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 7, 2019

Contents

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 keep 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, since the amount of empirical research is expanding, meta-analysis can be part in any field where repeated experiments are performed. The usefulness of meta-analysis expands therefore well over clinical 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. Systematic reviews may also incorporate results that can not be included in a meta-analysis, such as adverse effects reported in one trial, but a meta-analysis is the central part of a systematic review if results from multiple studies are available. The authors usually take their conclusions largely based on the results of the meta-analysis.

However, there are problems that potentially limit the validity of meta-analysis; studies at hand can be biased or heterogeneity between study results can be large and the number of studies small. The importance and the issues of meta-analysis are the reasons why they have been chosen as one general topic of the masters thesis. One particular problem will furthermore be investigated in more detail: reporting bias and meta-analysis. Not only will the methods to deal with issues as reporting bias be discussed, but also will they be applied on a dataset of systematic reviews that can be used for meta-analysis. So at the end of the report, the reader will not only have an impression of the technical issues caused by reporting bias, but also of the abundance and extent of it in the dataset. Since the dataset is very large and of good quality, results might also be representative to some extent for reporting bias in clinical science.

1.0.1 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 ommitted

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).

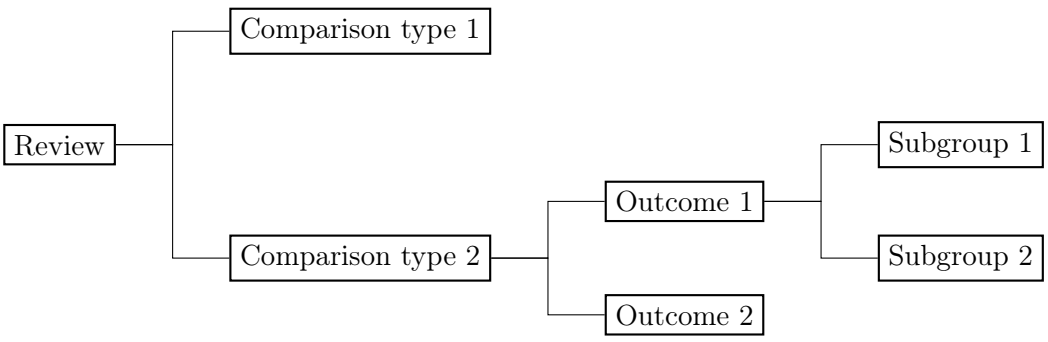

Figure 2.1: Structure of a hypothetical review with two different comparisons

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

Table 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

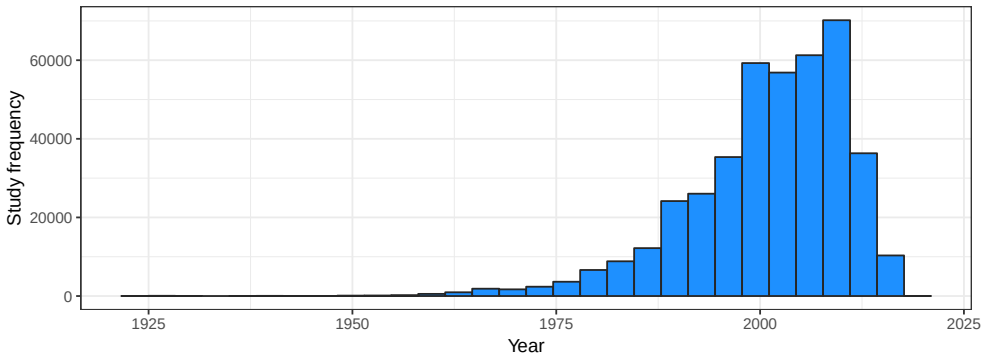

Figure 2.2: Frequencies of study publication years in the dataset. 44655 were excluded due to likely wrong indications

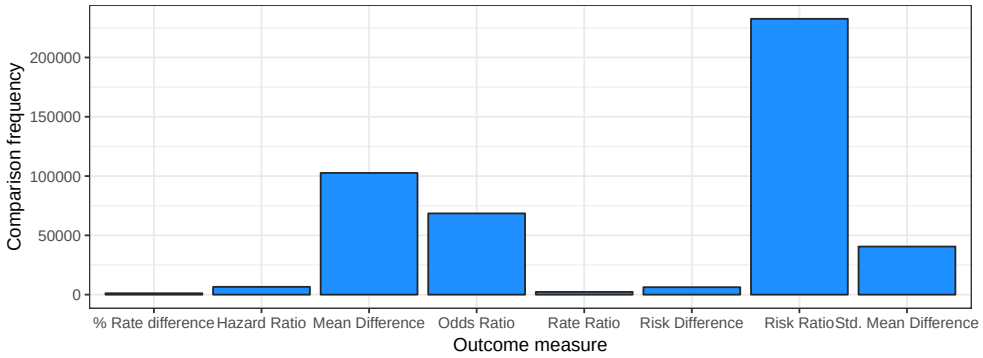

Figure 2.3: Frequencies of some outcome measures for the effects in the dataset. 5593 measures with other outcome measures are excluded

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

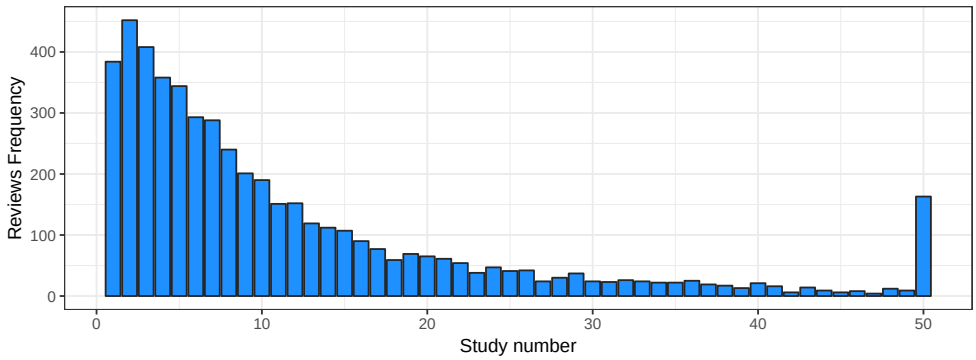

Figure 2.4: Empirical distribution of number of studies per review

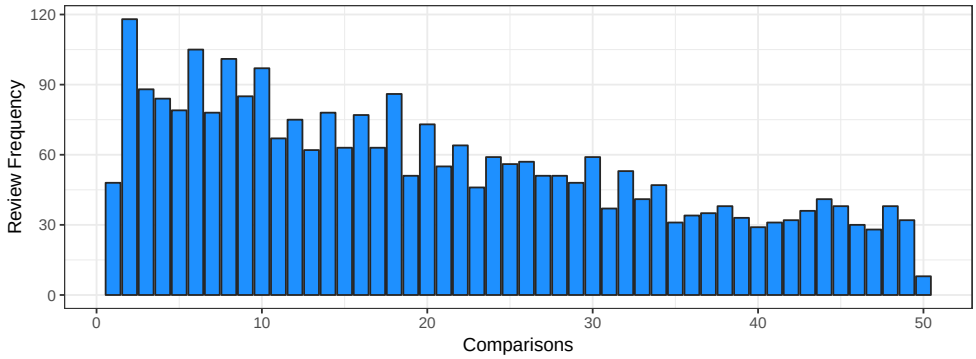

Figure 2.5: Empirical distribution of number of comparisons per review

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

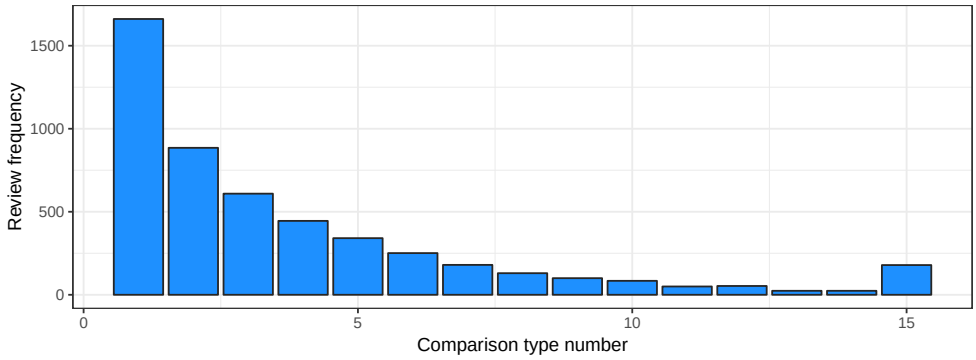

Figure 2.6: Empirical distribution of number of different comparison types per review

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 >= n

Chapter 3

Results

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 effect size decreases with increasing sample size of the comparisons in Figure 3.1. All effects are normalized by subtracting the mean effect size of the dataset and dividng through the standard deviation. Then, the mean of the absolute value for a given sample size is plotted against its sample size. 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.

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 that can be used, so all meta-analyses with less than ten studies have been excluded.

There are modifications to make tests more appropriate in case of binary outcomes, therefore, the results of the tests are separated in continuous outcome test results and dichotomoized outcome 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 dichotomous outcomes. The same is shown in Figure 3.3 for continuous outcome measures.

Furthermore, one can investigate how well the tests give the same results, i.e. classify the

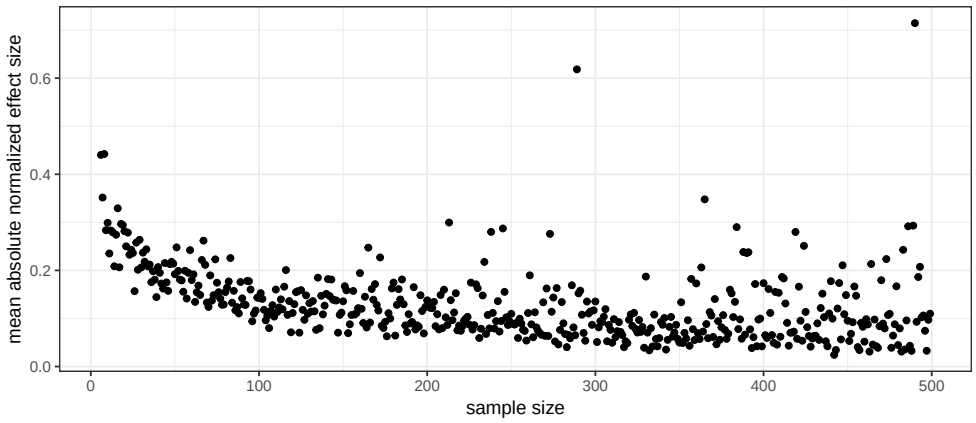

Figure 3.1: Mean of the absolute of the normalized effect size plotted against the total sample size.

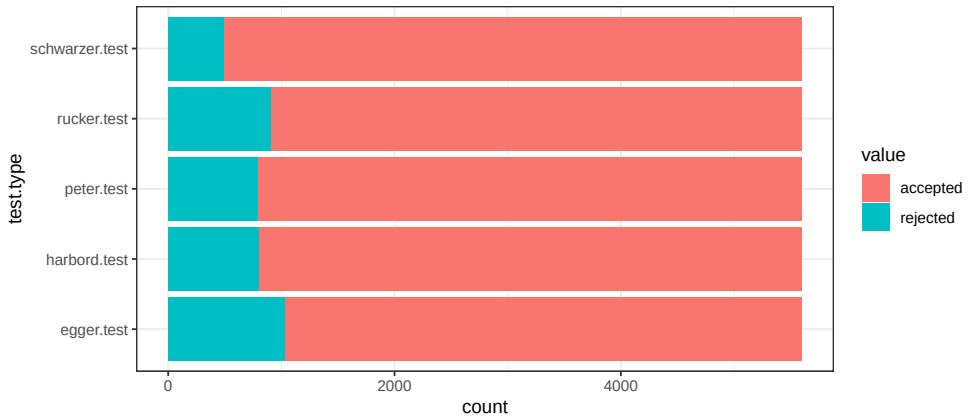

Figure 3.2: Proportion where the null hypothesis of no small study effect is rejected based on the 5% significance level for different tests (dichotomous outcomes).

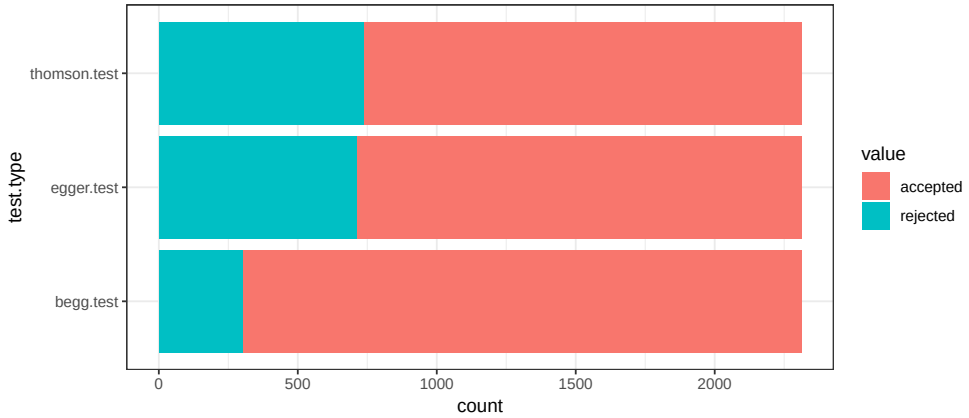

Figure 3.3: Proportion where the null hypothesis of no small study effect is rejected based on the 5% significance level for different tests (continuous outcomes).

same study collections as having publication bias. This is shown in Table 3.1 and Table 3.2, again separated for outcome types.

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

	Test Agreement	P-value Correlation
egger.schwarzer	0.84	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	0.67
harbord.peter	0.86	0.45

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

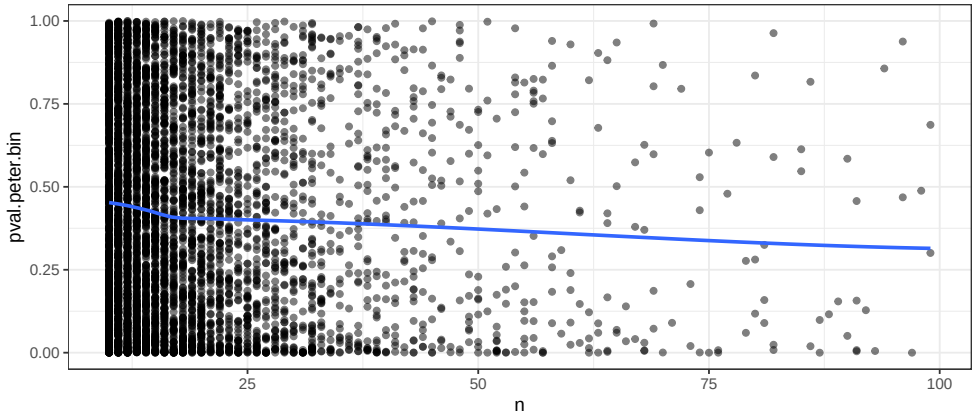

Figure 3.4: P-values of peters test for small study effects and their corresponding sample size (dichotomous outcomes)

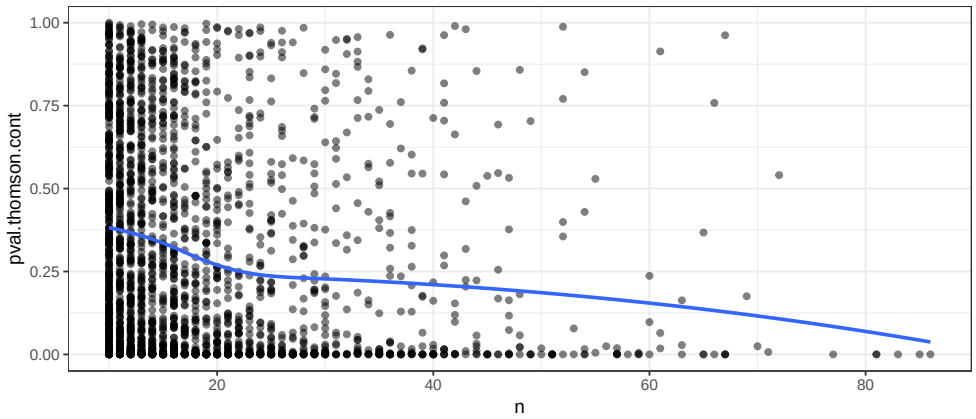

Figure 3.5: P-values of thomson and sharp's test for small study effects and their corresponding sample size (continuous outcomes)

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

	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 ar acceptance of the null hypothesis that there is no small study effect (continuous outcomes)

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.

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. A histogram with those fractions is shown in Figure 3.6 for continuous outcomes and 3.7 for dichotomous outcomes. The mean fraction of trimmed comparisons for binary outcomes is 0.19 and the median 0.17. 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

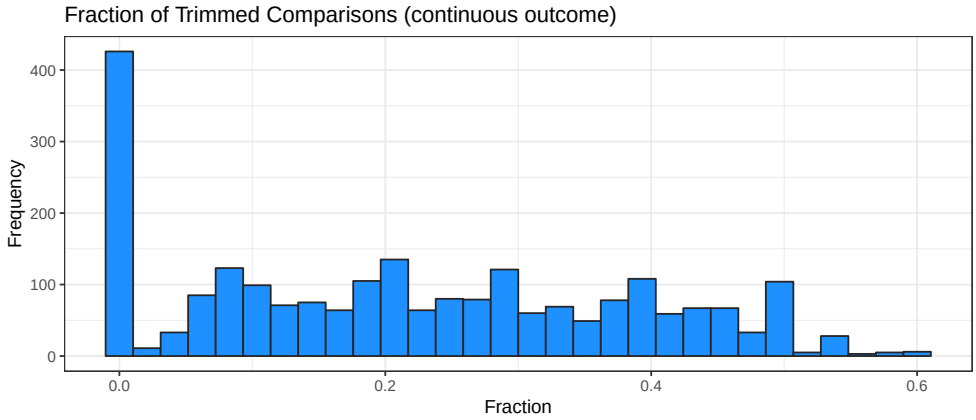

Figure 3.6: Histogram of fractions of trimmed comparisons from meta analyses with continuous outcomes.

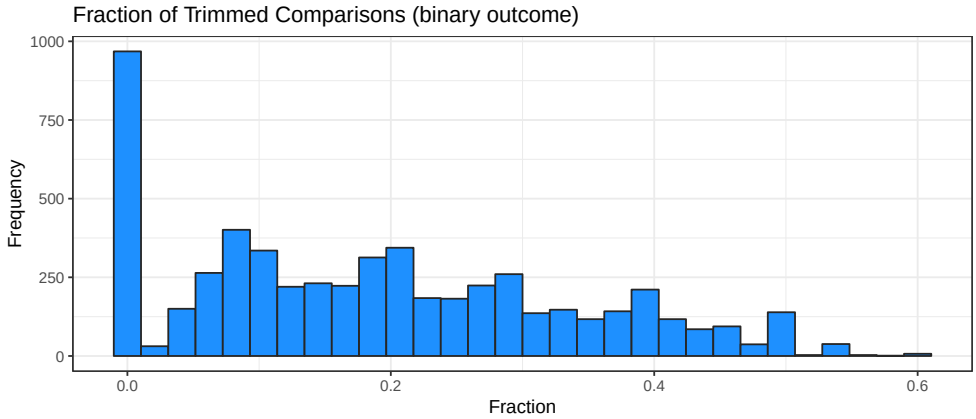

Figure 3.7: Histogram of fractions of trimmed comparisons from meta analyses with binary outcomes.

dichotomous outcomes.

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.

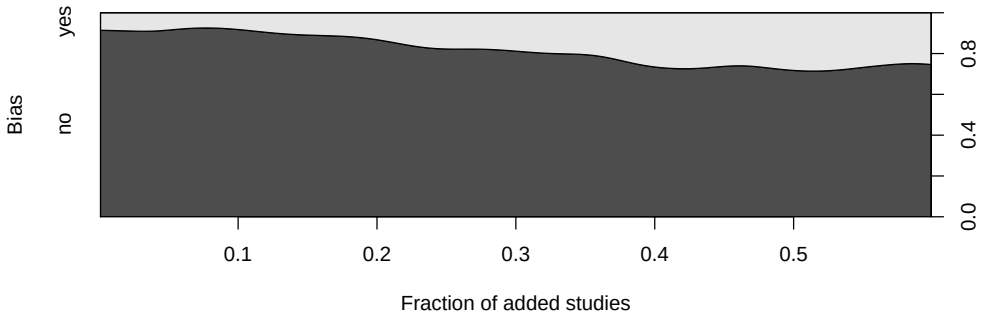

Figure 3.8: Fraction of added studies for small study effect correction by trim and fill and the corresponding small study effect test decision (based on 5% significance level) of peters test and dichotomous outcomes

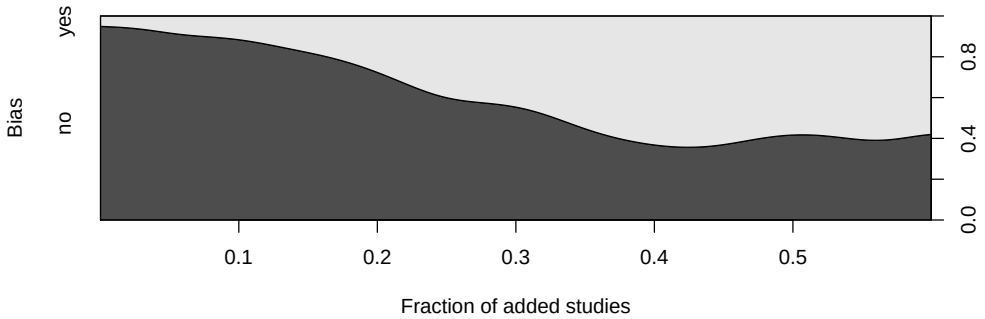

Figure 3.9: Fraction of added studies for small study effect correction by trim and fill and the corresponding small study effect test decision (based on 5% significance level) of thomson and sharp's test and continuous outcomes

Chapter 4

Discussion and Outlook

4.1 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. For convenience, notatotion will be identical whenever possible for both methods. 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 y_i be the effect size estimate of the study i and v_i be the corresponding variance of the effect size estimate. Effect measures have to be identical for all studies. Furthermore, let w_i be the inverse of the variance of the estimate from study i, $1/v_i$ and n the total number of studies.

The pooled estimate Δ_f of the fixed effects model is then simply the weighted average with the weights given by the inverses of the variances, w_i, given in 4.1. The variance ν_f is the reciprocal of the sum of the weights as in 4.2.

$$\Delta_f = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{4.1}$$

$$\nu_f = \frac{1}{\sum_{i=1}^{n} w_i} \tag{4.2}$$

The computation of random effects meta-analysis is more complicated. Random effects meta-analysis will give smaller studies with larger variance of their estimates more weight in the pooled estimate. The key principle is that the estimates are allowed to vary randomly around the true estimate Δ, and additionally, the estimates are subject to noise or sampling error themselves. Thus a within - and between study variance has to be computed. Let Q be the heterogeneity of study estimates as given in 4.3. The degrees of freedom d are equal to $n - 1$. C is defined in 4.4. The between study variance τ^2 is then defined as in **??**. The definition follows DerSimonian and Laird (1986), but other methods are available such as maximum likelihood or restricted maximum likelihood estimators.

$$Q = \sum_{i=1}^{n} w_i (y_i - \Delta_f)^2 \tag{4.3}$$

$$C = \sum_{i=1}^{n} w_i - \frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i} \tag{4.4}$$

$$\tau^2 = \max(0, \frac{Q - d}{C}) \tag{4.5}$$

The variance of a study estimate y_i of study i, $v_i^\star$ is then defined as in 4.6, with $w_i^\star$ being the inverse of $v_i^\star$. It is used to calculate a new kind of weighted mean to obtain a pooled estimate Δ_r as in 4.7. The variance of Δ_r, ν_r is then the sum of the reciprocal variances (4.8).

$$v_i^\star = v_i + \tau^2 \tag{4.6}$$

$$\Delta_r = \frac{\sum_{i=1}^{n} w_i^\star y_i}{\sum_{i=1}^{n} w_i^\star} \tag{4.7}$$

$$\nu_r = \frac{1}{\sum_{i=1}^{n} w_i^\star} \tag{4.8}$$

A p-value under the Null-hypothesis of $\Delta = 0$ can be obtained by calculating the Z-value (**??**) and the usual computation as in (4.10), Φ being the cumulative distribution function of a standard normal distribution.

$$Z = \frac{\Delta}{\sqrt{\nu}} \tag{4.9}$$

$$p = 2(1 - \Phi(|Z|) \tag{4.10}$$

4.2 Reporting Bias Test

Essentially, there are two kinds of tests for reporting bias, non-parametrical or rank-based tests or regression based tests. Five of them will be presented. First, tests for continuous outcome studies are described, and special modifications of those for binary outcomes will be introduced later.

Again, let y_i be the effect size estimate and v_i the variance estimate of study i. Then, the standardized effect size $y_i^\star$ is given in 4.12. Δ_f is the pooled estimate for fixed effect estimate defined in 4.1. Let $v_i^\star$ be the variance of $y_i - \Delta_f$ as defined in 4.11.

$$v_i^\star = v_i - 1/\sum_{i=1}^{n} v_i^- 1 \tag{4.11}$$

$$y_i^\star = (y_i - \Delta_f)/v_i^\star \tag{4.12}$$

Then, a rank correlation test based on Kendall's tau is used. The pairs of $y_i^\star$ and $v_i^\star$ 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 4.13. n is the number of studies.

$$Z = (u - l)/\sqrt{n(n-1)(2n+5)/18} \tag{4.13}$$

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

Alternatively, one can use Eggers test (Egger *et al.*, 1997) that is based on simple linear regression. Let $y_i^\star = y_i/\sqrt{v_i}$, $x_i = 1/\sqrt{v_i}$. Furthermore, let $\hat{y}_i = \beta_0 + \beta_1 x_i$. Using linear regression, one obtains a least-squares estimate for β_0, and the null hypothesis $\beta_0 = 0$ can be tested 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 given in 4.14.

$$p = 2 * (1 - t_{n-1}(\beta_0/se(\beta_0))) \tag{4.14}$$

A method proposed in Thompson and Sharp (1999) allows for between study heterogeneity. Let τ^2 be equal to 4.5. The effect size estimates are then assumed to be distributed as in 4.15. Then, a weighted regression is carried out with weigths $1/v_i^\star$ as given in 4.6. Analagous to Eggers test, β_0 is then tested with respect to the null hypothesis $\beta 0 = 0$.

$$y_i \sim N(\beta_0 + \beta 1 x_i, v_i + \tau^2) \tag{4.15}$$

A rank based alternative to Peters test for binary outcomes is the Harbords test (Harbord *et al.*, 2006). Let e_t be the number of events in a two-armed study i and n_t be the total number of patients in the treatment arm and the same for n_c and e_c for the control arm. Then 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 4.16 and 4.17.

$$r_i = e_t - (e_t - e_c)(e_t + (n_t - e_t))/(n_t + n_c) \tag{4.16}$$

$$v_i = \frac{(e_t + e_c)(e_t + (n_t - e_t))(e_c + (n_c - e_c))((n_t - e_t) + (n_c - e_c))}{(n_t + n_c)^2(n_t + n_c - 1)} \tag{4.17}$$

Similarly to Eggers or Peters Test, now a weighted linear regression can be performed on r_i/v_i depending on the standard error $1/\sqrt{v_i}$ with the $1/v_i$ used as weigths. r_i/v_i is also known as peto odds ratio.

4.3 Reporting Bias Adjustment

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 based on a funnel plot, on which the effect size estimates of studies are plotted against their standard error. If reporting bias is present, the estimate will shift in average towards higher or lower efffect sizes compared to the estimates of larger studies. Trim and Fill is a nonparametric approach.

The algorithm for the method tries to estimate the number of studies that are not available due to reporting bias, k. There are different estimators for k. First, Δ is estimated using a fixed or random effects model. Then, the k effect size estimates with the smallest standard errors are trimmed, and Δ 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 Δ, and Δ and its standard error is then computed to obtain an unbiased estimate.

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 Â£22.50. *Statistics in Medicine*, **7**, 817–818. 16

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. 15

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177 – 188. 15

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. 17

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. 16

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. 17

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, **18**, 2693–2708. 17