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 May 21, 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 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 quantiative 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 missleading 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 came 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 findigs 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.

```
## Error in mly.cont(data.ext, 0.05, min.study.number = 2):  could not find function
"mly.cont"
## Error in mly.bin(data.ext, 0.05, min.study.number = 2):  could not find function
"mly.bin"
## Warning in readChar(con, 5L, useBytes = TRUE): cannot open compressed file '/Users/p.gal
probable reason 'No such file or directory'
## Error in readChar(con, 5L, useBytes = TRUE): cannot open the connection
## Joining, by = c("n", "nn")
```

Chapter 2

The Cochrane Dataset

2.1 Cochrane Systematic Reviews

As has been mentioned before, the Cochrane Group has specialized on systematic reviews in clinical science. Certain knowledge of standards and principles of the Cochrane Group may help to assess the quality and the properties of the dataset. The following information stems from the Cochrane Handbook for Systematic Reviews (Higgins JPT, 2011).

The definition of a systematic review is that it "attempts to collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question." Thus, the "key properties of a review are":

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

At the end of a systematic review, "a systematic presentation, and synthesis, of the characteristics and findings of the included studies" is done.

53 Cochrane Review Groups prepare and maintain the reviews within specific areas of health care. A group consists of "researchers, healthcare professionals and people using healthcare services (consumers)".

The groups are supported by Method Groups, Centres and Fields. The Cochrane Method Groups aim to discuss and consult the groups in methodological questions concerning review preparation. The Centres 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 protocoll, specifying the research question, the methods to be used in literature search and analysis and the eligibility criteria of the study. Changes in protocolls 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 affec the outcome of the research study".

In order to avoid potential conflicts of interests, there is a code of conduct that all entities of the Cochrane Organization have to agree on: conflicts of interest must be disclosed and possibly be forwarded to the Cochrane Centre, 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 cicle 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 informations 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.

2.1.1 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, ususally the eligibility criteria follow. Usually, outcomes are not part of eligibility criteria, except for special cases such as adverse effect reviews.

The type of study is an important eligibility criterium. The Cochrane Collaboration focuses "primarily on randomized controlled trials", and also, the methods of study identification in literature search are focused on randomized trials. Furthermore, study characteristics such as blinding of study operators with respect to treatment and cluster-randomizing might be additional eligibility criteria which have to be chosen by the review authors.

After having specified the eligibility criteria, studies have to be collected. The central idea of systematic reviews, and also meta-analyses, is that the collected studies are a random sample of a population of studies, i.e. that they are representative and can be used to assess population properties. Therefore, the search process is crucial, as a selective search result may impose bias on the sample of studies available, making it a non-random sample. For this purpose, the Cochrane Groups are advised to go beyond MEDLINE !!cite!!, because a search restricted to it has been shown to deliver only 30% to 80% of available studies. "Time and budget restraints require the review author to balance the thoroughness of the search with efficiency in use of time and funds and the best way of achieving this balance is to be aware of, and try to minimize, the biases such as publication bias and language bias that can result from restricting searches in different ways." It is important to note that not only studies, but also study reports are occasionally used in the reviews, as they may provide useful information.

There are different sources that are being used to search for studies.

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

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

- EMBASE. 4,800 Journals publishing in 30 languages are indexed to EMBASE, which includes more than 11 million records from 1974 onwards. 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 databse.

- Regional or national and subject specific databases can additionally be consulted and often provide important information. Financial considerations may limit the use of such databases.

- General search engines such as Google Scholar, Intute and Turning Research into Practice (TRIP) database can be used.

- Citation Indexes. The database lists articles published in around 6,000 Journals with articles in which they have been cited and is available online as SciSearch. This form of search is known as cited reference searching.

- Dissertation sources. Dissertations are often listed in MEDLINE or EMBASE but one is advised to also search in specific dissertation sources.

- Grey Literature Databases. Approximately 10% of the results in the Cochrane Database stems from conference abstracts and other grey literature. The Institute for Scientific and Technical Information in France provides access to entries of the previously closed System for Information on Grey Literature database of the European Association for Grey Literature Exploitation). Another source is the Healthcare Management Information Consortium (HMIC) database containing records from the Library and Information Services department of the Department of Health (DH) in England and the King's Fund Information and Library Service. The National Technical Information Service (NTIS) gives access to the results of US and non-US government-sponsored research, as well as technical report for most published results. References from newsletters, magazines and technical and annual reports in behavioral science, psychology and health are provided in the PsycEXTRA database which is linked to PsycINFO database.

2.1.2 Structure and Content

The dataset consists of 5016 systematic reviews from the Cochrane Library with 52995 studies and 463820 results. A result compares clinical or medical interventions or treatments. Each study provides (multiple) results of clinical interventions.

In Table 2.1, two results from a systematic review about effects of barbiturates are shown as they are given in the dataset. As can be seen, the result is further specified by the variables in the columns. The comparison variable specifies what treatments or interventions are compared, the outcome variable how it is compared, and the subgroup variable (not given in table) indicates if the result belongs to a certain subgroup. Here, the result is of a binary outcome, so the events in the barbiturate treatment group and the total number of participants are given in columns "Events" and "Total" and the number of events in the control group "Events_c" and participants "Total_c". Events denotes here number of deaths at the end of follow up.

Study	Comparison	Outcome	Events	Total	Events_c	Total_c
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 2.1: Example of two results as given in the dataset. Events denotes the count of events in the treatment group while Events c the count of events in the group compared to. Further descriptive variables have been ommitted

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

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

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.
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
outcome.name/.nr	Specification by which outcome the interventions are compared and a unique number for the outcome
subgroup.name/.nr	Potentially indication of affiliation to subgroups and a unique number for the subgroup
outcome.measure	Indication of the quantification method of the effect (of one intervention compared to the other).
effect	Measure of the effect given in the quantity denoted by "outcome measure".
se	Standard error of the measure of the effect,
events1/events2	The counts of patients with an outcome *if* measurement/outcome is binary or dichotomous 2 (1 for treatment group and 2 for control group).
total1/total2	Number of patients in groups.
mean1/mean2	Mean of patient measurements *if* outcome is continuous.
sd1/sd2	Standard deviation of mean *if* outcome is continuous.

Table 2.2: Dataset variable names and descriptions

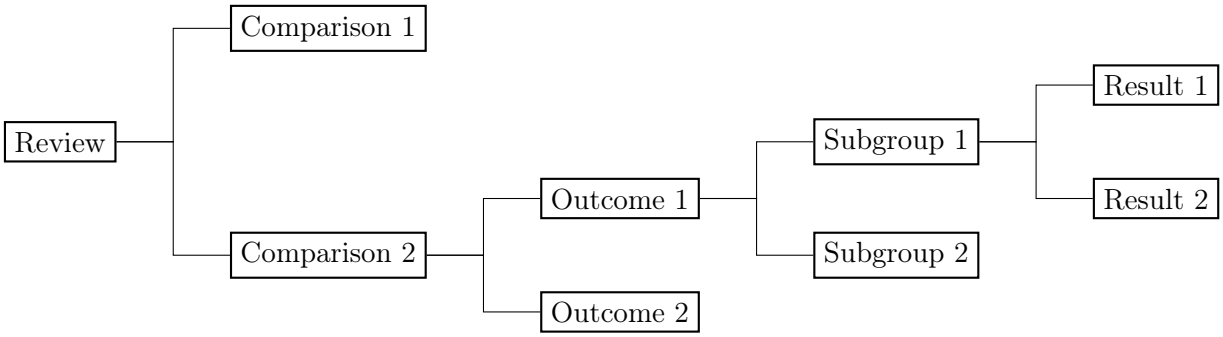

Figure 2.1: Structure of a hypothetical review with two different comparisons

The structure of a review will now be outlined based on an example of the dataset. Lets consider the previously mentioned barbiturate and head injury review. The aim was to "assess the effects of barbiturates in reducing mortality, disability and raised ICP (intra-cranial pressure) in people with acute traumatic brain injury" as well as to "quantify any side effects resulting from the use of barbiturates". Since there are arguments for and against use of barbiturates, the authors of the review did a comprehensive literature search and collected all available study

findings.

The review comprises five studies in total. Three of them compared barbiturate to placebo, one compared barbiturate to Mannitol and one Pentobarbital to Thiopental, which would be the comparison to speak in the previously introduced notion. The studies have different outcomes, for example, death or death and severe disability at follow up. One study split up outcomes for patients with and without haematoma, which would be subgroups.

The complete listing of outcomes is in table 2.3. The table also gives an illustration of the variety of data that can be included in a review. We have for example continuous (mean body temperature) and binary outcome data (death). Additionally to primary and secondary outcomes, often also dropouts and adverse effects are included in a review.

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 and outcome measure of the results are given

It is important not to confuse results with studies. A study can contribute multiple results to a systematic review, for example, primary and secondary outcomes and adverse effects.

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. In the case of continuous outcomes, the cases have been counted were effect sizes and standard errors of effect sizes are not available. This means that neither mean values of groups nor mean differences are given, or that neither standard deviation nor standard error is given for a result. Study years before 1920 and after 2019 have been counted as missing, as well as sample sizes below zero.

Missing mean values and mean differences	984
Missing standard deviations and standard errors	1300
Missing sample sizes	12173
Missing study year	44649

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

The studies that are included in the reviews and have been published are most often from the years after 1980 (5% quantile = 1982). The median of the publication years is 2003, the mean 2006.13 and the quartiles are 1996 and 2008. Only a handful ($n = 18$) have been published in 2018, none in 2019. No information is available concerning unpublished results and studies.

The results of a study are summarised in a result by a effect measure. This effect measure varies, depending on whether data is continuous, binary or time-to-event. The most abundant

effect measures used are summarised in Table 2.5. One can conclude of the table that roughly 30 % of outcomes in the dataset are continuous and the being some sort of discrete or binary outcomes, most often binary ($> 65\%$).

Outcome measure	n	Percentage
Risk Ratio	232583	50.1%
Mean Difference	102315	22.1%
Odds Ratio	49372	10.6%
Std. Mean Difference	40535	8.7%
Peto Odds Ratio	19122	4.1%
Hazard Ratio	6566	1.4%
Risk Difference	6234	1.3%
Rate Ratio	2283	0.5%
other	4810	1%

Table 2.5: 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,

The sample sizes amongst results vary to some extent. There are 5% of treatment group sample sizes that are smaller tha 8, the 5% quantile. The first quartile is 22, the median 48, the mean 302.03 and the third quartile 119. 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 = 15, first quartile = 44, median = 94 and third quartile = 229. The mean is 617.81.

The mean and median number of results per review are 12.42 and 7. There are 417 reviews with five or fewer results, and the quartiles are 16 and 102. Similarly, the number of reviews with a maximum of two studies included is 836, the mean study number is 12.42, the median 7 and the interquartile range 4 and 15. The discrepancy between mean and median is again due to large reviews with a high number of studies and results, most extreme in **?** which is a systematic review about antibiotic prophylaxisfor preventing infection after cesarean section, with 95 studies and 1497 results in total.

For results to be suitable for usage in meta-analysis, they have to be identical with respect to comparison and outcome. More speicifically, the studies in the dataset that have the same comparison, outcome and subgroup can be pooled in a meta-analysis since their research subject and experimental setup can be considered sufficiently homogeneous. The amount of studies that can be pooled for meta-analyses is of special interest in this masters thesis. Thus, the dataset is divided in groups whith identical experimental setup. The size of the group denotes how many results are included in a group.

Table 2.6 shows the number of *groups* of groups with equal or more than n results. Practically, this means that a given number of meta analyses can be performed with each having at least n results.

n	Number of groups	Cumulative sum of groups
1	102344	186300
2	31686	83956
3	16072	52270
4	9628	36198
5	6444	26570
6	4230	20126
7	2961	15896
8	2114	12935
9	1592	10821
10	1238	9229
11	921	7991
12	702	7070
13	585	6368
14	455	5783
15	5328	5328

Table 2.6: Cumulative number of groups with number of reproduction trials >= n

Chapter 3

Results

3.1 Small study effects

To provide an overview over the abundance of small study effects in the dataset, first it is shown how median absolute effect size decreases with increasing sample size of the results (Figure 3.1).

A clear trend of decrease of absolute effect size with increasing sample size (i.e. smaller variance) is visible. 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 effects.

The median absolute normalized effect size can be visualized for the different outcome measures separately (Figure 3.2). The plot confirms the trend of median effect size decrease towards lower sample size (e.g. for Risk Ratios: decrease > 0.1 standard deviations from 10 to 50 study participants). Instead of normalizing the effects for all effect sizes, the effects are normalized with respect to the effects of the same outcome measures.

3.1.1 Small Study Effect Tests

There are tests that can be applied to find out if small study effects are 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 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 3.3 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 ($n = 1383$) The same is shown in

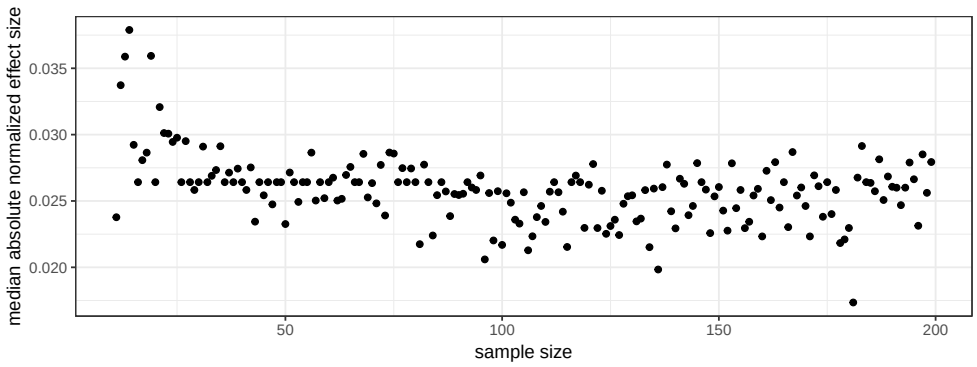

Figure 3.1: Median of the absolute value of the normalized effect size plotted against the total sample size.

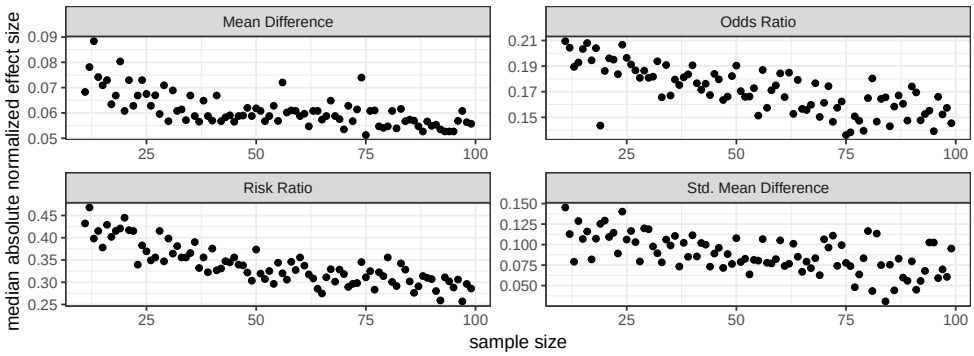

Figure 3.2: Median of the absolute value of the normalized effect size plotted against the total sample size, separated for outcome measures.

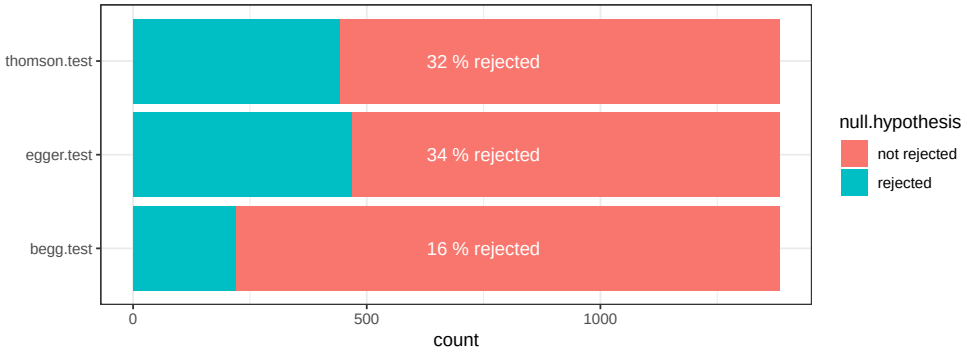

Figure 3.3: Proportion where the null hypothesis of no small study effect is rejected based on the 5% significance level for different tests (continuous outcomes).

Figure 3.4 for dichotomous outcome measures ($n = 3955$). The amount of studies varies from 5% (Schwarzer's Test) to 13 % (Egger's Test) for binary outcomes and 9% (Begg and Mazumdar's Test) to 25 % (Egger's Test) for continuous outcomes.

There is no substantive change in the fraction of positive test results, depending on if the pooled treatment effect size estimate is significant or not. Most substantively, the fraction of positive Egger Test results increases for significant pooled treatment effects by 0.1825716 (the minimal increase is -0.0032246 for Schwarzer's Test).

Furthermore one can look if the frequencies of tests that reject the null hypotheses change over time (mean publication year of the studies included in the meta analyses). The proportion

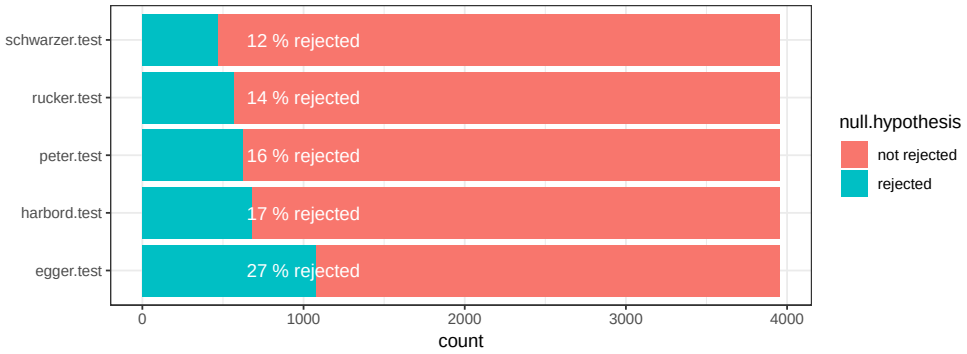

Figure 3.4: Proportion where the null hypothesis of no small study effect is rejected based on the 5% significance level for different tests (dichotomous outcomes).

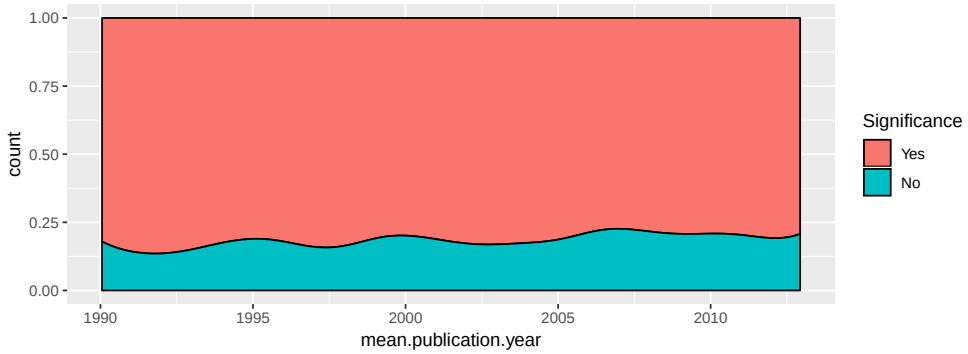

Figure 3.5: Proportion of test results where the null hypothesis of no small study effect is rejected over time (mean publication.year).

of the test results are shown in Figure **??**. The Figure suggests that the frequency of publication bias remains constant over time. The mean publication years have been restricted such that at least 180 meta-analyses are available per year, such that random fluctuation is restricted to some extent. The significance threshold for the p-values used is 0.05, and the small study effect test used is Thomson's test (with the arcsine variance stabilizing transformation function used in the case of binary outcomes).

The agreement of the tests, i.e. the proportion of meta-analyses where the rest 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 Agreement	P-value Correlation
egger.schwarzer	0.73	0.23
egger.peter	0.75	0.37
egger.rucker	0.75	0.34
egger.harbord	0.79	0.51
schwarzer.peter	0.81	0.26
schwarzer.rucker	0.81	0.25
schwarzer.harbord	0.86	0.47
rucker.peter	0.87	0.64
harbord.peter	0.84	0.50

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)

	Test Agreement	P-value Correlation
thomson.egger	0.85	0.69
thomson.begg	0.77	0.47
egger.begg	0.73	0.36

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 the Thompson and Sharp tests are shown with respect to the sample size of the meta-analysis in Figure 3.6. In the case of binary outcomes, the arcsine variance stabilizing function has been applied prior to use of Thompson and Sharp's test. A trend towards more rejections for larger sample sizes can be seen.

One can use the proportion of added studies by the trim-and-fill method from the overall

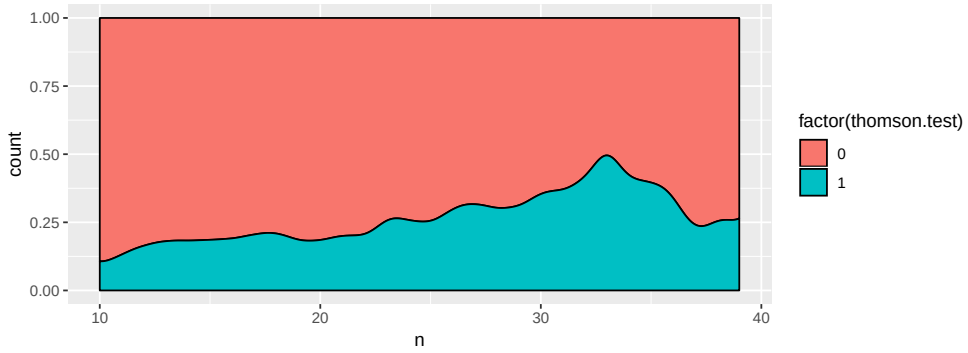

Figure 3.6: P-values of Thomson and Sharp's test for small study effects and their corresponding sample size.

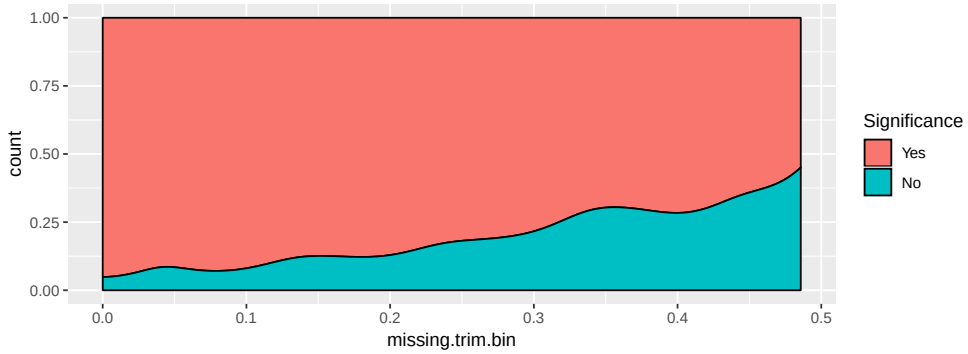

Figure 3.7: Fraction of added studies for small study effect correction by trim and fill and the corresponding small study effect test decision (based on 5% significance level) of Peters test and dichotomous outcomes

number of studies to further investigate the extent of small study effects. The mean fraction of trimmed comparisons for binary outcomes is 0.19 and the median 0.18. In Figure 3.7 and Figure 3.8, 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 dor dichotomous outcomes, there is less agreement with the trim-and fill method than in the case of Thomson and Sharp's test for continuos outcomes in the sense that the fraction of meta-analyses with rejected null hypothesises increases more clearly when there are more studies added by trim-and-fill.

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. Three of them will be applied to the meta-analyses shown previously that have ten or more study results and are therefore eligible for testing for publication bias.

The extent to what the results of the meta-analysis results are changed can be investigated. Because statistical significance is often used to decide if there is a treatment effect, a non-significant corrected effect size estimate can indicate that an observed treatment effect has been accepted because of small study effects. Therefore, the cases have been counted in which

1. Significance or non-significance of pooled estimate of meta-analysis did not change after correction for small study effects.

2. Significance of pooled estimate of meta-analysis did change to non-significance after correction for small study effects.

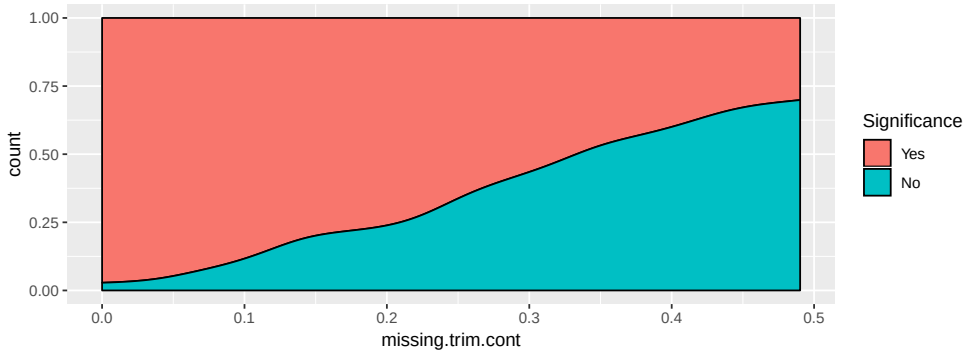

Figure 3.8: Fraction of added studies for small study effect correction by trim and fill and the corresponding small study effect test decision (based on 5% significance level) of Thomson and Sharp's test and continuous outcomes

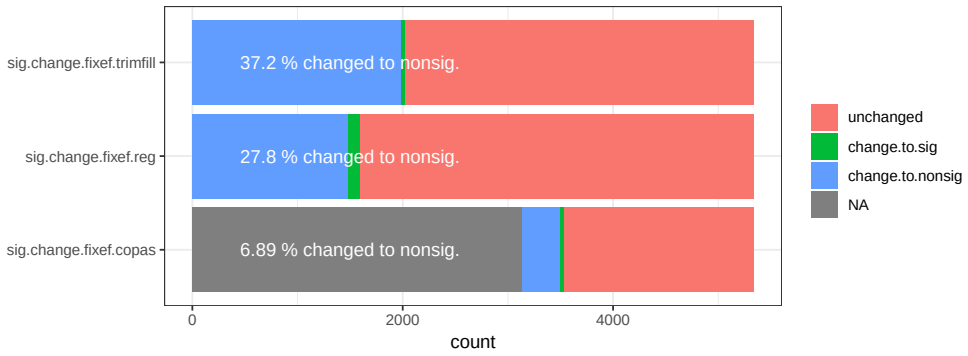

Figure 3.9: Change in signficance of fixed effects meta-analysis pooled estimate after correction.

3. Non-significance of pooled estimate of meta-analysis did change to significance after correction for small study effects.

The results of this can be seen in Figure 3.9 for all three methods, comparing the significance of the corrected pooled effect size estimate with the significance of the pooled effect size estimate of the fixed effects meta-analysis. The same for significance of random effects meta-analysis is shown in Figure 3.13. The significance threshold was chosen such that the p-value had to be < 0.05 for rejection of the null hypothesis of no treatment effect. The correction methods were trim-and-fill, copas selection model and regression with random effects and shrinkage of within-study-variance methods. More details to the applied correction methods and their application are in the methods section **??**. Notably, the correction methods has been applied to all meta-analyses, thus also for such that had no significant small study effect test result.

Since it has been previously seen in Figure **??** that the results of small study effects vary considerably between continuous outcomes, the results in significance change from fixed effects meta-analysis can be seen separately in Figure 3.11 for continuous and binary outcomes. The change in significance from random effects meta-analysis to significance of corrected estimate can be seen in Figure 3.11

Because the real amount of publication bias in the dataset is not known, the correction method can also be applied only to meta-analyses that have publication bias according to the small study effect tests in the previous section. Because the test developed by **?** has been applied to both binary and continuous outcome meta-analyses (in the case of binary outcomes to arcsine variance stabilized proportions), it is used as a criterium to distinguish biased from unbiased meta-analyses. The proportions of significance tests of pooled treatment effects that turned

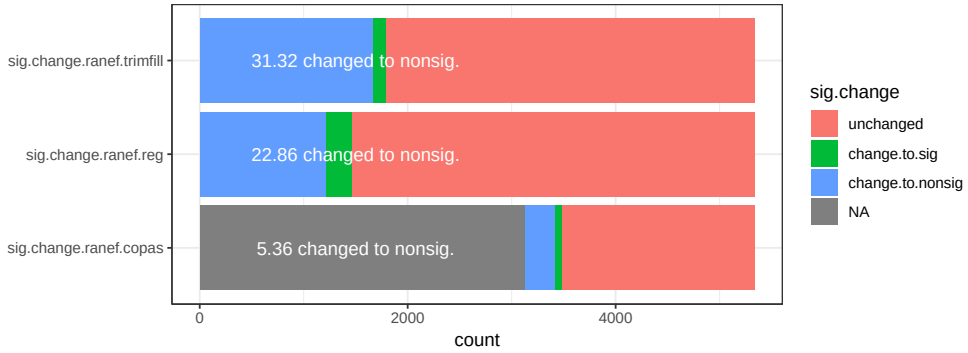

Figure 3.10: Change in signficance of random effects meta-analysis pooled estimate after correction.

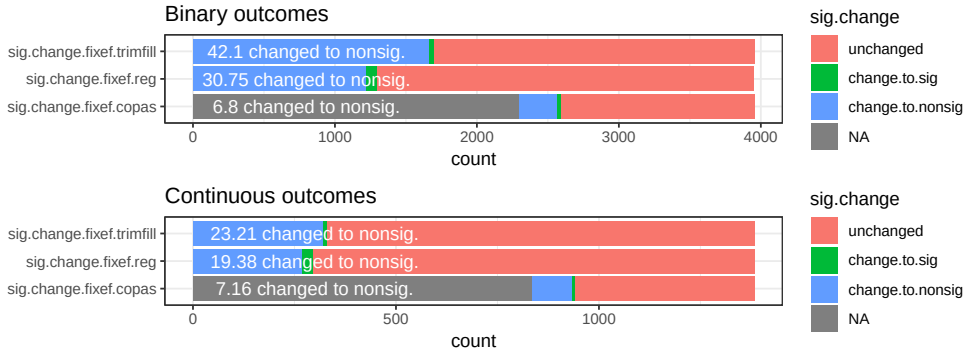

Figure 3.11: Change in signficance of fixed effects meta-analysis pooled estimate after correction, separated for continuous and binary outcomes.

from significant to non-significant, non-significant to significant etc. are shown in Figure 3.11. Fixed effects meta-analysis has been used to determine signficance of the uncorrected estimate.

Similarly, the number of missing studies per meta-analysis, i.e. those which have not been included because of small study effects, are estimated by the copas and trim-and-fill method and their empirical distribution is shown in histograms in Figure 3.14. For visualisation, the fraction of unpublished studies from the total fraction of available studies is shown.

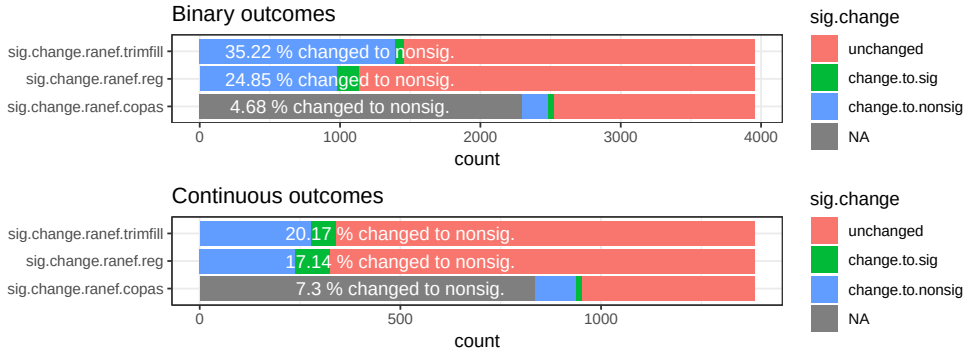

Figure 3.12: Change in signficance of random effects meta-analysis pooled estimate after correction, separated for continuous and binary outcomes.

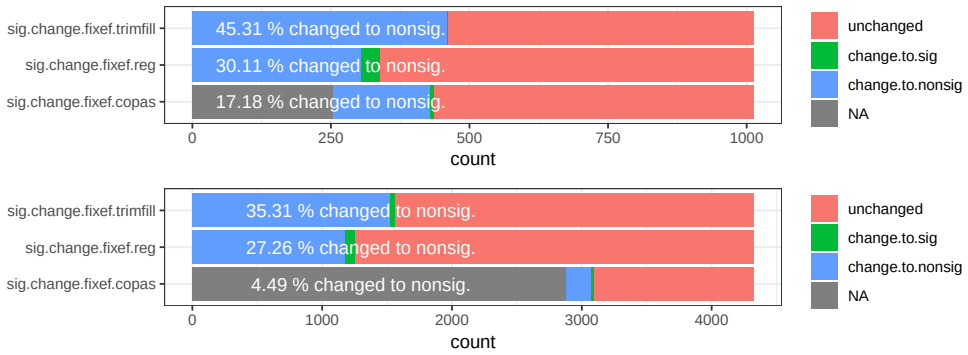

Figure 3.13: Change in signficance of random effects meta-analysis pooled estimate after correction, separated for continuous and binary outcomes.

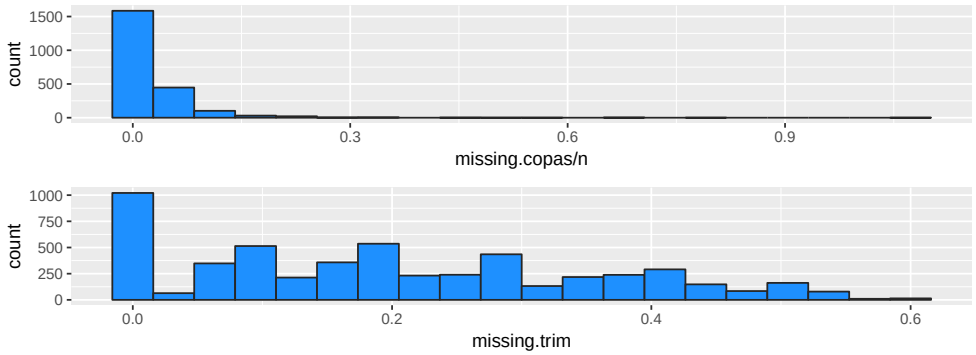

Figure 3.14: Fraction of missing studies estimated of the number of total studies included in the meta-analysis for copas selection and trim-and-fill method.

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 Δ usually denotes the summarized or pooled effect estimate of the meta-analysis, and η 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 Transformation between effect sizes

Binary and continuous outcome measures can both be transformed into correlations and fishers z-scaled correlations in order to be compared.

4.3 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). Let θ be a log odds ratio. A standardized mean difference d (also known as Cohen's d) is calculated by multiplying θ with $\sqrt{3}/\pi$.
We get from a standardized mean difference d to a correlation r by using

$$r = \frac{d}{\sqrt{d^2 + a}}$$

where a is a correction factor if $n_t \neq n_c, a = (n_c + n_t)^2/n_c n_t$

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 Δ 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

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

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

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, τ^2, the variance of true effects, can be calculated. τ^2 is on the same scale as the effect size and reflects the absolute amount of dispersion. In practice, τ^2 can be smaller than zero, then it is set to zero.

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

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

The estimation method for τ^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

$$I^2 = (Q-n+1)/Q \tag{4.5}$$

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

4.4 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 Δ_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 η_f is the reciprocal of the sum of the weights as shown in 4.7.

$$\Delta_f = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{4.6}$$

$$\eta_f = \frac{1}{\sum_{i=1}^{n} w_i} \tag{4.7}$$

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 Δ, and additionally, the estimates are subject to noise or sampling error themselves.

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

$$v_i^\star = v_i + \tau^2 \tag{4.8}$$

$$\Delta_r = \frac{\sum_{i=1}^{n} w_i^\star y_i}{\sum_{i=1}^{n} w_i^\star} \tag{4.9}$$

$$\eta_r = \frac{1}{\sum_{i=1}^{n} w_i^\star} \tag{4.10}$$

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), Φ being the distribution function of a standard normal distribution.

$$Z = \frac{\Delta}{\sqrt{\nu}} \tag{4.11}$$

$$p = 2(1 - \Phi(|Z|) \tag{4.12}$$

4.5 Small Study Effects Tests

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.

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

$$y_i^\star = (y_i - \Delta_f)/v_i^\star \tag{4.13}$$

$$v_i^\star = v_i - 1/\sum_{i=1}^{n} v_i^{-}1 \tag{4.14}$$

A rank correlation test based on Kendall's tau is then used. The pairs $(y_i^\star, v_i^\star)$ 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 4.15.

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

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 $y_i^\star = y_i/\sqrt{v_i}$ and $x_i = 1/\sqrt{v_i}$. Using $y_i^\star$ as dependent, and x_i as explanatory variable in linear regression, one obtains an intercept β_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.

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

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 τ^2 be equal to 4.4. The effect size estimates are then assumed to be distributed as in 4.17.

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

Then, a weighted regression is carried out with weights $1/v_i^\star$ based on the inverse of the variance as in 4.8. Analogous to Egger's test, β_0 is then tested with respect to the null hypothesis $\beta_0 = 0$.

4.5.2 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

$$y_i = \log(e_t * (n_c - e_c)/e_c * (n_t - e_t)) \tag{4.18}$$
$$v_i = 1/(e_t + (n_t - e_t) + 1/(e_c + (n_c - e_c) \tag{4.19}$$

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 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.

$$r_i = e_t - (e_t - e_c)(e_t + (n_t - e_t))/(n_t + n_c) \tag{4.20}$$

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

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

$$y_i = \arcsin e_t/n_t - \arcsin e_c/n_c \tag{4.22}$$

$$v_i = 1/4n_t + +/4n_c \tag{4.23}$$

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

4.6 Small Study Effect Adjustment

4.6.1 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, 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.

4.6.2 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

$$y_i \sim N(\mu_i, \sigma_i^2) \tag{4.24}$$
$$\mu_i \sim N(\mu, \tau^2) \tag{4.25}$$

corresponding to a standard random effects meta-analysis. μ is the overall mean effect, σ_i^2 the within study variance and τ^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 σ). Only a proportion

$$P(\text{select}|s) = \Phi(a + b/s) \tag{4.26}$$

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. 4.26 can be rewritten as

$$z = a + b/s + \delta \tag{4.27}$$

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

$$y_i = \mu_i + \sigma_i \epsilon_i \tag{4.28}$$
$$\mu_i \sim N(\mu, \tau^2) \tag{4.29}$$
$$z_i = a + b/s_i + \delta_i \tag{4.30}$$

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

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

$$P(\text{select}|s, y) = P(a, b, s, y) = \Phi\left(\frac{u + \tilde{\rho}((y - \mu)/\sqrt{(\tau^2 + \sigma^2)})}{\sqrt{1 - \tilde{\rho^2}}}\right) \tag{4.31}$$

It can also be shown that the expected value

$$\mathbb{E}(y|s, select) = \mu + \rho\sigma\lambda(u) \tag{4.32}$$

which shows that the expected value for a study is larger for larger σ.

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 μ 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 4.32 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 :

$$y_i = \mu_i + \beta s_i + \sigma_i \epsilon_i \tag{4.33}$$

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 β 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 μ that have p-values over a predefined significance threshold can be used for inference of the effect size.

4.6.3 Adjustment by Regression

There are multiple ways to adjust for small study effects by regression. The general idea is to extrapolate 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 β_0 and on its standard error $\sqrt{v_i}$ as in 4.37.

$$y_i = \beta_0 + \beta_1(\sqrt{v_i + \tau^2}) + \epsilon_i(\sqrt{v_i + \tau^2}), \epsilon_i \overset{\text{iid}}{\sim} N(0,1) \tag{4.34}$$

β_1 represents the bias introduced by small study effects, as can be seen when looking at 4.35

$$\mathbb{E}((y_i - \beta_0)/\sqrt{v_i}) \to \beta_1 \text{ if } \sqrt{v_i} \to \infty \tag{4.35}$$
$$\mathbb{E}(y_i) \to \beta_0 + \beta_1 \tau \text{if} \sqrt{v_i} \to 0 \tag{4.36}$$

After estimating τ^2, one can estimate β_0 and β_1 as seen before e.g. in Thompson and Sharp's Test with weights also equal to Thompson and Sharp's Test (see 4.5.1).

To diminuish the random variation within studies, but keep the variation between studies, we change 4.37 to a scenario where each study has M-fold increased precision:

$$y_{M,i} = \beta_0^\star + \beta_1^\star(\sqrt{v_i/M + \tau^2}) + \epsilon_i(\sqrt{v_i/M + \tau^2}) \tag{4.37}$$

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

$$y_{\infty,i} = \beta_0^\star + \tau(\beta_1^\star + \epsilon_i), \epsilon_i \overset{\text{iid}}{\sim} N(0,1) \tag{4.38}$$

Note that:

$$y_{\infty,i} = \beta_0^\star + \tau\beta_1^\star = \beta_0 \tag{4.39}$$

β_0 is termed the limit meta analysis expectation. Now, the random errors from 4.37 are rewritten as:

$$\epsilon_i = \frac{y_i - \beta_0^\star}{\sqrt{v_i + \tau^2}} - \beta_1^\star \tag{4.40}$$

Assuming ϵ_i to be fixed, we can plug it into 4.38 and get

$$y_{\infty,i} = \beta_0^\star + \sqrt{\frac{\tau^2}{v_i + \tau^2}}(y_i - \beta_0^\star) \tag{4.41}$$

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

Chapter 5

Discussion

It is tempting to claim to have a perspective that spans the entire clinical research when looking at the cochrane dataset. Several points can be made to argue in this favor (see chapter 2 CITE for details).

- Size: Over 52,000 studies, 20,000 comparisons and 47 millions study participants. A vast diversity of topics and studies that stem from different institutions and researchers and have been conducted in different countries.
- Quality: An organization of field and analysis experts controls quality of both studies and methods.

The dataset is certainly unique with respect to this features. The effort of the Cochrane Organization in reprocession, classification and integration of studies in a larger scientific framework is unmatched, not only in clinical science, but in empirical science in general CITE?CITE. Regardless whether it is representative of the field of clinical research, it thus provides an exceptional perspective, not only on clinical science, but on empirical science in general.

Concerning the generalizability of results of this thesis on clinical science, there are important caveats. First of all, intrinisically, the Cochrane Organization focuses on established research of public interest. By definition, it can not comprise the full diversity of the field, and especially where research and evidence is new, sparse and not established, it is almost certainly not included. The aim of Cochrane is to concentrate scientifical knowledge, thus it can not incorporate studies that are unique in their research subjects and are on the "periphery" of science.

It is said in the Cochrane Handbook that Cochrane Field groups orchestrate the efforts of their groups and set emphasis on certain research. It is thus clear, that the Cochrane Library only comprises a selective fraction of clinical research.

Although Cochrane strives to include unpublished study results, it almost certainly provides a selective view by including more published studies. Apart from language issues, publication bias might be the largest source of bias. This leads us to the main topic of this masters thesis: small study effects and publication bias in the Cochrane Library. Thus, the research question of this thesis is in fact connected to the question of how representative the Cochrane Library is for clinical research. The answer to this question is insofar important, as that the issues treated in this thesis have to be addressed to the Cochrane Organization or to clinical science in general. This distinction is very important and will be encountered throughout the next pages.

5.1 Meta-analyses

If one investigates the success of a treatment, the first way of doing so is to investigate to what extent the treatment effect estimate differs from no treatment effect. This is usually done by a hypothesis test, where the null hypothesis of no treatment effect is rejected if the p-value of

the treatment effect estimate is above a given threshold. One decides in this case in favor of the treatment.

The aim of a meta-analysis is to re-calculate the treatment effect based on multiple results. The treatment effects are pooled and the pooled treatment effect estimate can then again be tested based on the null-hypothesis. Although not commonly accepted, this may be a stronger argument for the success of treatment, because meta-analysis can capture consistency among treatment effect estimates among different studies and will result in rejection of a null hypothesis if consistency is given. It is thus in line with the empirical principle that experimental findings should be validated based on their reproducibility. Consequently, a meta-analysis may provide a significant treatment effect estimate, even if the single results do not.

It was the aim of the first section CITE of the results part to compare significance of the treatment effect of single studies with the significance of the pooled effect. So the question to be answered was: How does significance in single studies relate to significance when all evidence is included? The first kind of significance has been termed primary significance and the second, secondary significance. One can imagine 3 scenarios:

- A primary significant study belongs to a group which, overall, can not reject the null-hypothesis of no treatment effect in meta-analysis. The study result is then overruled by the result of the meta-analysis (if one is to accept that meta-analysis is a more stringent rule for decision about treatment effects).
- The opposite: A non-significant primary study belongs to a group which has a significant pooled treatment effect estimate. The study failed to find evidence for a "true" effect.
- Primary and secondary significance remain unchanged after meta-analysis.

The amount of overlap between primary and secondary significance was moderate, as it had to be expected based on the methodology of meta-analysis. The vast amount of non-significant primary results that contribute in the end to a significant secondary meta-analysis result is compelling. This links to the topic of the masters thesis of small study effects and publication bias. One ad-hoc interpretation of publication bias is that results that are not significant are less likely to be published (arguably because they show no treatment effect).

If one is to prefer, or accept meta-analysis as a way to assess treatment effect evidence, the result provides a strong argument for publication of non-significant results, because they can contribute to evidence for treatment effects in meta-analysis (in a Cochrane review). Significant treatment effects that are found may find their into clinical practice and potentially benefit there patients and consumers.

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. 21

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. 20

Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis . *Biostatistics*, **1**, 247–262. 23

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. 23

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. 23

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. 23

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. 22

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. 22

Higgins JPT, G. S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration. 3, 11

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. 22

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

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. 23

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, **18**, 2693–2708. 22