

**First line of title**  
**second line of title**

---

Master Thesis in Biostatistics (STA495)

by

Name of student  
Matriculation number

supervised by

Name of responsible supervisor (with title)  
Name of supervisor (with title and affiliation if external)

Zurich, month year



*p*-values:  
their use, abuse and proper use  
illustrated with seven facets

Mäxli Musterli

Version April 13, 2019



# Contents

<b>Preface</b>	<b>iii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 The Cochrane Dataset</b>	<b>3</b>
<b>3 Results</b>	<b>9</b>
3.1 Small Study Effects . . . . .	9
<b>4 Methods</b>	<b>15</b>
4.1 Basic notation . . . . .	15
4.2 Heterogeneity . . . . .	15
4.3 Meta Analysis . . . . .	16
4.4 Small Study Effects Tests . . . . .	17
4.5 Small Study Effect Adjustment . . . . .	19
<b>5 Conclusions</b>	<b>23</b>
<b>A Appendix</b>	<b>25</b>
<b>Bibliography</b>	<b>27</b>



# Preface

Howdy!

Max Muster  
June 2018





# Chapter 1

## Introduction

Meta-analysis is at the core of evidence based medicine because it allows to summarise evidence over multiple studies and provide a more broad view on success and effectivity of clinical treatments. The necessity of meta-analyses is also increased by the abundance of data and publications. Especially when the findings differ or even contradict between studies, meta-analysis is the only way to go if one wants to make decisions based on quantitative and scientific criteria.

For this, meta-analyses do not only benefit research, but also clinical practice, and may lead to better health care and prevention. However, the usefulness of meta-analysis does not restrict to clinical science, but to any empirical and quantitative science.

Usually, a meta-analysis is part of a systematic review where researchers decided to summarise all research in a given field or more specifically, that concerns a given question. Meta-analysis can be applied to all studies that are approximately identical in their experimental setup and the way the outcome of the experiments is measured. In systematic reviews where meta-analyses are used, the conclusions are most often strongly based on the results and the interpretation of the meta-analysis.

However, there are problems that potentially limit the validity of meta-analysis; the number of studies available can be incomplete or the results of the studies can be biased. Some of those problems can be solved or asserted by special statistical methods.

### 1.0.1 Small Study Effects or Publication Bias

When study sample size decreases, the probability of extreme and misleading results in a study increases. This becomes a problem if results are selectively published, and therefore available, based on their results. When this is the case, one speaks of a small study effect or of “Publication bias”.

The issue has been discussed extensively in the last years, most often in the context of what has come to be known as the replication crisis. The reasons for small study effects are manifold, but originate most often in the myopical acting of agents in science and the lack of statistical education. Studies are reported by scientists, published by journals and noticed by readers more often if their findings are positive and find e.g. a substantive difference or effect. When doing a meta-analysis, one again obtains biased results.

The reason why that is less of an issue for larger studies is that extreme results are in general less likely and that due to larger effort, a result is published although there has been no clear and positive findings.

While there is generally no way to assert poor study quality, small study effect can in principle be asserted and corrected for statistically. This masters thesis will mainly be about statistical methods to detect and adjust for small study effects. It can furthermore be divided in two parts:

- Methodological part: Collection and discussion of statistical tests and correction methods for small study effects.

- Applied part: Application of the methods to studies of the Cochrane Library of systematic Reviews. Subsequent discussion of the implications of the results for clinical science.

In contrast to simulation studies, it is not possible to assess critical properties of the methods such as the power of a test, since the truth is not known. But based on the amount of data, one can of course try to make extrapolation to tendencies in clinical science in general. Moreover, it is still interesting to see how the methods behave in general, especially with respect to each other. It may, as an example, be possible to answer the question which statistical test is most conservative and which pooling method is most optimistic on average. Comparison with results from simulations may allow to speculate about the reasons when simulation and real world results diverge.

### 1.0.2 Cochrane and the Cochrane Database of Systematic Reviews

The Cochrane Organization has specialized on systematic reviews in clinical science. It publishes and maintains a library with a large number of systematic reviews that are available in some countries to the public.

The data analyzed in this thesis stems completely from the Cochrane Library of systematic Reviews (cite).

The reviews are arguably of good quality, since the authors are following elaborated guidelines, and there are control-mechanisms within the organisation that should prohibit conflicts of interests. This might further improve the validity and precision of findings and conclusions that have been made based on this data.

## Chapter 2

# The Cochrane Dataset

### 2.0.1 Structure and Content

The dataset consists of 5016 systematic reviews from the Cochrane Library with 52995 studies. Each study provides data of (multiple) comparisons of clinical interventions. In Table 2.1, two comparisons from a systematic review about effects of barbiturates are shown as they are given in the dataset. As can be seen, the comparison is further specified by the variables in the columns. One row of the dataset is one comparison.

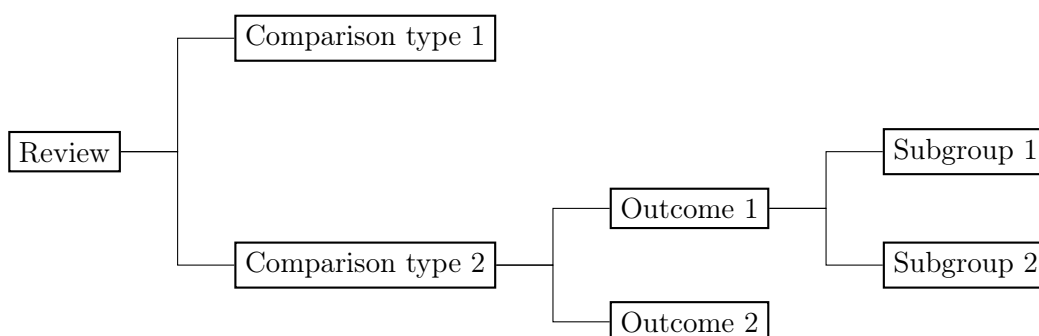
Study	Comparison_type	Outcome	Events	Total	Events_c	Total_c
Bohn 1989	Barbiturate vs no barbiturate	Death at the end of follow-up	11.00	41.00	11.00	41.00
Ward 1985	Barbiturate vs no barbiturate	Death at the end of follow-up	14.00	27.00	13.00	26.00

**Table 2.1:** Example of two comparisons 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 omitted

A complete listing of the variables is given in Table 2.2. They can roughly be separated into variables that specify the review in which the comparison is contained and variables that specify the comparison itself (separated by a horizontal line in Table 2.2).

The structure of a review is shown in Figure ???. The comparison type variable specifies what is compared, the outcome variable how it is compared, and the subgroup variable indicates if the comparison belongs to a certain subgroup. If desired, Figure ??? can be compared to Table 2.3 where an exemplary review is listed.

It is important to not confuse comparisons with studies. A study can contribute multiple comparisons to a systematic review. Also, despite a comparison has variables concerning event counts and means, it can only have one of the two, either means (if the outcome measure is continuous) or event counts (for binary outcomes).



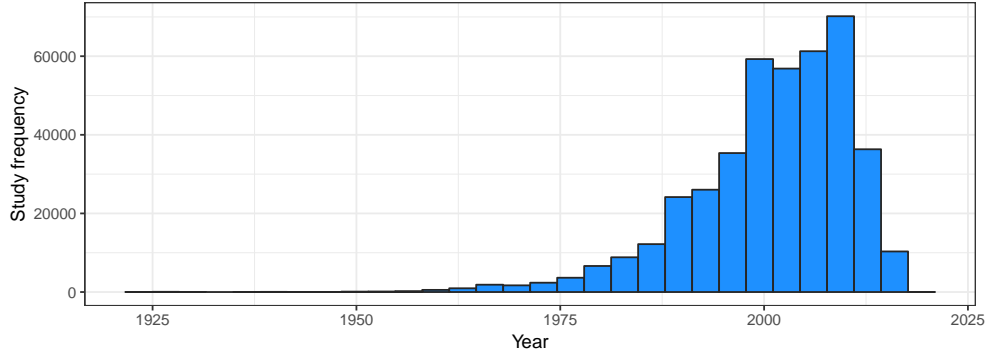
**Figure 2.1:** Structure of a hypothetical review with two different comparisons

Variable	Description
<b>file.nr</b>	The number of the file from which the review data has been gathered. This file corresponds to a file available in the Cochrane library
<b>doi</b>	Digital object identifier. A unique id of the review such that the full text of the review can be found on the web.
<b>file.index</b>	Internal index of the file in the Cochrane library.
<b>file.version</b>	Denotes the version of the review, since the reviews are occasionally updated.
<b>comparison.name/.nr</b> and a unique number for the comparison	Specification of the interventions compared in the study
<b>outcome.name/.nr</b> and a unique number for the outcome	Specification by which outcome the interventions are compared
<b>subgroup.name/.nr</b> unique number for the subgroup	Potentially indication of affiliation to subgroups and a
<b>study.name</b>	Name of the study to which the comparison belongs
<b>study.year</b>	Year in which the study was published
<b>outcome.measure</b>	Indication of the quantification method of the effect (of one intervention compared to the other).
<b>effect</b>	Measure of the effect given in the quantity denoted by “outcome measure”.
<b>events1/events2</b>	The counts of patients with an outcome <i>if</i> measurement/outcome is binary or dichotomous 2 (1 for treatment group and 2 for control group).
<b>total1/total2</b>	Number of patients in groups.
<b>mean1/mean2)</b>	Mean of patient measurements <i>if</i> outcome is continuous.
<b>sd1/sd2</b>	Standard deviation of mean <i>if</i> outcome is continuous.

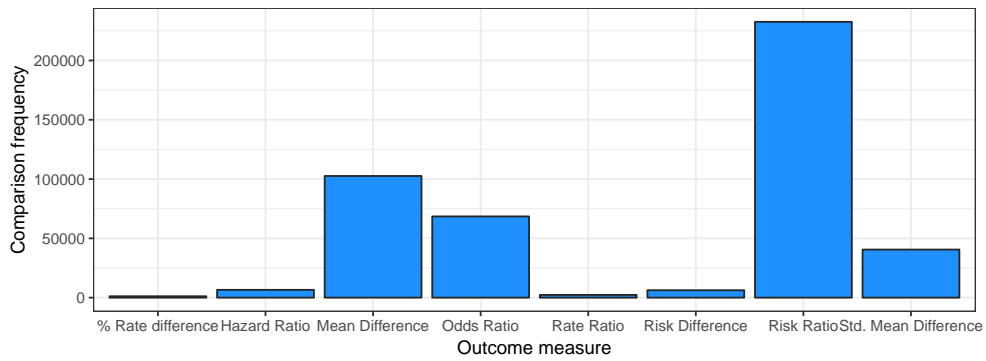
Table 2.2: Dataset variable names and descriptions

Study	Comparison	Outcome
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 2.3: Barbiturate and head injury review. In the columns, study names, comparison types and outcome measure of the comparisons are given



**Figure 2.2:** Frequencies of study publication years in the dataset. 44655 were excluded due to likely wrong indications



**Figure 2.3:** Frequencies of some outcome measures for the effects in the dataset. 5593 measures with other outcome measures are excluded

Having provided an overview over the dataset, now, some more specific information is provided. The dataset consists of 463820 comparisons and has 26 variables that specify the comparisons. Information about missing values in the dataset is given in Table 2.4. For variables as research subject, outcome and subgroup name and event counts there are no missing values. The relative amount of missing values is very low except for study years.

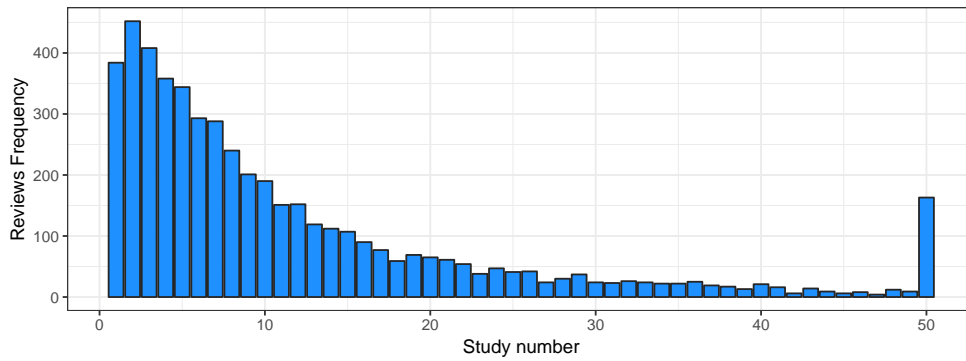
Missing mean values	1287
Missing standard deviations	999
Missing effects	158
Missing study year	27234

**Table 2.4:** Number of missing variables and measurements in the dataset

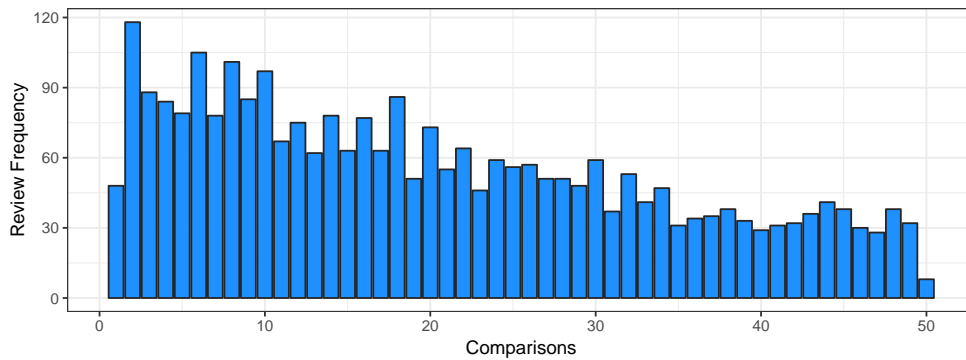
More properties of the reviews, the studies and the comparisons in the dataset will be provided on the following pages. The publication dates of the studies included in the dataset are shown in Figure 2.2. Most studies were published after 2000.

Figure 2.3 provides the frequencies of outcome types of the comparisons. Note that the abundance of mean differences and standardized mean differences can also give an impression of the proportion of continuous outcome comparisons vs. binary outcome comparisons in the dataset.

It is also possible to look at the properties of the reviews. One question could be how many studies or comparisons that a review comprises. The former is shown in Figure 2.4 and the latter in Figure 2.5. It can be seen that while almost 400 reviews consist of one study only, there are more than 150 with equal or more than 30 distinct studies. A similar variance between reviews



**Figure 2.4:** Empirical distribution of number of studies per review



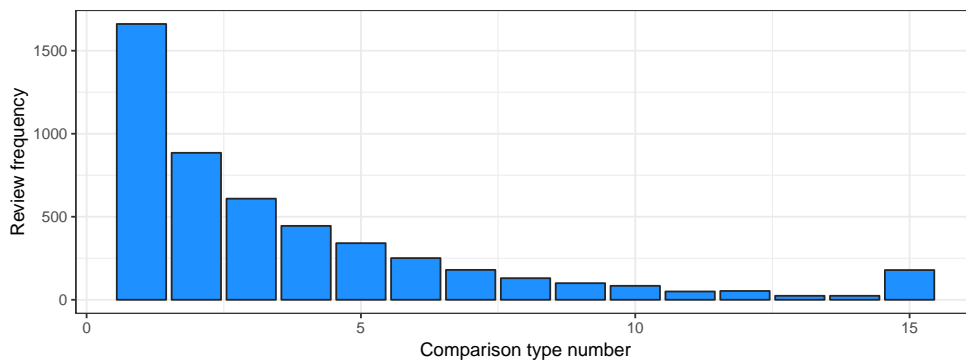
**Figure 2.5:** Empirical distribution of number of comparisons per review

can also be observed when looking at the number of comparisons.

A question not to be mistaken with the previous would be how many comparison *types* there are per review. This gives an additional impression of the scope of a review. Analogously to the previous figures, the empirical distribution of comparison types is depicted in Figure 2.6.

For comparisons to be suitable for usage in meta-analysis, they have to be somewhat identical (same comparison type, outcome measure and possibly subgroup). For an analysis of reporting bias, again a certain number of studies is required in order for reporting bias to be detectable by the methods. One question would therefore be: How many groups of identical comparisons of a certain size are given in the dataset? This depends on which degree of similarity between comparisons is considered to be sufficient.

In Table 2.5, two different similarity criteria have been used. One is based on the same



**Figure 2.6:** Empirical distribution of number of different comparison types per review

comparison type and outcome measure, the other includes additionally subgroup affiliation of comparisons, i.e. only comparisons in the same subgroups are considered to be similar enough.

Table 2.5 shows the cumulative number of *groups* of comparisons with equal or more than  $n$  comparisons. Practically, this means that this number of meta analyses can be performed with each having at least  $n$  comparisons.

n	Cumulative sum (without subgroups)	Cumulative sum (with subgroups)
1	109191	186300
2	67699	83956
3	47800	52270
4	36169	36198
5	28090	26570
6	22702	20126
7	18547	15896
8	15475	12935
9	13008	10821
10	11008	9229
11	9362	7991
12	8057	7070
13	6988	6368
14	6044	5783
15	5328	5328

**Table 2.5:** Cumulative number of groups with number of reproduction trials  $\geq n$





# Chapter 3

## Results

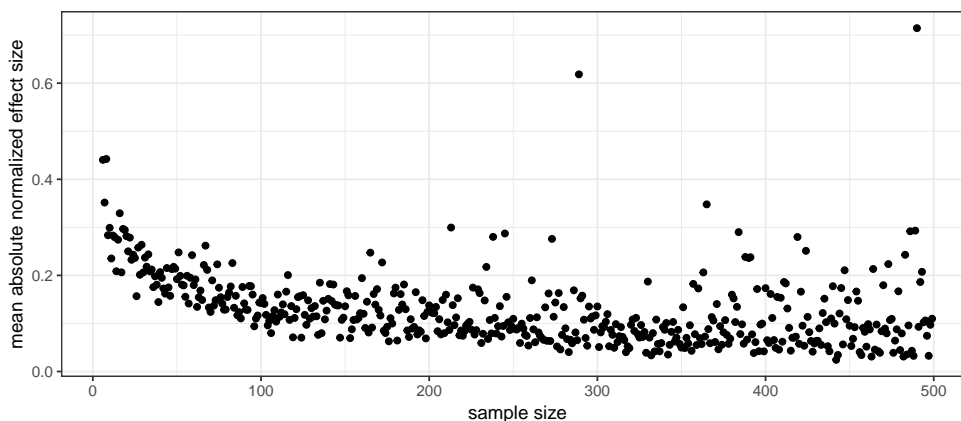
### 3.1 Small Study Effects

One crucial assumption in meta analysis is that the availability and publication of studies does not depend on their effect and the variance of the effect. If this is not given, one often speaks of publication bias. In fact, there can also be other reasons for this (see discussion section). A more appropriate term for the phenomenon is small study effect. If small study effects are present in a meta-analysis, the classical approaches to merge single study results in to an overall intervention effect fails to provide an appropriate estimate of the treatment effect.

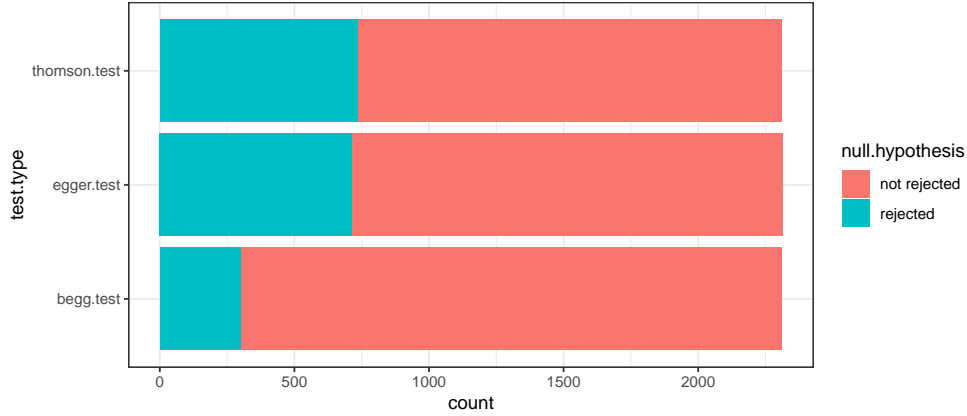
To provide an overview over the issue, first it is shown how mean absolute effect size decreases with increasing sample size of the comparisons in Figure 3.1. The decrease is particularly substantive from very small trials ( $n = 5$ ) to medium sample size ( $n = 100$ ), afterwards the trend is less pronounced. With increasing sample size, there are fewer results, therefore, the variation between means increases. All effects are normalized by subtracting the mean effect size of the dataset and dividing through the standard deviation. Note that various types of outcome measures are included, such as mean difference and risk ratios, and are normalized with respect to all sample sizes.

#### 3.1.1 Small Study Effect Tests

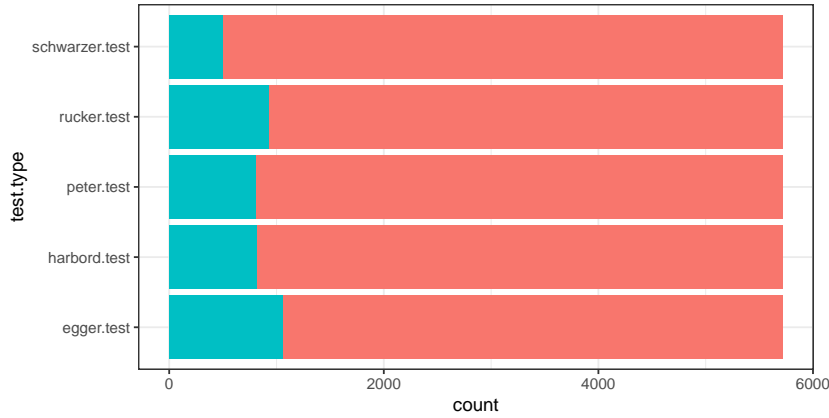
There are tests that can be applied to find out if reporting bias is present in the meta analysis. For the precise description, see the methods section. Application of the tests is only recommended if there are ten or more studies ([Higgins JPT, 2011](#)) that can be used, so all meta-analyses with



**Figure 3.1:** Mean of the absolute of the normalized effect size plotted against the total sample size.



**Figure 3.2:** Proportion where the null hypothesis of no small study effect is rejected based on the 5% significance level for different tests (continuous outcomes).



**Figure 3.3:** Proportion where the null hypothesis of no small study effect is rejected based on the 5% significance level for different tests (dichotomous outcomes).

less than ten studies have been excluded.

There are modifications to make tests more appropriate in case of binary outcomes, therefore the results have been separated in continuous and dichotomous outcome test results. In Figure 3.2 the proportion of test results that led to rejection of the null hypothesis of no small study effect based on the 5 % level are shown for continuous outcomes. From 2312 meta-analyses with continuous outcomes, most meta-analyses with rejected null hypotheses have been found by the Thompson and Sharp test (0.319), followed closely by Eggers test (0.309) and Begg and Mazumdar's test (0.131). The same is shown in Figure 3.3 for dichotomous outcome measures ( $n = 5717$ ). The test that rejects most null hypotheses is here Egger's test (0.186), followed by Rücker's test (0.163), Harbord's test (0.144), Peters test (0.142) and Schwarzer's test (0.088).

The agreement of the tests, i.e. the proportion of meta-analyses where the test results are equal between tests, is shown in Table 3.1 and Table 3.2, again separated for outcome types. Agreement in tests for binary outcomes is better than continuous outcomes, with some variation between tests (binary outcomes: 83 to 91%). Correlation varies more between tests, both for continuous and binary outcome tests.

Test performance depends on the sample size, despite having restricted sample size to a minimum of 10 studies. The p-values of two tests are shown with respect to the sample size of the studies in Figure 3.4 for peters test based on dichotomous outcomes and in Figure 3.5 for thomson and sharp's test based on continuous outcomes. A trend towards more significant p-values is visible for larger sample sizes.

	Test Agreement	P-value Correlation
egger.schwarzer	0.83	0.38
egger.peter	0.84	0.44
egger.rucker	0.83	0.39
egger.harbord	0.91	0.75
schwarzer.peter	0.85	0.24
schwarzer.rucker	0.83	0.30
schwarzer.harbord	0.88	0.52
rucker.peter	0.89	0.66
harbord.peter	0.86	0.45

**Table 3.1:** Proportion of tests that agree in rejection or acceptance of the null hypothesis that there is no small study effect (dichotomous outcomes)

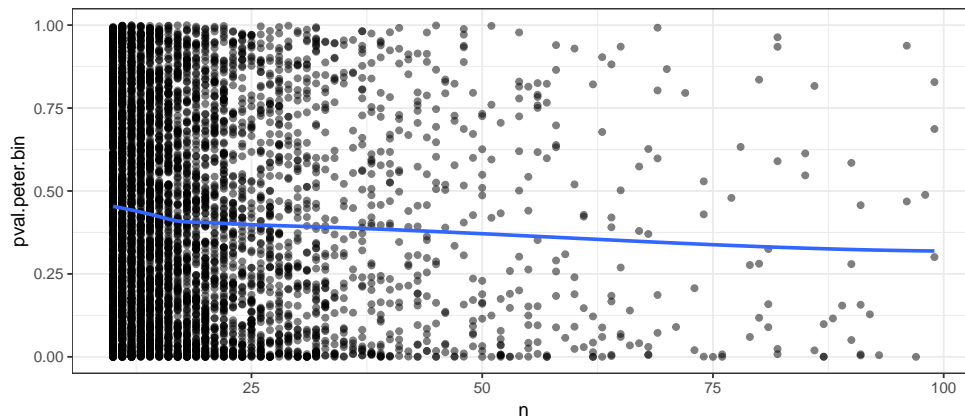
	Test Agreement	P-value Correlation
thomson.egger	0.85	0.67
thomson.begg	0.75	0.45
egger.begg	0.75	0.38

**Table 3.2:** Proportion of tests that agree in rejection or acceptance of the null hypothesis that there is no small study effect (continuous outcomes)

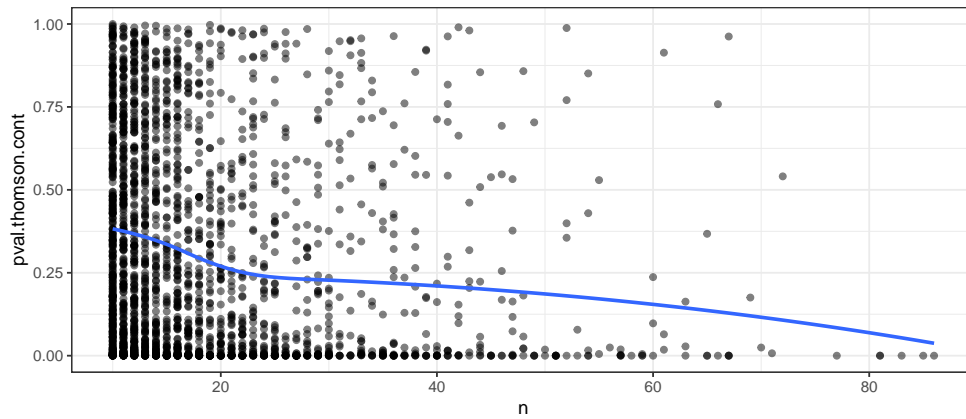
One can use the proportion of added studies by the trim-and-fill method from the overall number of studies to further investigate the extent of small study effects. The mean fraction of trimmed comparisons for binary outcomes is 0.17 and the median 0.15. A histogram with those fractions is shown in Figure 3.6 for continuous outcomes and 3.7 for dichotomous outcomes. In both cases, the method very commonly adds supposedly unpublished studies to the meta-analyses. In Figure 3.8 and Figure 3.9, the relationship between fraction of added studies by trim-and-fill and the hypothesis test decisions of the small study effects tests is shown for continuous and dichotomous outcomes. In the case of Peters test for dichotomous outcomes, there is less agreement with the trim-and fill method than in the case of Thomson and Sharp's test for continuous outcomes in the sense that the fraction of meta-analyses with rejected null hypotheses increases more clearly when there are more studies added by trim-and-fill.

The mean fraction of trimmed comparisons for continuous outcomes is 0.22 and the median 0.2.

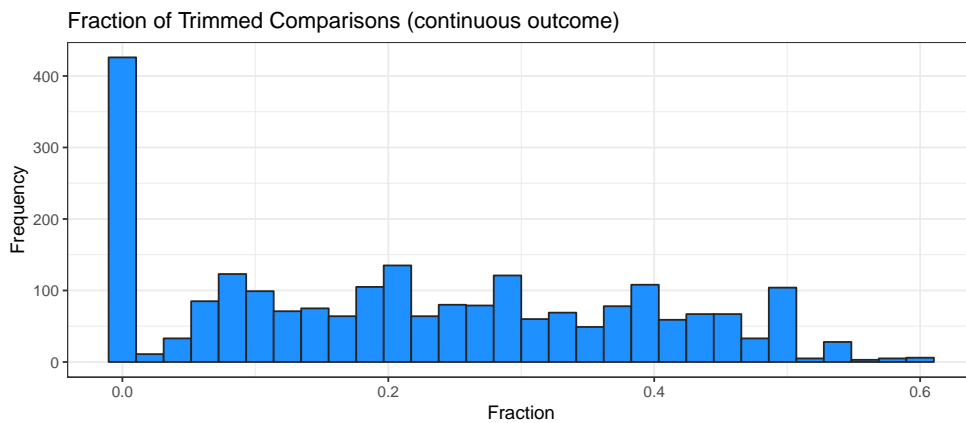
The same is repeated for binary outcomes in figure 3.7.



**Figure 3.4:** P-values of Peters test for small study effects and their corresponding sample size (dichotomous outcomes)



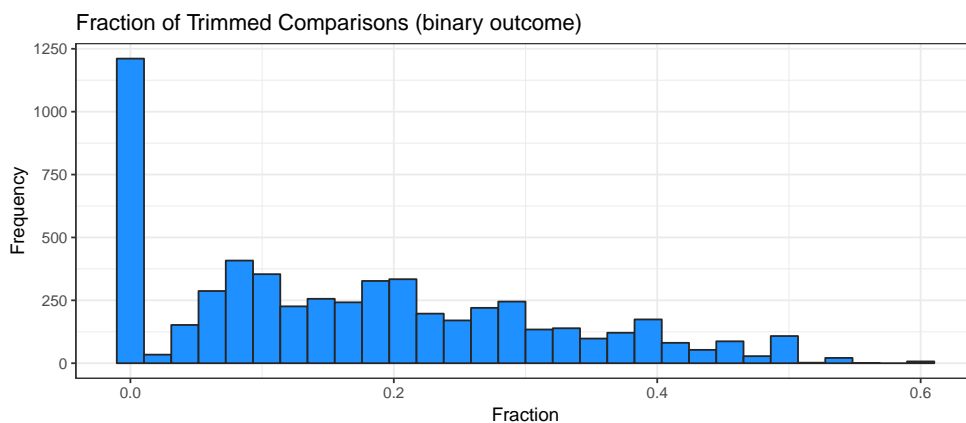
**Figure 3.5:** P-values of thomson and sharp's test for small study effects and their corresponding sample size (continuous outcomes)



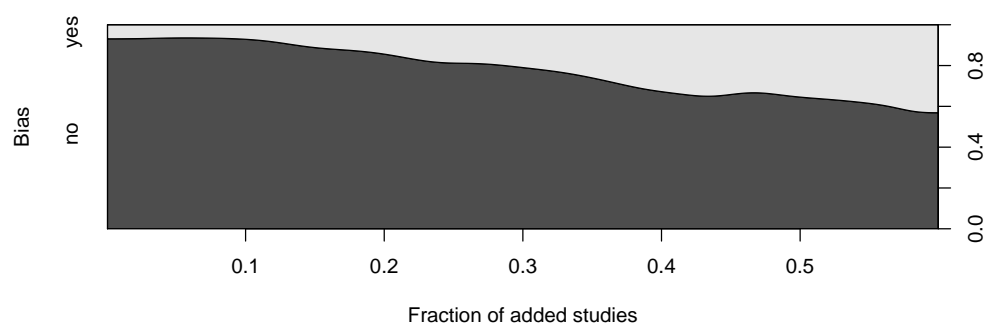
**Figure 3.6:** Histogram of fractions of trimmed comparisons from meta analyses with continuous outcomes.

### 3.1.2 Small Study Effect Correction

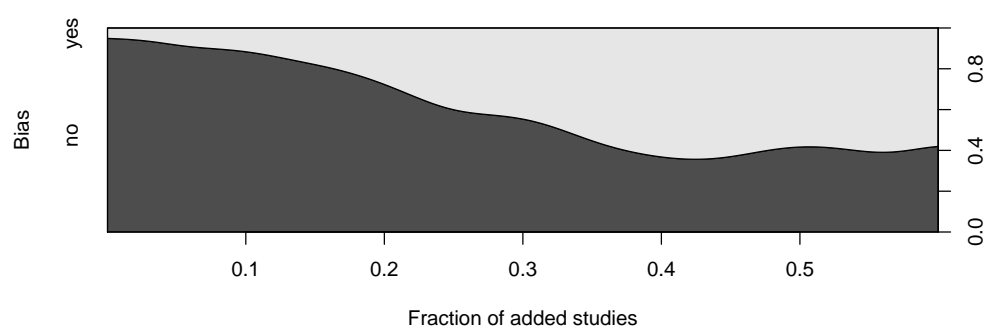
Multiple methods are available to correct for the effects of small study effects in order to get an unbiased estimate.



**Figure 3.7:** Histogram of fractions of trimmed comparisons from meta analyses with binary outcomes.



**Figure 3.8:** Fraction of added studies for small study effect correction by trim and fill and the corresponding small study effect test decision (based on 5% significance level) of Peters test and dichotomous outcomes



**Figure 3.9:** Fraction of added studies for small study effect correction by trim and fill and the corresponding small study effect test decision (based on 5% significance level) of Thomson and Sharp's test and continuous outcomes



# Chapter 4

## Methods

### 4.1 Basic notation

The notation used here will be used throughout the chapter and exceptions will be noted. Let  $i$  be the number of a study of a meta analysis with  $n$  being the total number of studies.  $y_i$  is then the effect size estimate (usually log odds ratio or mean difference) and  $v_i$  the variance of the estimate of study  $i$ .  $w_i$  is used for weights which are defined when necessary, and  $\Delta$  usually denotes the summarized or pooled effect estimate of the meta-analysis, and  $\eta$  the variance thereof.

In the case of binary outcomes, let  $e_t$  be the number of events and  $n_t$  be the total number of patients in the treatment arm and  $n_c$  and  $e_c$  analogously for the control arm in a two-armed study  $i$ .

### 4.2 Heterogeneity

In addition to sampling error, there can be additional, “real” variation between estimates of different studies, indicating real differences between the studies. This is called between study variation in contrast to within study variation (noise).

The  $Q$  statistic is a weighted sum of squares that quantifies the deviation from the weighted mean of study effect estimates. Let  $w_i$  be the inverse of the variance and  $\Delta$  be a summarized effect estimate of your choice as for example a variance-weighted mean 4.6. Then  $Q$  can be calculated as in 4.2

$$\Delta = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \quad (4.1)$$

$$Q = \sum_{i=1}^n w_i (y_i - \Delta)^2 \quad (4.2)$$

Because  $Q$  is a standardized measure, it does not depend on the effect size, but only on the study number  $n$ . Under the assumption of equal effect sizes of all studies, the expected value of  $Q$  is  $n - 1$ , so the excess dispersion is just  $Q - n + 1$ . To test the assumption of equal effect sizes one uses that  $Q$  follows a central Chi-squared distribution with  $n - 1$  degrees of freedom under the null hypothesis of equal effect sizes.  $1 - F(Q)$  will provide the  $p$ -value for the significance test with  $F$  being the cumulative distribution function of the Chi-squared distribution with the corresponding degrees of freedom.

Since  $Q$  is a standardized metric, it gives no impression of the real dispersion of the effect sizes. For this purpose,  $\tau^2$ , the variance of true effects, can be calculated.  $\tau^2$  is on the same scale as the effect size and reflects the absolute amount of dispersion. In practice,  $\tau^2$  can be smaller than zero, then it is set to zero.

$$C = \sum_{i=1}^n w_i - \frac{\sum_{i=1}^n w_i^2}{\sum_{i=1}^n w_i} \quad (4.3)$$

$$\tau^2 = \max(0, \frac{Q - d}{C}) \quad (4.4)$$

The estimation method for  $\tau^2$  is known as DerSimonian and Laird method, but others, such as restricted maximum likelihood can be used. Note that their estimate can differ substantially and consequently also the estimate of the pooled effect size estimate.

To estimate the proportion of real variance between effect estimates of the observed variance, the  $I^2$  can be used. The calculation is given in 4.5

$$I^2 = (Q - n + 1)/Q \quad (4.5)$$

There are ways to compute confidence intervals for  $I^2$  and  $\tau^2$  that are not shown (see (Borenstein *et al.*, 2011, 122)).

### 4.3 Meta Analysis

There are numerous methods to pool the estimates of multiple studies into one estimate, and two will be introduced here; fixed and random effects meta-analysis. First the fixed effect meta-analysis will be explained. Note that both methods can be used for continuous or dichotomous outcomes. For more details about the methods, see chapter 11 and 12 in Borenstein *et al.* (2011)

Let  $w_i = 1/v_i$  be the inverse of the variance of the estimate from study  $i$ . The pooled estimate  $\Delta_f$  of the fixed effects model is then the weighted average with the weights given by the inverses of the variances,  $w_i$ , given in 4.6. The variance  $\eta_f$  is the reciprocal of the sum of the weights as shown in 4.7.

$$\Delta_f = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \quad (4.6)$$

$$\eta_f = \frac{1}{\sum_{i=1}^n w_i} \quad (4.7)$$

The computation of random effects meta-analysis is more complicated. Random effects meta-analysis will give smaller studies with larger variance more weight in the pooled estimate. Shortly, the idea is that the estimates are allowed to vary randomly around the true estimate  $\Delta$ , and additionally, the estimates are subject to noise or sampling error themselves.

The variance of a study estimate  $y_i$  of study  $i$ ,  $v_i^*$  is defined as in 4.8, with  $w_i^*$  being the inverse of  $v_i^*$ . It is used to calculate a new weighted mean to obtain a pooled estimate  $\Delta_r$  as in 4.9. The variance of  $\Delta_r$ ,  $\eta_r$  is then the sum of the reciprocal variances (4.10).

$$v_i^* = v_i + \tau^2 \quad (4.8)$$

$$\Delta_r = \frac{\sum_{i=1}^n w_i^* y_i}{\sum_{i=1}^n w_i^*} \quad (4.9)$$

$$\eta_r = \frac{1}{\sum_{i=1}^n w_i^*} \quad (4.10)$$

A p-value under the Null-hypothesis of  $\Delta = 0$  can be obtained by calculating the Z-value (4.11) and using the distribution function of a standard normal as shown in (4.12),  $\Phi$  being the distribution function of a standard normal distribution.



$$Z = \frac{\Delta}{\sqrt{\nu}} \quad (4.11)$$

$$p = 2(1 - \Phi(|Z|)) \quad (4.12)$$

## 4.4 Small Study Effects Tests

### 4.4.1 Continuous Outcome Tests

#### 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  $y_i^*$  can be computed as in 4.13.  $v_i^*$  is the variance of  $y_i - \Delta_f$  as defined in 4.14.  $\Delta_f$  is the pooled estimate for fixed effect estimate defined in 4.6.

$$y_i^* = (y_i - \Delta_f)/v_i^* \quad (4.13)$$

$$v_i^* = v_i - 1/\sum_{i=1}^n v_i^{-1} \quad (4.14)$$

A rank correlation test based on Kendall's tau is then used. The pairs  $(y_i^*, v_i^*)$  that are ranked in the same order are enumerated. Let  $u$  be the number of pairs ranked in the same order, and  $l$  the number of pairs ranked in the opposite order (e.g. larger standardized effect size and smaller variance). Then the normalized test statistic  $Z$  is given in 4.15.

$$Z = (u - l)/\sqrt{n(n-1)(2n+5)/18} \quad (4.15)$$

The changes in the case of ties are negligible (Begg, 1988, 410).

#### Egger's Test: Linear Regression Test

Alternatively, one can use Eggers test (Egger *et al.*, 1997) that is based on linear regression. Let  $y_i^* = y_i/\sqrt{v_i}$  and  $x_i = 1/\sqrt{v_i}$ . Using  $y_i^*$  as dependent, and  $x_i$  as explanatory variable in linear regression, one obtains an intercept  $\beta_0$  and a slope.

If  $\beta_0 \neq 0$ , the null hypothesis of no small study effect may be contested, using that  $\beta_0 \sim t_{n-1}$ ,  $n-1$  being the degrees of freedom of the  $t$ -distribution. The p-value for  $\beta_0 = 0$  (no reporting bias) is then given by 4.16.

$$p = 2 * (1 - t_{n-1}(\beta_0/se(\beta_0))) \quad (4.16)$$

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

A method proposed in Thompson and Sharp (1999) allows for between study heterogeneity. Let  $\tau^2$  be equal to 4.4. The effect size estimates are then assumed to be distributed as in 4.17.

$$y_i \sim N(\beta_0 + \beta_1 x_i, v_i + \tau^2) \quad (4.17)$$

Then, a weighted regression is carried out with weights  $1/v_i^*$  based on the inverse of the variance as in 4.8. Analogous to Edger's test,  $\beta_0$  is then tested with respect to the null hypothesis  $\beta_0 = 0$ .

#### 4.4.2 Dichotomous Outcomes Tests

The issue with dichotomous outcomes is that effect size and variance of effect size are not independent. Consequently, the tests above will tend to reject the null-hypothesis too often, i.e. that they are not conservative enough. A number of solutions to this problem are existing in the literature.

##### Peters Test: Weighted Linear Regression Test

A modification of the weighted linear regression test that takes into account effect size and variance interdependence for dichotomous outcomes is proposed in [Peters \*et al.\* \(2006\)](#).

Let  $y_i$  be the log-odds ratio estimate [4.18](#) and  $v_i$  its variance [4.19](#)

$$y_i = \log(e_t * (n_c - e_c) / e_c * (n_t - e_t)) \quad (4.18)$$

$$v_i = 1/(e_t + (n_t - e_t)) + 1/(e_c + (n_c - e_c)) \quad (4.19)$$

and  $x_i$  be the total sample size  $n_t + n_c$ . Instead of taking the variance as explanatory or independent variable in regression as in Egger's Test, the inverse of the total sample size  $x_i$  is used, and the variance  $v_i$  is used as a weight. The subsequent test procedure is then identical to Egger's test.

Peters test is a small modification of Macaskill's test where the explanatory variable is the sample size instead of its inverse.

##### Harbord's Test: Score based Test

A rank based alternative to Peters test for binary outcomes is the Harbord's test ([Harbord \*et al.\*, 2006](#)). The score  $r_i$  (the first derivative of the log-likelihood of a proportion with treatment effect equal 0) and its variance  $v_i$  can be computed as shown in [4.20](#) and [4.21](#).

$$r_i = e_t - (e_t - e_c)(e_t + (n_t - e_t)) / (n_t + n_c) \quad (4.20)$$

$$v_i = \frac{(e_t + e_c)(e_t + (n_t - e_t))(e_c + (n_c - e_c))((n_t - e_t) + (n_c - e_c))}{(n_t + n_c)^2(n_t + n_c - 1)} \quad (4.21)$$

Similarly to Egger's or Peters Test, now a weighted linear regression can be performed on  $r_i/v_i$  with the standard error  $1/\sqrt{v_i}$  as explanatory variable and  $1/v_i$  as a weight. Note that  $r_i/v_i$  is also known as peto odds ratio.

##### Schwarzer's Test: Rank Correlation Test

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

The standardized cell count deviation  $(e_t - \mathbb{E}(E_t))/\sqrt{v_i}$  and the inverse of  $v_i$  is then used in the way as before in Begg and Mazumdar's test.

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

The correlation between variance and effect size of dichotomous outcome measures can be abolished by the variance stabilizing transformation for binomial random variables. Let

$$y_i = \arcsin e_t / n_t - \arcsin e_c / n_c / v_i = 1/4n_t + +/4n_c \quad (4.22)$$

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

## 4.5 Small Study Effect Adjustment

### 4.5.1 Trim and Fill

One method to account for reporting bias in meta-analysis is to apply the Trim and Fill adjustment method (Duval and Tweedie, 2000). It is a nonparametric test based on a funnel plot, on which the effect size estimates of studies are plotted against their standard error.

The algorithm for the method tries to estimate the number of studies  $k$  that are not available due to reporting bias (different estimators are available for  $k$ ). First,  $\Delta$  is estimated using a fixed or random effects model. Then, the  $k$  effect size estimates with the smallest standard errors are trimmed, and  $\Delta$  is estimated again. The procedure is repeated until  $k$  is 0 and the funnel plot is symmetric. The total number of missing studies is then mirrored with respect to the final effect size estimate  $\Delta$ , and  $\Delta$  and its standard error is then computed to obtain an unbiased estimate.

### 4.5.2 Copas Selection Model

A method proposed in Copas and Shi (2001, 2000); Copas and Malley (2008) assumes that there is a population of studies of which only a part has been published dependent on the variance and size of their estimated effects. Studies with small variance and large effect sizes are more likely to be published than studies with large variance and small effect sizes. Note that small effect size means here a treatment effect close to the control effect.

Let  $y_i$  be the effect size estimate of study  $i$ . Then

$$y_i \sim N(\mu_i, \sigma_i^2) \quad (4.23)$$

$$\mu_i \sim N(\mu, \tau^2) \quad (4.24)$$

corresponding to a standard random effects model.  $\mu$  is the overall mean effect,  $\sigma_i^2$  the within study variance and  $\tau^2$  the between study variance. This is the *population model*.

The *selection model* is defined as follows. Suppose a selection of studies with reported standard errors  $s$  (likely different from  $\sigma$ ). Only a proportion

$$P(\text{select}|s) = \Phi(a + b/s) \quad (4.25)$$

of the selection will be published, with  $a$  defining the overall proportion of published studies and  $b$  (assumed to be positive) defining how fast this proportion increases with  $s$  becoming smaller. 4.25 can be rewritten as

$$z = a + b/s + \delta \quad (4.26)$$

with  $\delta \sim N(0, 1)$ . The study with standard error  $s$  is only selected if  $z$  is positive. Therefore, the larger  $z$ , the more likely the study is selected. Combining population and selection model for study  $i$ , we have

$$y_i = \mu_i + \sigma_i \epsilon_i \quad (4.27)$$

$$\mu_i \sim N(\mu, \tau^2) \quad (4.28)$$

$$z_i = a + b/s_i + \delta_i \quad (4.29)$$

where  $(\epsilon_i, \delta_i)$  are standard normal residuals and jointly normal with correlation  $\rho = \text{cor}(y_i, z_i)$ . Every given study  $i$  in the meta-analysis has  $z_i > 0$ . If  $\rho$  is large and positive and  $z_i > 0$ , then the estimate of a study  $i$  that is selected is likely to have positive  $\epsilon_i$  and  $\delta_i$ . Thus, the true mean  $\mu$  is likely to be overestimated.

Let  $u = a + b/s$ ,  $\lambda(u) = \phi(u)/\Phi(u)$  ( $\phi$  is the standard normal density function) and  $\tilde{\rho} = \sigma/\sqrt{(\tau^2 + \sigma^2)\rho}$ . The probability of a study being selected is

$$P(\text{select}|s, y) = P(a, b, s, y) = \Phi\left(\frac{u + \tilde{\rho}((y - \mu)/\sqrt{(\tau^2 + \sigma^2)})}{\sqrt{1 - \tilde{\rho}^2}}\right) \quad (4.30)$$

It can also be shown that the expected value

$$\mathbb{E}(y|s, \text{select}) = \mu + \rho\sigma\lambda(u) \quad (4.31)$$

which shows that the expected value for a study is larger for larger  $\sigma$ .

One can compute a likelihood function based on the distribution of  $y$  conditional on  $z > 0$ . The likelihood can be maximized for any given pair  $a, b$  (can not be estimated since the number of missing studies is not known), and a maximum likelihood estimate  $\hat{\mu}$  for the true mean  $\mu$  can be obtained. One can then perform a sensitivity analysis. First, one looks how  $\hat{\mu}$  changes for different values of  $a, b$ . One can then compare the fitted values in 4.31 with the real values. To test the fit of the model (while keeping all other parts unchanged), the model can be extended in the following way :

$$y_i = \mu_i + \beta s_i + \sigma_i \epsilon_i \quad (4.32)$$

If we accept  $\beta = 0$ , then we accept that the selection model has satisfactorily explained any relationship between  $y$  and  $s$ . Only if the value is large enough, typically  $p > 0.05$ , one concludes that the selection model has explained the observed data. The p-value is obtained by a likelihood ratio test comparing the maximum of the likelihood with the  $\beta$  term added and without it, and by a likelihood ratio test.

To find out if the null-hypothesis of, say,  $\mu = 0$  can be rejected, another likelihood ratio test can be performed, this time with imputing  $\mu = 0$  and comparing the two maximum likelihoods.

In practice, only a range of values for  $a, b$  are reasonable. For those values, the quantities above can be calculated and illustrated. Values for  $\mu$  that have p-values over a predefined significance threshold can be used for inference of the effect size.

### 4.5.3 Adjustment by Regression

There are multiple ways to adjust for small study effects by regression. The general idea is to regress the effect size of a study with a variance of zero based on the given effects and variances.

Rücker *et al.* (2011) use a random effects model together with shrinkage procedure to obtain an unbiased estimate. Similarly to what has been seen in Copas selection model, we let  $y_i$  depend on the intercept  $\beta_0$  and on its standard error  $\sqrt{v_i}$  as in 4.36.

$$y_i = \beta_0 + \beta_1(\sqrt{v_i + \tau^2}) + \epsilon_i(\sqrt{s_i + \tau^2}), \stackrel{\text{iid}}{\sim} N(0, 1) \quad (4.33)$$

$\beta_1$  represents the bias introduced by small study effects, as can be seen when looking at ??

$$\mathbb{E}((y_i - \beta)/\sqrt{v_i}) \rightarrow \beta_1 \text{ if } \sqrt{v_i} \rightarrow \infty \quad (4.34)$$

$$\mathbb{E}(y_i) \rightarrow \beta_0 + \beta_1 \tau \text{ if } \sqrt{v_i} \rightarrow 0 \quad (4.35)$$

After estimating  $\tau^2$ , one can estimate  $\beta_0$  and  $\beta_1$  as seen before e.g. in Thompson and Sharp's Test with weights also equal to Thompson and Sharp's Test.

To diminish the random variation within studies, but keep the variation between studies, we change 4.36 to a scenario where each study has  $M$ -fold increased precision:

$$y_i = \beta_0^* + \beta_1^*(\sqrt{v_i/M + \tau^2}) + \epsilon_i(\sqrt{s_i/M + \tau^2}) \quad (4.36)$$

Letting  $M \rightarrow \infty$ , we obtain:

$$y_{\infty,i} = \beta_0^* + \tau(\beta_1^* + \epsilon_i), \epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1) \quad (4.37)$$

Note that:

$$y_{\infty,i} = \beta_0^* + \tau\beta_1^* = \beta_0 \quad (4.38)$$

$\beta_0$  is termed the limit meta analysis expectation. Now, the random errors from 4.36 are rewritten as:

$$\epsilon_i = \frac{y_i - \beta_0^*}{\sqrt{v_i + \tau^2}} - \beta_1^* \quad (4.39)$$

Assuming  $\epsilon_i$  to be fixed, we can plug it into 4.37 and get

$$y_{\infty,i} = \beta_0^* + \sqrt{\frac{\tau^2}{v_i + \tau^2}}(y_i - \beta_0^*) \quad (4.40)$$

By estimating  $\tau^2$ ,  $v_i$  and  $\beta_0^*$ , we can use the formula to obtain a new study means, adjusted for small study effects and shrunk to a common mean.



## Chapter 5

## Conclusions





## Appendix A

# Appendix

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



# Bibliography

- Begg, C. B. (1988). Statistical methods in medical research p. armitage and g. berry, blackwell scientific publications, oxford, u.k., 1987. no. of pages: 559. price £22.50. *Statistics in Medicine*, **7**, 817–818. [17](#)
- 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. [16](#)
- Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis . *Biostatistics*, **1**, 247–262. [19](#)
- 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. [19](#)
- 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. [19](#)
- 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. [19](#)
- 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. [17](#)
- 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. [18](#)
- Higgins JPT, G. S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration. [9](#)
- Peters, J., Sutton, A., R Jones, D., Abrams, K., and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA : the journal of the American Medical Association*, **295**, 676–80. [18](#)
- Rücker, G., Carpenter, J. R., and Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal*, **53**, 351–368. [20](#)
- 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. [18](#)
- Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, **18**, 2693–2708. [17](#)

