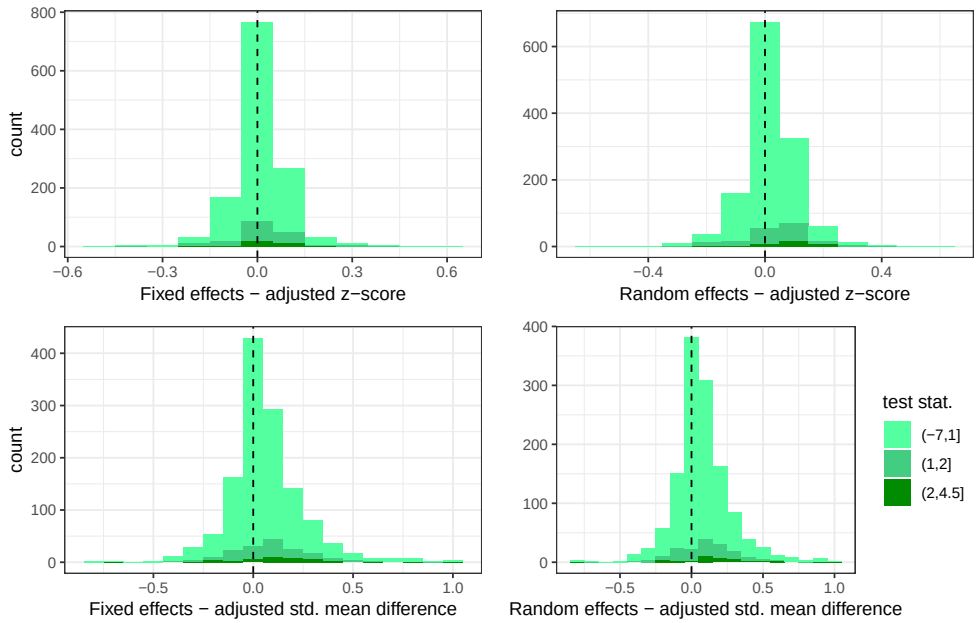

Figure 4.6: Histogram of the treatment effect differences between meta-analysis and regression adjusted meta-analysis. Negative differences indicate greater adjusted effect sizes than meta-analysis effect sizes. The bins are centered at zero and binwidth is equal to 0.1. Deeper green color indicates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

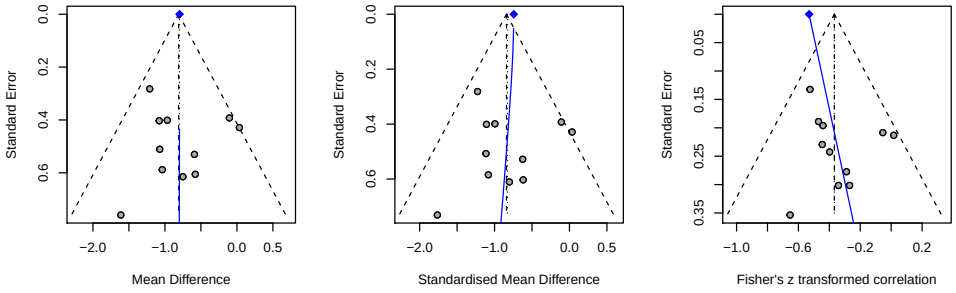

Figure 4.7: Funnel plots for a meta-analysis based on three different effect size measures: Mean differences, std. mean differences and Fisher's z transformed correlations and corresponding standard errors. Vertical dashed lines indicate meta-analysis estimates, the rhombus with the curved blue line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

differences (the original measure), standardized mean differences and Fisher's z scores; there is no change upon adjustment using mean differences (blue line), reduction of the effect 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).

Figure 4.8 shows the differences between Meta-analysis and adjusted effect sizes adjusted by Copas selection model; the 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

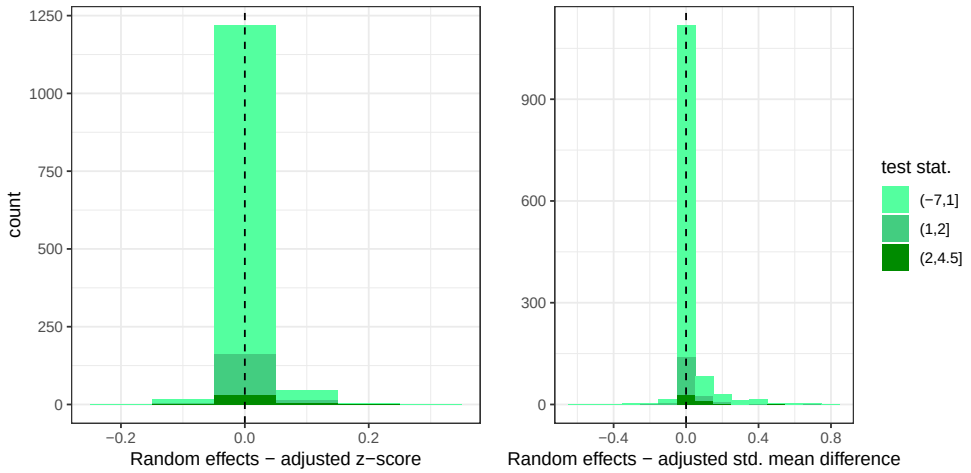

Figure 4.8: Histogram of the treatment effect differences between meta-analysis and Copas adjusted meta-analysis. Negative differences indicate greater adjusted effect sizes than meta-analysis effect sizes. The bins are centered at zero and binwidth is equal to 0.1. Deeper green color indicates more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

better 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 publication bias test statistics and adjustment, *i. e.* positive differences are accompained by large positive test statistics, 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 std. mean difference 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 69 meta-analyses have their z-score reduced by more than 0.13 by regression adjustment (and 5% or 69 increased by -0.11 or more, fixed effects reference). Also, std. mean difference is reduced by 0.39 compared to fixed effects estimates in 5% or 69 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	0.00	8.14	50.14	42.00	65.49
z: Random - Copas	-0.00	0.00	0.00	0.00	0.04	0.00	67.15	85.23	18.08	65.49
z: Fixed - Regression	-0.11	-0.03	0.01	0.05	0.13	0.01	0.00	52.02	52.02	0.00
z: Random - Regression	-0.13	-0.02	0.02	0.06	0.16	0.02	0.00	56.56	56.56	0.00
d: Fixed - Copas	-0.05	-0.01	0.00	0.01	0.11	0.01	19.38	56.41	37.03	56.92
d: Random - Copas	-0.00	0.00	0.00	0.00	0.16	0.02	60.09	85.01	24.93	56.92
d: Fixed - Regression	-0.21	-0.03	0.04	0.14	0.40	0.06	0.00	60.16	60.16	0.00
d: Random - Regression	-0.20	-0.02	0.05	0.17	0.44	0.08	0.00	62.68	62.68	0.00
IV: Fixed - Copas	-0.06	-0.01	0.00	0.01	0.07	0.00	23.97	42.15	66.12	59.50
IV: Random - Copas	-0.00	0.00	0.00	0.01	0.08	0.01	61.98	28.93	90.91	59.50
IV: Fixed - Regression	-0.24	-0.02	0.03	0.11	0.36	0.03	0.00	66.12	66.12	0.00
IV: Random - Regression	-0.25	-0.02	0.04	0.13	0.49	0.05	0.00	68.59	68.59	0.00

Table 4.3: Quantiles and means of the differences between meta-analysis combined 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 std. mean difference (d). Separate rows give the results for IV outcomes, where the original effect sizes (log rate ratios, hazard ratios, etc.) are used.

id	comparison.nr	subgroup.nr	z fixed	z random	z Copas	z reg.	smd fixed	smd random	smd Copas	smd reg.
CD000370	8	2	0.66	0.66	0.66	0.82	1.59	1.59	1.40	0.28
CD001183	7	0	-0.48	-0.48	-0.48	-0.21	-1.10	-1.12	-0.50	-0.11
CD002307	2	1	-0.10	-0.08	-0.10	0.03	-0.47	-0.42	-0.41	-3.00
CD008625	2	2	-0.60	-0.60	-0.60	-0.39	-1.72	-2.02	-1.01	-0.69
CD010060	1	0	0.31	0.32	0.31	0.23	0.50	0.52	0.20	-0.49

Table 4.4: Missing meta-analysis combined treatment effect and adjusted treatment effects. Abbreviations are used for z-score (= z) and std. mean difference (= d).

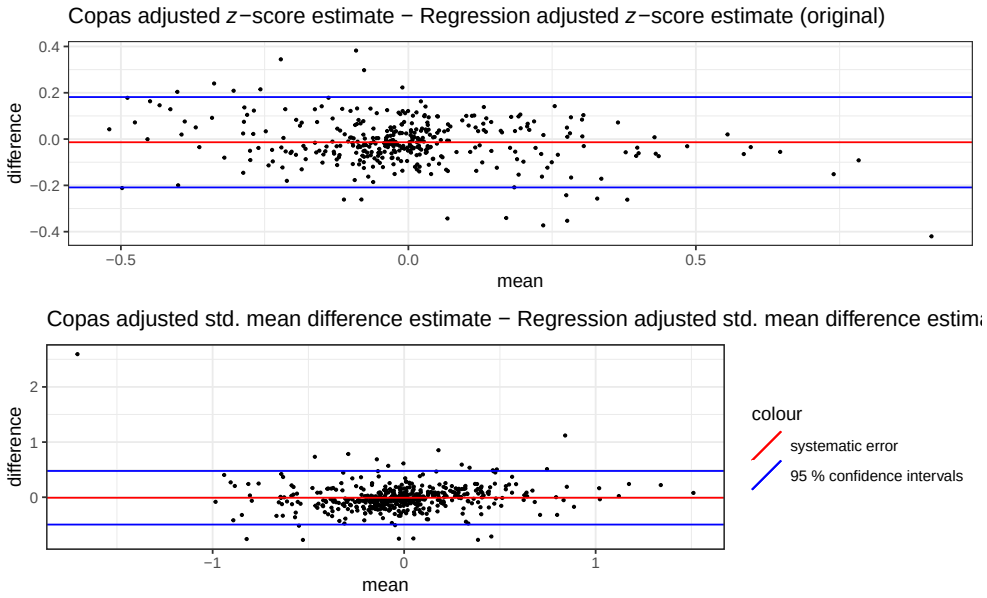

Figure 4.9: Mean - difference plots for publication bias adjustment methods. The mean of the adjusted treatment effects is displayed on the x-axis and the difference on the y-axis. Blue and red lines display the systematic error and the confidence intervals of the systematic error (limits of agreement). Two values have been omitted in the middle plot for std. mean difference and one for z-score (see Table 4.4).

4.2.2 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 in Figure 4.9. A Tukey mean difference plot can reveal systematic differences and biases between the two measurement methods. Note that only 26 out of 121 `IV` outcome data is included when using z-score and std. mean differences, since not all effects could be transformed.

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 4.9 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, indicating that regression estimates 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.

4.2.3 Change in Evidence for Treatment Effects

Adjustment for small study effects in meta-analysis will provide new effect sizes and standard errors. The evidence against the null hypothesis of no treatment effect can be computed newly.

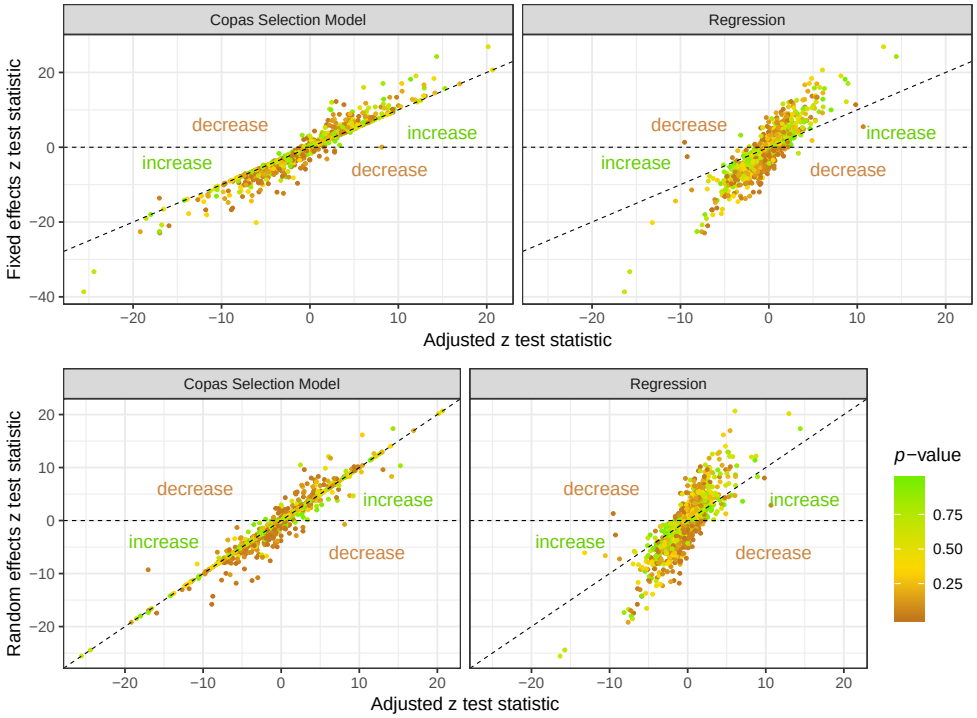

Figure 4.10: Test statistics of meta-analyses before and after adjustment. The color indicates the evidence for publication bias (p-value of Rücker's or Egger's test), and the dotted line depicts the diagonal through the origin. Together with the horizontal dotted lines, the area which it borders indicates the consequences of adjustment for the evidence for a treatment effect. "decrease" means that the test statistic is smaller upon adjustment, "increase" that it is larger.

The test statistics $\frac{\theta}{\text{se}(\theta)}$ of random effects and fixed effects meta-analysis and the corresponding adjusted test statistics are shown in Figure 4.10. The original scale of the effects sizes is used, since it is the measure on which policy-makers assess the treatment efficacy.

The areas that are bordered by the diagonal and horizontal lines indicate which are the consequences of adjustment for a dot within. "decrease" indicates fewer evidence for treatment effect is given after adjustment, "increase" that adjustment led to larger evidence for treatment effects. It can be seen that the alignment on the diagonal changes depending on the adjustment method. The effect of adjustment by copas can be seen when comparing it to random effects meta-analysis. There, less of an effect of adjustment on the evidence can be seen. The adjustment is more likely to be more reliable than the unadjusted estimate when publication bias is strong (see **??**). The dots with darker color are thus more likely to provide less biased estimates.

The decrease in evidence is larger when using adjustment with regression, which can be expected as the uncertainty of the additional parameter from the linear regression fit to estimate publication bias is included in the uncertainty of the estimate. In contrast, as the Copas selection model algorithm does a sensitivity analysis, the uncertainty of additional model parameters does not affect the estimate as they are not estimated but treated as fixed (see section 3.2.3 and section 2.6.2).

Large adjustment in effect sizes and test statistics is not necessarily accompanied by evidence for publication bias (color of the dots) in regression adjustment, as already discussed. But the dark dots can be found more often far away from the diagonal. Some cases which have more evidence for treatment efficacy after adjustment have evidence for publication bias, although the publication bias tests applied are one-sided and only test for bias towards large effects. This is because the algorithm to detect the side on which publication bias was expected (section 3.2.3) has failed,

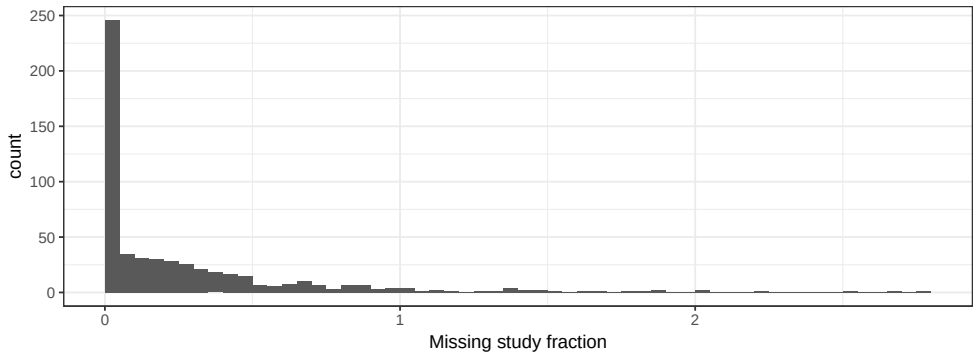

Figure 4.11: Histogram of the fraction of missing studies from the total number of studies in a meta-analyses (only data shown where Copas estimate was obtained, thus $n = 560$)

e. g. because few significant effects are given (used to define side of bias) and the weighted mean effect is close to zero, in which case it is difficult to decide upon the side of publication bias. There are some adjustments with large increase in evidence and clear dark orange color, as in the plots on the right hand side in Figure 4.10. Two of them are investigated in detail.

- z-value of regression adjustment 10.5 vs fixed effects z-value 5.7: When the funnel plot of these meta-analyses are investigated, it becomes clear that there is no reason to assume publication bias since the smaller studies are near to risk ratios equal to one, and larger studies show larger ratios. Publication bias tests disagree in their findings (Rücker's and Harbord's tests: 0.058 and 0.05, and Peters and Schwarzer's test: 0.467 and 0.79).

- z-value of regression adjustment 9.24 vs fixed effects z-value 2.51: Asymmetry of the funnel plot could be confirmed when analyzing it with mean differences. The change to a larger treatment effect is rather due to large between study heterogeneity and the specific routine of the regression adjustment. Because it takes into account between-study heterogeneity (bias parameter times τ^2), while fixed effects meta-analysis does not, it returns a larger treatment effect. The adjusted effect (MD = - 10.3) lies between the fixed effect estimate (-2.56) and the random effects estimate (-18.7).

In the case of Copas selection model and when both test statistics are large (> 10), the explanation lies in a specialty of the model; the adjusted treatment effect estimates are not larger than the unadjusted, but their standard errors are different and sometimes smaller, because they are obtained from the Fisher information matrix of the log-likelihood (see subsection 2.6.2).
The Copas selection model also allows to compute the number of missing (*i. e.* unpublished) studies in a given meta-analysis, given that the model's assumptions are correct. It finds that 2,618 are missing, which corresponds to 11.4% from all 22,916 analysed studies. Figure 4.11 shows a histogram of the overall fraction of missing studies. Note that random effects meta-analysis substitutes have been excluded (828 out of 1,388 to limit the size of the bin at no. missing studies = 0.

We can see that in some occasions, the method finds more than half of all studies in a meta-analysis are missing. In most occasions, the estimate of missing studies is zero, as can be seen in Table 4.5, where both the absolute number of missing studies and the fraction of missing studies in a meta-analysis are given. The discrepancy between mean and median may indicate that the estimate of 11.4% missing studies depends somewhat on these extreme cases. As can be written of from Table 4.5, 5%, *i. e.* 28 meta-analyses have 17.6 or more studies missing: in fact, these 5% most extreme make up for 1,060, more than 30% of all missing studies.

	= 0	5%	25%	50%	75%	95%	mean
Missing fraction	226	0	0	0.1	0.4	1.0	0.3
Missing study number	226	0	0	1.5	5.9	20.1	4.7

Table 4.5: Fraction of missing studies and estimates of missing studies with their zero counts ("= 0"), quantiles and means.

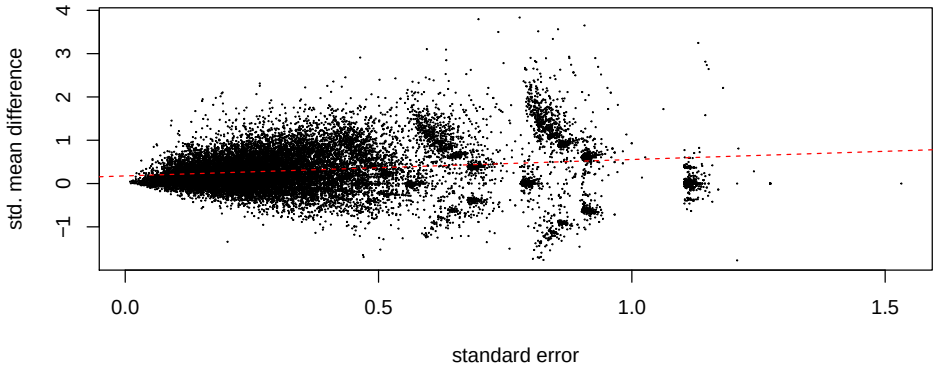

Figure 4.12: Std. mean differences of all meta-analyses plotted against their standard error, with the fit of a generalized linear model (dotted red line).

4.3 Mixed Effect Models and Publication Bias over Time

To test if publication bias in meta-analyses has decreased over time and newer meta-analyses have less bias, a generalized linear model has been set up to analyze all meta-analyses jointly and estimate a single parameter denoting publication bias.
The dependent variable is the standardized mean differences(therefore, from `outcome.flag` = `IV`, only the std. mean differences are used). As before in section 4.2.1, the std. mean difference is mirrored to one side by multiplying with the sign of the expected side of bias, *i. e.*-1 or 1.
In a first step, the explanatory variable is the standard error, and the weights are the inverse of the variance of the std. mean difference. Additionally, random effects for the meta-analysis and for the review are added, the former being nested in the latter. Figure 4.12 shows the complete dataset with the fit of the previously described model (dotted red line). The details and evidence for the small study effects in the generalized linear model are given in Table 4.6, the coefficients in Table 4.7

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
null.fit	1	4.0	18768.2	18800.1	-9380.1			
publication.bias.fit	2	5.0	18230.5	18270.4	-9110.3	1 vs 2	539.7	<0.00001

Table 4.6: Anova table for a generalized linear model for small study effects compared to the null model ('null.fit'). 'publication.bias.fit' denotes the model with random intercepts for meta-analyses and reviews.

It is possible to include meta-analysis specific random slopes. Before continuing, it is tested if incorporation of random slopes improves model fit. Table 4.8 shows the anova table for a model with and without random slopes. Model diagnostics indicate that there is no benefit in including random slopes for the single meta-analysis.
To test if publication bias varies over time, we include the year of the publication of the review as an additional explanatory variable. Ultimately, it is of interest if there is an interaction between

	estimate	2.5%CL	97.5%CL
(Intercept)	0.18	0.16	0.19
se.smd.pool	0.38	0.35	0.41

Table 4.7: Coefficients and 95% confidence limits of the generalized linear model. 'se.smd.pool' denotes the standard error of the std. mean difference.

the small study effect (publication bias) and time of publication, i.e. if the slope varies depending on the year. Because such a model is nested in a more simpler model where the study year is only an additive effect, all these models are fitted. Table 4.10 displays model fit diagnostics.

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
publication.bias.fit	1	5.0	18230.5	18270.4	-9110.3			
publication.bias.rs.fit	2	7.0	18233.1	18288.9	-9109.5	1 vs 2	1.5	0.5

Table 4.8: Anova table for two generalized linear model fits. 'publication.bias.fit' denotes the model with random intercepts for meta-analyses and reviews, 'publication.bias.rs.fit' the model with additional random slopes per review.

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
publication.bias.fit	1	5.0	18230.5	18270.4	-9110.3			
review.year.fit	2	6.0	18231.6	18279.5	-9109.8	1 vs 2	0.9	0.34
review.year.int.fit	3	7.0	18233.6	18289.4	-9109.8	2 vs 3	0.0	0.87

Table 4.9: Anova table for three generalized linear model fits. 'publication.bias.fit' denotes the model with the standard error as explanatory variable, 'review.year.fit' the model with standard error and the centered (- 2013) review year as explanatory variable, and 'review.year.int.fit' the model with interaction between the two.

Because neither AIC, BIC nor the F-test do show any improvement of model fit, we can not reject the null hypothesis that publication bias has not decreased or increased over the years. The review publication year is however only one measure of time, and possibly imprecise. Thus, the models are re-fitted with the mean centered study publication year of a meta-analysis. The mean of the study years is 2000.3.

The analysis of model fit indicates that there is an improvement when using the study year as a covariate, but only as an additive effect. It finds that effect sizes are in general little bit smaller in latter mean publication years (after the year 2000). The coefficients of the model are given in Table 4.11. AIC and BIC are smaller and the p-value of the F-test is very small. However, not the size of the effect, but the slope of the linear regression fit would have to be decreased in latter years to ultimately conclude a decrease of publication bias over the years.

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
publication.bias.fit	1	5.0	18115.2	18155.1	-9052.6			
study.year.fit	2	6.0	18117.2	18165.0	-9052.6	1 vs 2	0.0	0.958789
study.year.int.fit	3	7.0	18118.9	18174.7	-9052.5	2 vs 3	0.3	0.591536

Table 4.10: Anova table for three generalized linear model fits. 'Small study fit' denotes the model with the standard error as explanatory variable, 'study.year.fit' the model with standard error and the centered (- 2000) mean study year of the meta-analysis as explanatory variable, and 'study.year.int.fit' the model with interaction between the two.

	estimate	2.5%CL	97.5%CL
(Intercept)	0.1758	0.1575	0.1940
se.smd.pool	0.3800	0.3482	0.4118
mean.study.year	-0.0001	-0.0020	0.0019

Table 4.11: Coefficients and 95% confidence limits of the generalized linear model. 'se.smd.pool' denotes the standard error of the std. mean difference.

Chapter 5

Discussion

5.1 Results in the light of the Literature

Before discussing the results of the previous analyses, some results of similar studies are discussed briefly and compared to the results of this study, if possible. When clear methodological drawbacks are found in the analysis, the reader is referred to chapter 2 for comments on the disadvantages. The amount of studies analyzing publication bias makes it not possible to discuss each of them, and thus only some interesting and/or representative examples are shown.

5.1.1 Egger *et al.* (1997)

The Cochrane Library (1996 issue 2) and publications from the four leading medicine journals (the Lancet, BMJ, JAMA and Annals of Internal Medicine) between 1993 and 1996 were used to find systematic reviews with randomized controlled trials for the application of Egger's small study effect test. They included 38 meta-analyses with at least 5 studies and binary outcomes from the Cochrane Library and 37 from the journals. Five (13%) meta-analyses from Cochrane and 13 (38%) from the journals had significant two sided small study effect test results (p-value < 0.1). Also, they found that test-statistics were more often negative, which corresponded in their setup to larger effects in small studies. (63.2 among Cochrane meta-analyses and 70.3 for journal meta-analyses). The results from this report with Rücker's test are: 14.8%, and when using regression adjustment to decide about the direction of publication bias, we get 80.7%.

5.1.2 (Leicester)

Out of 397 systematic reviews from the Cochrane Library (complete number for 1998, issue 3), 49 had more than ten included studies and binary outcomes and 48 compared two treatments. They were analysed by trim-and-fill method (Duval and Tweedie, 2000) to detect and adjust for funnel plot asymmetry (similar to small study effect tests, but the method is known to overestimate bias). 23 were found to have missing studies, and eight had more than three missing studies which was considered to be significant publication bias. Additionally, they found that three estimates of random effects meta-analysis became non-significant after adjustment, and one became significant (by negative adjustment). The results are difficult to compare, but the methodological limitations make this findings unreliable.

5.1.3 Ioannidis and Trikalinos (2007a)

Data processing steps were however different, and only binary outcomes as used in two-by-two tables were analysed. The approximately corresponding numbers are put in parentheses for comparison. The Cochrane Library from 2003 (issue 2) was used. After removal of duplicates and intractable meta-analyses, they had 6,873 meta-analyses with more than two studies left

(20,219). When only using one meta-analysis per review, this reduced to 846 (3,555. Then, the criteria that have also been applied in this masters study are applied: $I^2 < 0.5$, variance ratio of smallest and largest effects > 4, at least one significant study result and at least 10 studies to be used. Afterwards, they applied Harbord's, Egger's and Begg's Test to the dataset. The reader can compare some corresponding numbers in Table 5.1.

	Ioannidis	Study
Wider dataset (n)	6873	20219
study number > 10	13%	12%
variance ratio > 4	72%	75%
study sig. number > 1	55%	52%
all exclusion criteria	5%	5%
harbord test p-value < 0.1	12%	16%

Table 5.1: Comparison of results from Ioannidis et.al. (2007) to the results of this study. The percentage of meta-analysis which match all exclusion criteria denotes the ones that apply to all criteria in the table plus the small heterogeneity criterium. Harbord test is two-sided.

5.1.4 Souza *et al.* (2007)

Reviews of the World Health Organization (WHO) Reproductive Health Library (RHL), issue 9, were analysed with the trim-and-fill method. The RHL reproduces and expands reviews from the Cochrane Library with implications for developing countries. 21 of 105 reviews contained more than ten studies and were used. Trim-and-fill found asymmetry in 18 of 21 studies, and 10 had more than 3 missing studies ("significance"). Two of those and one with one missing studies found no evidence for treatment effects after the 0.05 p-value threshold after adjustment by trim-and-fill.

5.1.5 Kicinski *et al.* (2015)

The author uses a Bayesian hierarchical selection model, but does not analyse treatment effects, but the parameters of the weight function of the selection model, which is estimated with a Bayesian approach and MCMC sampling.
The author provides an estimate of the probability of including significant findings versus non-significant findings in Cochrane meta-analyses over time. The data is from the Cochrane Library from 2013 (issue number not provided). The author excluded treatment - treatment comparisons and analysed safety and efficacy meta-analyses separately (how this was achieved is not documented in the paper).
From 3845 reviews, the author separated 907 reviews with more than ten studies. From those, 539 compared placebo to treatment. After removing duplicates and sensitivity analyses, 358 analyses with 1297 meta-analyses remained. From these 191 were excluded because they comprised overall mortality and withdrawal, because these could not clearly be specified as safety or efficacy, respectively.
1106 meta-analyses from 329 meta-analyses, containing 802 efficacy and 304 safety meta-analyses. The median publication year per meta-analysis was 1997 for efficacy meta-analyses and 1999 for safety meta-analyses. Then, a Bayesian two-step hierarchical selection model was applied (Kicinski (2013), Silliman (1997)). It assumes that the selection process is a two-step weight, which assigns different probabilities to non-significant and significant effects. Thus, a ratio of publication probability between significant and non-significant study estimates using a two-step weight function. The model was fitted with the Monte-Carlo Markov-Chain algorithm STAN and the

geometric mean was used as an estimate for the publication probability ratio. Simulations performed in (Kicinski, 2013) indicate that the method performs well if the true mean effect size was not small, and also robust to small study effects. It outperformed Egger's test and Begg's test in assessing publication bias, especially when small study effects were absent, and had lower false-positive rates.
The results showed a clear publication bias for significant results for efficacy meta-analyses (27% higher for significant studies, 95%CI credible intervals: 1.18 to 1.36). The probability was more than twice as high in 27% of the meta-analyses (95% CI: 23% to 31%). But the probability decreased: from 1.65 (95% CI: 1.31 to 2.15) in 1980 (average publication year) to 1.36 (95% CI: 1.17 to 1.62) in 1990 to 1.18 (95% CI: 1.04 to 1.33) in 2000. For sake of completeness, the probability of inclusion was 1.78 (95% CI: 1.51 to 2.13) larger for non-significant safety effect estimates, but again, decreased with time (1.77, 95% CI: 1.46 to 2.21 in 2000). The results are in line with the results from this study, but it was not possible to reaffirm the finding of weaker publication bias in reviews published more frequently.

5.1.6 van Aert *et al.* (2019)

The Authors of this recent study sample 366 meta-analyses randomly from the Cochrane Library (supposedly 2018 or 2019, not mentioned). They exclude any meta-analysis which include effect sizes identical to effect sizes in other meta-analyses (the larger meta-analysis is retained). Additional criteria were: $I^2 < 0.5$ and at least 5 studies. The meta-analyses were analysed using standardized mean differences and four different tests for publication bias: Egger's test, p-uniform test (Van Assen *et al.*, 2015), Begg and Mazumdar's rank correlation test and the excess significance test.
The authors found, based on the significance threshold of $p < 0.1$, the following results (own results in brackets): 12.2 % significant results from Egger's test (16.3), 8.5 % significant results from Begg and Mazumdar's rank correlation test (10.5) and 4.4 % significant excess significance test results (9.7). It is known that the publication bias tests are lacking power in general as sample size is usually small. Decreasing sample size will result in lower power, especially if the maximal sample size is $n \geq 5$. Additionally, it was also not taken into consideration that it may be difficult to state that publication bias is present in a meta-analysis when no result within it is statistically significant (only 18.8 % of the effects of the sampled meta-analyses were). This, together that it has not been tried to exclude safety outcomes, may account for the large differences between the results.
The author's come to the conclusion in their study that in contrast to other studies, they find few evidence for publication

5.1.7 Further Studies on Publication Bias

Zhang *et al.* (2013) find publication bias in critical care studies with similar methods. Nüesch *et al.* (2010) report publication bias in clinical osteoarthritis research because of funnel plot asymmetry and pledge for routine assessment of publication bias. Dechartres *et al.* (2013) analyse publication bias based on sample size in meta-analyses from top journals and Cochrane reviews in 93 Meta-analyses and find that, for example, effects in trials with less than 50 patients were 48% larger than in larger trials. Onishi and Furukawa (2014) find that in 36 Meta-Analyses without comprehensive literature research, there are 19.4% significant publication bias tests (Egger's test).

5.2 Interpretation

In this study, evidence for publication bias has been found in approximately 20% of the meta-analyses. The results are often times consistent among different tests and subsets of the data. A mixed effects meta regression confirms these findings with very large evidence for small study effects (indicating publication bias). Since the results of two-sided tests do not differ substantially from previous findings, the results are in line with previous research.
Presence of publication bias in meta-analyses can ultimately lead to patient harm if decision makers in clinical practice use the meta-analyses to decide about application of a treatment. On the swiss Cochrane website, it is stated that Cochrane is the "single most reliable source of evidence on the effects of health care". However, publication bias does in general lead to too large confidence in the results of the meta-analyses. In some cases, it might be that the evidence for a treatment in a meta-analysis is an artifact of publication bias. This also leads to a waste of resources in science.
The term publication bias is frequently used in this study, but 1.7% of the analysis dataset (1) are unpublished results. This mitigates the estimated extent of publication bias in this study since it has been shown repeatedly that effects in the grey literature are smaller.

5.3 Limitations

After application of the criteria of Ioannidis and Trikalinos (2007a), there are left 11.6 % of the reviews and 32.9% of the studies. It is not clear if the results also apply for the excluded data, *i. e.* if the extent of publication bias is similar. The most restrictive criterium of exclusion was that a meta-analysis had to have at least ten studies. The included studies are thus likely to be part of enforced and sustained research efforts. When fewer and unparalleled studies are available, the quality of the research is not necessarily better. Publication bias may be even more of an issue in more discovery oriented research Ioannidis (2005). By excluding meta-analyses where $I^2 \geq 0.5$, the most extreme cases of publication bias are possibly removed since publication bias increases the between-study heterogeneity.
It also has to be said that the removal of safety outcomes is only partially successful. This is an issue, because it is not well known if and how the publication probability is affected by the size of safety outcomes. Kicinski *et al.* (2015) suggests that they may inversely affect publication (*i. e.* non-significant effects are preferred).
Also, it has been previously discussed that the method to detect the side on which bias is to be expected can fail and provide meaningless results. One could also discuss in more general terms if one can speak of publication bias in meta-analyses that have only one significant effect out of ten ($n = 417$), *i. e.* if statistical significance is largely absent in a meta-analysis.
There is also criticism that one can address on the methods. The small study effect which is used as a indication for publication bias can be caused by other biases such as; selective outcome reporting and poor methodological quality of smaller studies. It can also be caused by true heterogeneity between studies; early and small studies are more likely to include high-risk patients for which treatment can have larger benefits.

It is possible that the results are caused by systematic differences in the population of study participants only. This is considered rather unlikely. Other criticism is more severe. The small study effect tests applied do not account for statistical significance of effects directly. They work because smaller studies with larger errors need larger effects to be significant. The only test that is applied that takes into account statistical significance per se is the excess significance test, which is underpowered.
The criticism also applies for adjustment methods, as they also rely on small study effects. A weakness of the regression adjustment is that it also adjusts if the uncertainty in the small study

effect is large. To compare it in this case with the meta-analysis estimate is difficult because the uncertainty in estimating the small study effect affects the uncertainty of the treatment effect. Copas selection model uses the p-value threshold of 0.1 to decide if adjustment is necessary or not. One could argue that this threshold is arbitrary. Additionally, the threshold is used in two manners: first to decide if adjustment is necessary at all, and secondly, if the adjusted model does suitably account for publication bias ($p > 0.1$). Hypothesis tests against the null hypothesis can however not be used to decide in favor of the null-hypothesis, and a separate test would be needed there.
With Copas selection model, it is in general the issue that the model is not estimating the selection process parameters, but rather applying some criteria of parsimony (the model with least publication bias/selection with p-value for small study effect > 0.1 is chosen).
Publication bias can also be present if a meta-analysis lacks a small study effect, for example if by chance, the smaller studies were published irrespectively of the statistical significance of their findinfs. Methods as the excess significance test of Ioannidis and Trikalinos (2007b) are applied to overcome this issue, but the method is known to lack power. No methods exist so far to adjust for excess significance.
Therefore, the results are not exact and could be improved if more suitable methods are applied. The Bayesian approach taken by Kicinski *et al.* (2015) has many advantages over the applied methods, *e. g.* by accounting for statistical significance, but relies on assumptions over the weight function and is at least somewhat sensible on the choice of the prior distributions.

5.4 Outlook

Given the size and abundance of information in the dataset, the possibilities to investigate sources of bias in the Cochrane Library of Systematic Reviews are not exhausted (see for example in Fanelli *et al.* (2017)). In particular, one could easily come up with more hypotheses that could be tested in meta meta-regression, as introduced in the last section of the results. Many researchers have come up with hypotheses about the reasons and circumstances that promote publication bias, which could be tested if the information is extended.
There are numerous suggestions for different measures of publication bias and small study effects, which could be applied on the dataset (see Mueller *et al.* (2016)). This study mainly relies on suggestions from Sterne *et al.* (2001b), Ioannidis and Trikalinos (2007a) and Rücker *et al.* (2011). It uses most often methods that are well integrated in packages (Viechtbauer (2010), Schwarzer (2007)). Although there are well addressed, mathematical justifications for these methods, there might be better methods not yet tested on large datasets. An evaluation of all suggested methods was unfortunately beyond the scope of this study.

5.5 Implications

This study is so far the largest assessment of publication bias by small study effects and uses data that has been collected after major efforts have been made to curb publication bias. It is the first analysis of the Cochrane Library of Systematic Reviews also analyzing publication bias using hazard and rate ratios. The results indicate, similarly to other studies, that the efforts did not yet resolve the issue.
Also, it is the first study to extensively adjust the combined effect sizes from meta-analyses for publication bias. It has been found repeatedly that publication bias might threaten the validity of the findings of some meta-analyses from Cochrane systematic reviews. Cochrane did so far not extend their protocols to publication bias tests or adjustment methods. Especially when a lot of studies are available, the use of adjustment methods might resolve this issue.
Since Cochrane uses results from grey literature to limit the effects of publication bias, but still

does not manage to abolish it, there is also the need for journal editors to change their publication policies.

Bibliography

Abbasi, K. (2004). Compulsory registration of clinical trials. *BMJ*, **329**, 637–638. 1

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. 11

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, 17

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, **10**, 101–129. 1

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

Dechartres, A., Trinquart, L., Boutron, I., and Ravaud, P. (2013). Influence of trial sample size on treatment effect estimates: meta-epidemiological study. *BMJ (Clinical research ed.)*, **346**, f2304–f2304. 47

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. 11, 45

Dwan, K., Gamble, C., Williamson, P. R., and Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias — an updated review. *PLOS ONE*, **8**, 1–37. 1

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. and Smith, G. D. (1995). Misleading meta-analysis. *BMJ*, **310**, 752–754. 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, 11, 45

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2007). *Regression*. Springer. 8

Fanelli, D., Costas, R., and Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, **114**, 3714–3719. 49

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. 11

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, **53**, 799–813. 31

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. 12

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. 17

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, 19

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, 25, 45, 48, 49

Ioannidis, J. P. and Trikalinos, T. A. (2007b). An exploratory test for an excess of significant findings. *Clinical Trials*, **4**, 245–253. 13, 30, 49

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, **2**, . 48

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

Kicinski, M. (2013). Correction: Publication bias in recent meta-analyses. *PLOS ONE*, **8**, . 46, 47

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, 46, 48, 49

Lee, K., Bacchetti, P., and Sim, I. (2008). Publication of clinical trials supporting successful new drug applications: a literature analysis. *PLoS Med*, **5**, e191. 1

(Leicester), A. J. S., Song, F., Gilbody, S. M., and Abrams, K. R. (2000). Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research*, **9**, 421–445. 45

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

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. 16

Moreno, S. G., Sutton, A. J., Ades, A., Stanley, T. D., Abrams, K. R., Peters, J. L., and Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, **9**, 2. 11

Mueller, K. F., Meerpohl, J. J., Briel, M., Antes, G., von Elm, E., Lang, B., Motschall, E., Schwarzer, G., and Bassler, D. (2016). Methods for detecting, quantifying, and adjusting for dissemination bias in meta-analysis are described. *Journal of Clinical Epidemiology*, **80**, 25 – 33. 49

US Public Law (2007). *United States Code 110–85*. Food and Drug Administration Amendments Act. 1

Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A. W. S., Tschannen, B., Altman, D. G., Egger, M., and Jüni, P. (2010). Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ*, **341**, . 47

Onishi, A. and Furukawa, T. A. (2014). Publication bias is underreported in systematic reviews published in high-impact-factor journals: metaepidemiologic study. *Journal of Clinical Epidemiology*, **67**, 1320 – 1326. 47

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. 27

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. 14, 16, 27, 49

Rücker, G., Schwarzer, G., Carpenter, J. R., Binder, H., and Schumacher, M. (2010). Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics*, **12**, 122–142. 14

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News*, **7**, 40–45. 16, 27, 49

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. 12

Silliman, N. P. (1997). Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association*, **92**, 926–936. 46

Souza, J. P., Pileggi, C., and Cecatti, J. (2007). Assessment of funnel plot asymmetry and publication bias in reproductive health meta-analyses: an analytic survey. *Reproductive Health*, **4**, 3. 46

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

Sterne, J. A. C., Egger, M., and Smith, G. D. (2001b). Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, **323**, 101–105. 49

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, **18**, 2693–2708. 12

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, 27, 47

Van Assen, M. A., van Aert, R., and Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological methods*, **20**, 293. 47

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. 27, 49

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. 27

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1. 27

Zhang, Z., Xu, X., and Ni, H. (2013). Small studies may overestimate the effect sizes in critical care meta-analyses: a meta-epidemiological study. *Critical care (London, England)*, **17**, R2–R2. 47