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 June 17, 2019

Contents

Preface **iii**

1 Introduction **1**

2 The Cochrane Dataset **3**
2.1 Cochrane Systematic Reviews 3
2.2 Data Tidying and Processing 9

3 Results **13**
3.1 Small study effects 13
3.2 Small Study Effect Tests 13

4 Methods **19**
4.1 Introduction and Notation 19
4.2 Effect Measures and p-values 19
4.3 Fixed and Random Effects Meta-Analysis 21
4.4 Small Study Effects Tests 22
4.5 Small study effect and Publication Bias Adjustment 26
4.6 Transformation between effect sizes 29

A Appendix **31**

Bibliography **33**

Preface

Howdy!

Max Muster
June 2018

Chapter 1

Introduction

Typically, we expect empirical scientific results to be distorted by random noise, such that the results are not equal to the real process that they describe. Additionally, one can reasonably assume that the results are also biased to some extent, as the statistical paradigm of precision versus bias would predict: The smaller the noise part of distortion, the more the result should be distorted by bias. However, the trade-off is not inevitable and can be minimized by scientific rigor and foresightfull experimental design. Large efforts have been made both in theory and practice to improve the quality of experiments and their analysis, such as defining experimental settings that prohibit influence of expectations of the researcher (e.g. blinding) or increase sample size in experiments. The introduction of randomized clinical trials is for example seen nowadays as a benchmark in clinical science, heavily improving the reliability of its findings.
It is more and more considered a new scientific field by itself to think of and argue about circumstances that improve the quality of scientific findings. Statistics is in some sense predestined to contribute to it, because probability and chance concepts are key to understanding of any empirical science.
There are also issues, in which statistics is not thought to be able to contribute much to understanding and to be of little help for solving them. For example, the selective attention that the agents in science are paying to new, strong and clear findings in each science is considered merely a psychological and socialissue. It is often explained by the affinity of humans for stories rather than mere facts and for novelties. The issue that scientific findings get more attention and are more likely to be published, read and cited is known for many years, but has gained new traction in meta science when the so-called reproducibility crisis emerged in (2003 to 2005). Studies such as ... showed that even rigid prosecution of study protocols did often not lead to reproduction of previous results when experiments were repeated, which struck empirical science in its core. Most reasonably, some concluded that it was a large misunderstanding of statistical tools such as the p-value in the scientific community that lead to this situation (...). Although that some measures have been taken since then, and some progress is made, the non-reproducibility of scientific findings remain an acute threat to the relevance and reliability of science. Some even argue that certain scientific fields sooner or later will repeat the experience that for example psychology or medicine have made, and are yet to encounter their own reproducibility crisis (...).
Clearly, reproduction of experimental results is the gold standard of testing empirical science, because it reassures the universality and reliability by which certain procedures or effects appear if measured in the correct experimental context. However, there are also problems there. One can almost never fully exclude chance events to have played a major role in negating or reaffirming the experimental findings (which is of course always the case where noise is present in the data). Furthermore, reproduction studies are generally costly and the precise experimental conditions might be hard to reproduce because of lack of information or other reasons.
There is another way to assess the strength of scientific results if the experiment has been conducted more than once (however not with exactly identical protocol as in reproduction studies):

Meta-analyses. In a meta-analyses, results of multiple (fairly identical) experiments (studies) are summarized to a single result. Very often, meta-analyses are not only regarded to be a synthesis of results, but of evidence, thus reflecting the overall and summarized evidence regarding to a scientific question. It is on purpose that selective attention to positive, strong results in the scientific literature have been mentioned beforehand, because it is clear how they will affect this second way of reassurance of scientific findings: The meta-analysis will again lead to irreproducible findings if it is based on a set of results that is itself biased. Thus, meta-analysis will in this case not reach its purpose of assessing reliability of science, but worsen the problem by reinforcing the confidence in the overestimated over-optimistic effects. However, and this is the main topic in this masters thesis, there are some ways to detect irregularities in the body of results that can provide indirect hints that a selective rather than a random sample from a hypothetical population of experiments is present. The abundance of the methods to link some features of the sample of experiments to publication bias, as the tendency of scientific literature to over-proportionally include large effects is generally called, also speaks for the relevance of the topic. A review in 2017 (...) for example identified 147 (!) conceptually differing methods.
This speaks as well for the inevitable difficulties that such methods encounter. First of all, it is most often impossible to estimate the real selection process, that is the rate by which smaller effects go unnoticed because of selective publication and reporting, because the number of missing results in the studies is not known at the first place and impossible to retrieve. Secondly, there is almost no real world data of complete and unbiased meta-analyses, such that evaluations of methods is dependent on simulations. So we have, after applying the methods, only indirect information of publication bias, and the extent to what it might influences scientific findings. Almost all the time, alternative explanations for the results of the methods might relieve the operators and publishers from the reproval of publication bias. However, given the large body of evidence for publication bias collected in other ways, those can probably be expelled most of the time, such that those methods indeed give a quantitative measure of publication bias.
In this masters thesis, the answers for two questions are investigated: What is the extent and effect of publication bias in clinical science? For this purpose, a dataset from the Cochrane Organization is analysed with commonly used techniques to detect and adjust for publication bias (more precisely, small study effects, as it will be seen later). The discussion of the results of the analysis will result in the light of the literature to publication bias.

Chapter 2

The Cochrane Dataset

2.1 Cochrane Systematic Reviews

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 a clinical or medical intervention or treatments to a control. 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, further specifications are provided by the variables in the columns.
The comparison variable specifies *what kind* of treatments or interventions are compared, the outcome variable *how* it is compared, and the subgroup variable (not indicated in table) if the result belongs to a certain subgroup. Here, the result is of a binary type, so the counts of 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". A event is here death 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

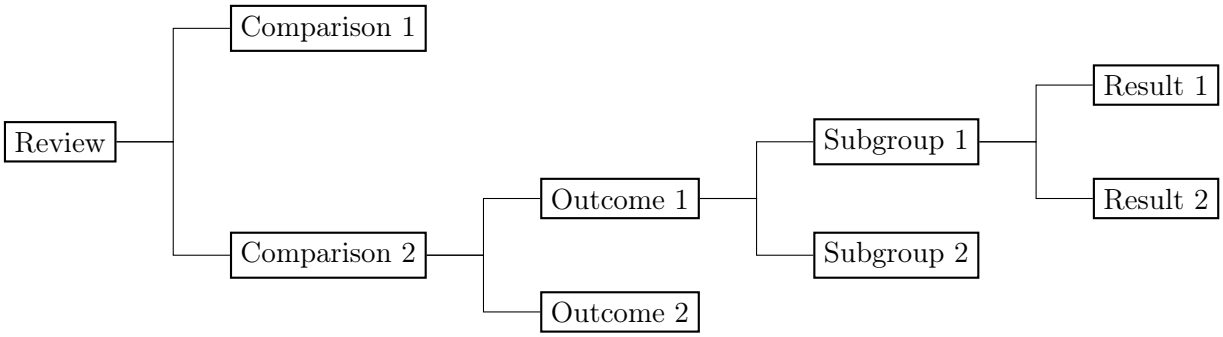

Figure 2.1: Structure of a hypothetical review with two different comparisons

The structure of a review will now be outlined based on an example of the dataset. Let us consider the previously mentioned barbiturate and head injury review. The aim was to "assess the effects of barbiturates in reducing mortality, disability and raised ICP (intra-cranial pressure) in people with acute traumatic brain injury" as well as to "quantify any side effects resulting from the use of barbiturates".
The review comprises five studies in total. Three of them compared barbiturate to placebo,

one compared barbiturate to Mannitol and one Pentobarbital to Thiopental. The studies have different outcomes, for example, death or death and severe disability at follow up, but also dropout counts or adverse effects (secondary outcomes). We have continuous (e.g. mean body temperature) and binary outcome data (e.g. death/no death). One study split up outcomes for patients with and without haematoma, which would be subgroups. Thus, it is important not to confuse results with studies. A study can contribute multiple results to a systematic review, for example, primary and secondary outcomes and adverse effects.

Study	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

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. For continuous outcomes, the cases have been counted were no treatment effect and standard error is available: neither mean values and standard deviations, nor mean differences and standard error. Also Study years before 1920 and after 2019 are declared 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 2001 and the quartiles are 1996 and 2008. Only a handful ($n = 18$) have been published in 2018, none in 2019.
The top treatment effect measure (risk ratio, mean difference, hazard ratio etc.) abundances are summarized in Table 2.5. One can conclude of the table that roughly 30 % of outcomes in the dataset are continuous and the rest being some sort of discrete or binary outcomes, most often binary ($> 65\%$).

The sample sizes among results vary to some extent. There are 5% of treatment group sample sizes that are smaller than 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

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,

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 due to large reviews with a high number of studies and results, most extreme in **?** which is a systematic review about antibiotic prophylaxis for preventing infection after cesarean section, with 95 studies and 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 specifically, 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. Importantly, this distinction is also used by the Cochrane Organization itself, i.e. the meta-analyses are identical to the meta-analyses done in the systematic reviews.
The dataset is divided in meta-analyses with identical experimental setup. The size of a meta-analysis denotes how many results are included in a group. Table 2.6 shows the number of meta-analysis with size $\geq n$ results. Practically, this 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

2.2 Data Tidying and Processing

The original dataset was provided by .. (some words how the dataset was obtained and processed by C.R.).
The information in the previous pages are from a tidied and processed version of this dataset.

2.2.1 Modification of old variables

In some cases, reasonable assumptions led to small changes in the originally provided variables:

- **study.year**: It was assumed that studies that are declared to have been published before 1920 ($n = 30664$) are mis-specified, as well as after 2019 ($n = 384$), so these have been set to NA.

- **mean1** and **mean2**: When both "mean1" and "mean2" are equal to zero or NA, "outcome.measure.new" is equal to "(Std.) Mean Difference", but a mean difference is given in "effect", "mean1" is set equal to "effect" ($n = 1593$).

2.2.2 Newly Introduced Variables

Some new variables are added to the obtained dataset:

- **outcome.measure.new**: Outcome measure specifications given in "outcome.measure" were standardized whenever they were supposed to denote the same outcome measure. An example would be the formally different notations for odds ratios: "Odds Ratio", "odds ratio", "OR", which were all denoted as "Odds Ratio".

- **outcome.type**: A simplification of "outcome.measure.new". The outcome.type "bin" indicates if "outcome.measure.new" is either one of "Risk Ratio", "Odds Ratio", "Risk difference" or "Peto Odds Ratio". "cont" is equivalent to "outcome.measure.new" equal to "Std. Mean Difference" or "Mean Difference". "surv" equal to "Hazard Ratio" and "rate" equal to "Rate Ratio" or "Rate Difference".

- **lrr** and **var.lrr**: log risk ratio and variance of the log risk ratio for outcome.type "bin".

- **smd** and **var.smd**: Hedges g and the variance of Hedges g for outcome.type "cont". If not computable from "mean1", "mean2", "sd1" and "sd2", and "outcome.measure.new" was equal to "Std. Mean Difference", it was set equal to "effect" (and "var.smd" equal to "se" squared, $n = 6938$).

- **smd.ordl** and **var.smd.ordl**: Cohen's d and its variance as obtained by transformation of a log odds ratio for outcome.type "bin".

- **cor.Pearson** and **var.cor.Pearson**: Pearson correlation coefficient and variance as obtained from the d (for outcome.type "bin") or g (for outcome.type "cont") to r transformation.

- **z** and **var.z**: Fisher's z score and it's variance obtained from the Pearson correlation r to z transformation.

- **pval.single**: p-value against the null hypothesis of no treatment effect, derived by a t-test for outcome.type "cont" or Wald test for outcome.type "bin".

- **events1c** and **events2c**: Correction of "events1" and "events2" zero event counts or event counts = patient number. When no events occurred, 0.5 was added, and when all patients experienced the event, 0.5 was subtracted. When one of "events" had zero counts while the other had maximum counts, no adjustment occurred.

- **meta.id**: Meta-analysis ID variable to uniquely identify any potential meta-analysis in the dataset. Consistent to what has been discussed before, all results that share a common comparison, outcome and subgroup (optional, subgroups not given in any case) may be combined in a meta-analysis.

2.2.3 Eligibility criteria for Publication Bias Test and Adjustment

Initially, the analysis is restricted to binary, continuous and survival outcomes. More formally:

- **Outcome measures**: The analysis has been restricted on the following outcome measures (from "outcome.measure.new"): "Odds Ratio", "Risk Ratio", "Risk difference", "Peto Odds Ratio", "Std. Mean Difference", "Mean Difference", "Hazard Ratio" ($n = 456727$ out of a total of 463820 results)

The suggestions of criteria for eligibility for publication bias tests of Ioannidis and Trikalinos (2007) have largely been followed:

- **Sample size**: A meta-analysis is comprised of at least ten studies ($n = 5797$ remaining).

- **Heterogeneity**: The I^2 statistic of a given meta-analysis is smaller than 0.5, thus, the proportion of between study variance of the overall variance is smaller than 0.5 ($n = 4038$ remaining).

- **Study size**: The ratio between largest variance of an estimate and smallest variance of an estimate is larger than four ($n = 4006$ remaining).

- **Significance**: At least one treatment effect has a p-value below the significance threshold 0.05 ($n = 2673$ remaining)

After having excluded meta-analyses with less than ten studies, additional criteria for excluding inconvenient meta-analyses are:

- **Zero events**: In the case of binary outcomes, meta-analyses with zero events in any study and any group are excluded ($n = 141$ out of meta-analyses with at least ten studies).

- **Missing means and sd's**: In the case of continuous outcomes, meta-analyses with means and sd's in any group being equal to zero are excluded ($n = 306$ out of meta-analyses with at least ten studies).

- **Single patient data**: Meta-analyses that are comprised of single patient data are excluded ($n = 8$ out of meta-analyses with at least ten studies)

–Make Flowchart–

Exclusions due to Computational Errors

In exceptional cases, the applied meta-analysis methods and publication bias tests and adjustment algorithms failed due to various reasons. Here, those cases are listed and if known, the reasons why computation failed are given.
Binary Outcomes: Out of the meta-analyses that are of "outcome.type" bin, and have at least ten studies, two studies hat to be excluded:

- meta.id = 62301 (file.nr = 1531, comparison.nr = 8, outcome.nr = 6, subgroup.nr = 1): Largest study (se = 0.1) has risk ratio 1 and smallest study (se = 1) risk ratio 0.07. Copas publication bias adjustment methods has likelihood optimization issues.

- meta.id = 94519 (2519, 14, 1, 2): Largest study (se = 0.05) has risk ratio 0.9 and smallest study (se = 0.4) risk ratio 3. Again Copas publication bias adjustment methods has likelihood optimization issues.

There were no issues with "outcome.type" "cont" and "surv". For the subsequent information, it is important to know that meta-analyses are only repeated with Hedges g, Pearson correlation coefficient and Fisher's Z-scores, if the outcome type is not survival and all of the above criteria are met ($n = 0$).
Pearson Correlation Coefficient: Three studies had to be excluded when analyzing the Pearson correlation coefficients:

- meta.id = 157083 (file.nr = 5061, comparison.nr = 1, outcome.nr = 5, subgroup.nr = 2): A meta-analysis with all results from the same study, and equal sample size (5 each group). The Copas selection model algorithm issued an error when optimizing the likelihood.

- meta.id = 159329 (5183, 3, 13): One very small study with group sizes of 2 and 1 patients and small risk ratio of 0.222 (compared to the largest of 1.05, se = 0.04). Also, the Copas selection model algorithm issued an error when optimizing the likelihood.

- meta.id = 182298 (6211, 1, 8, -): One very small study with 8 patients in each group and 0 event counts in both. The Copas selection model algorithm issued an error when optimizing the likelihood.

Fisher's Z score: The same issues as encountered for Pearson correlation coefficients (meta.id = 157083 and 159329). The variance of the z score s_z^2 can be calculated as: $s_z^2 = \frac{1}{n-3}$, n being the total sample size. Thus, studies that have total sample size $n \geq 3$ are discarded in the meta-analyses ($n = 72$ studies). 2 meta-analyses have then less than 10 studies left for pooling:

- meta.id = 10671: (file.nr = 5061, comparison.nr = 1, outcome.nr = 5, subgroup.nr = 2), 7 studies left.

- meta.id = 43324: (1019, 1, 3, 2), 9 studies left.

The Publication Bias Test and Adjustment Dataset

The dataset that was ultimately tested and adjusted for publication bias comprises 2673 meta-analyses and 42789 results. ..

Chapter 3

Results

3.1 Small study effects

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

The median z-score for a given sample size of a trial is shown in Figure 3.1. It is clearly visible that the absolute value decreases with increasing sample size, i.e. that the effect size is becoming smaller. The trend flattens off after $\sim$ sample size = 400 (not shown). – Appendix –

Only for illustration, the same trend is reproduced in Figure 3.2 with a similar method, using the original effect size measures "Odds Ratio", "Risk Ratio", "Mean Difference" and "Std. Mean Difference" (the most commonly used in the dataset). The "normalized effect" is the original effect size, normalized with respect to all other effect sizes of the same measure (i.e. subtraction of mean and division through standard error of the mean).

3.2 Small Study Effect Tests

The results of the test for small study effects for the meta-analyses fullfilling the criteria from chapter **??**, section 2.2, are presented in Figures 3.3, 3.4 and 3.5. The histogram of the p-values appears most often skewed to the right, indicating evidence for small study effects. Because different tests are applied depending on outcome type, the results are displayed separately for binary, continuous and survival outcome types (as previously defined in the dataset variable "outcome.type", see chapter **??**, section 2.2.
For continuous and survival outcomes, the names refer to:

- Egger's test, the weigthed linear regression test as described in section 4.4.1
- Thompson and Sharp's test, the weighted linear regression test adjusted for between-study heterogeneity, section 4.4.1
- Begg and Mazumdar's test, the rank test described in section 4.4.1

For binary outcomes, the names refer to:

- Harbord's test, the likelihood score based test (section 4.4.2)
- Peter's test, the weighted linear regression with inverse sample size as explanatory variable (study size proxy) described in section 4.4.2

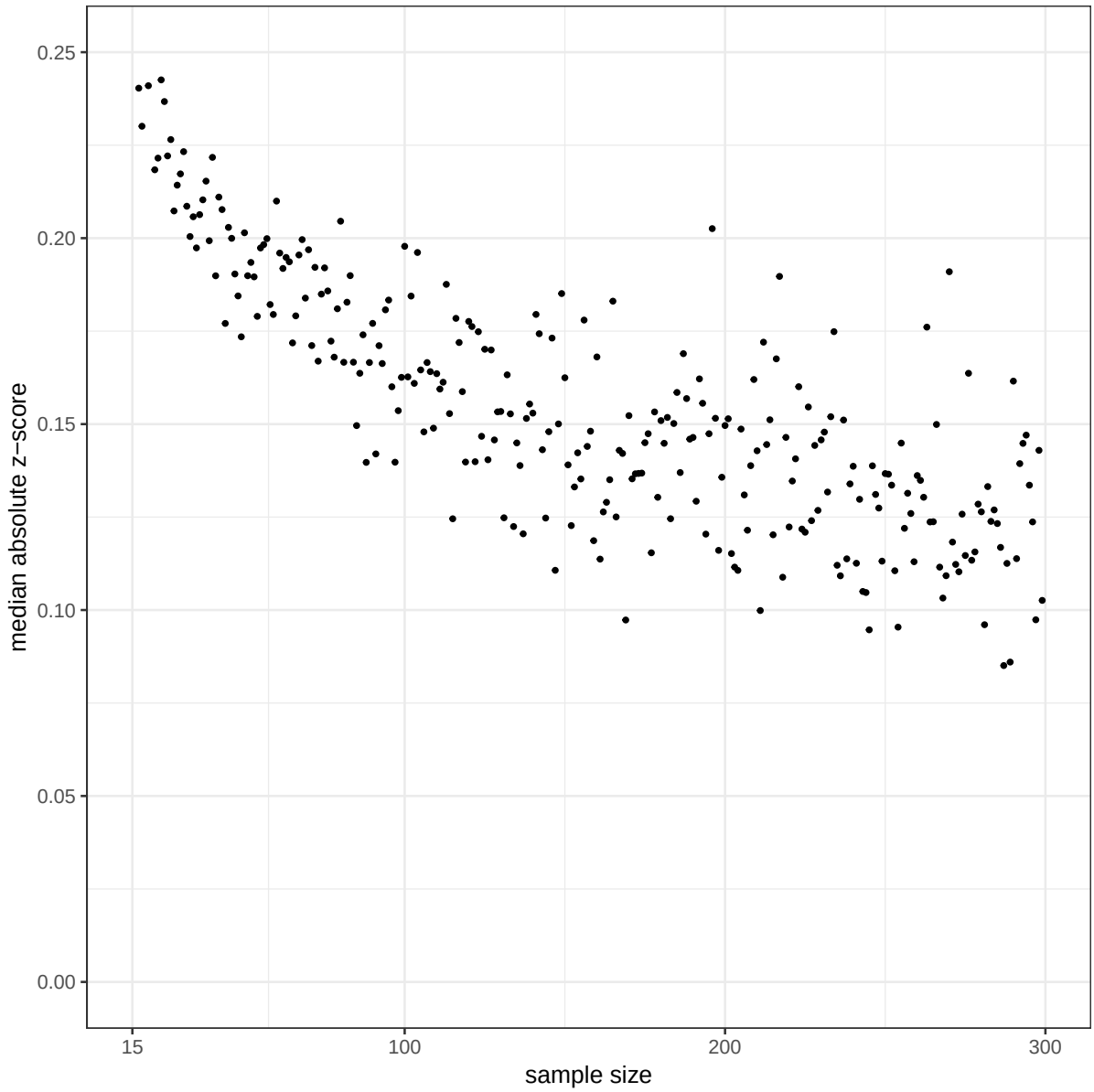

Figure 3.1: Median of the absolute z-score plotted against the total sample size.

- Rücker's test, the test based on the arcsine transformation of proportions, in combination with Thompson and Sharp's regression test (section 4.4.2)
- Schwarzer's test, the rank based test using the expected event counts computed with the hypergeometric distribution (section 4.4.2)

3.2.1

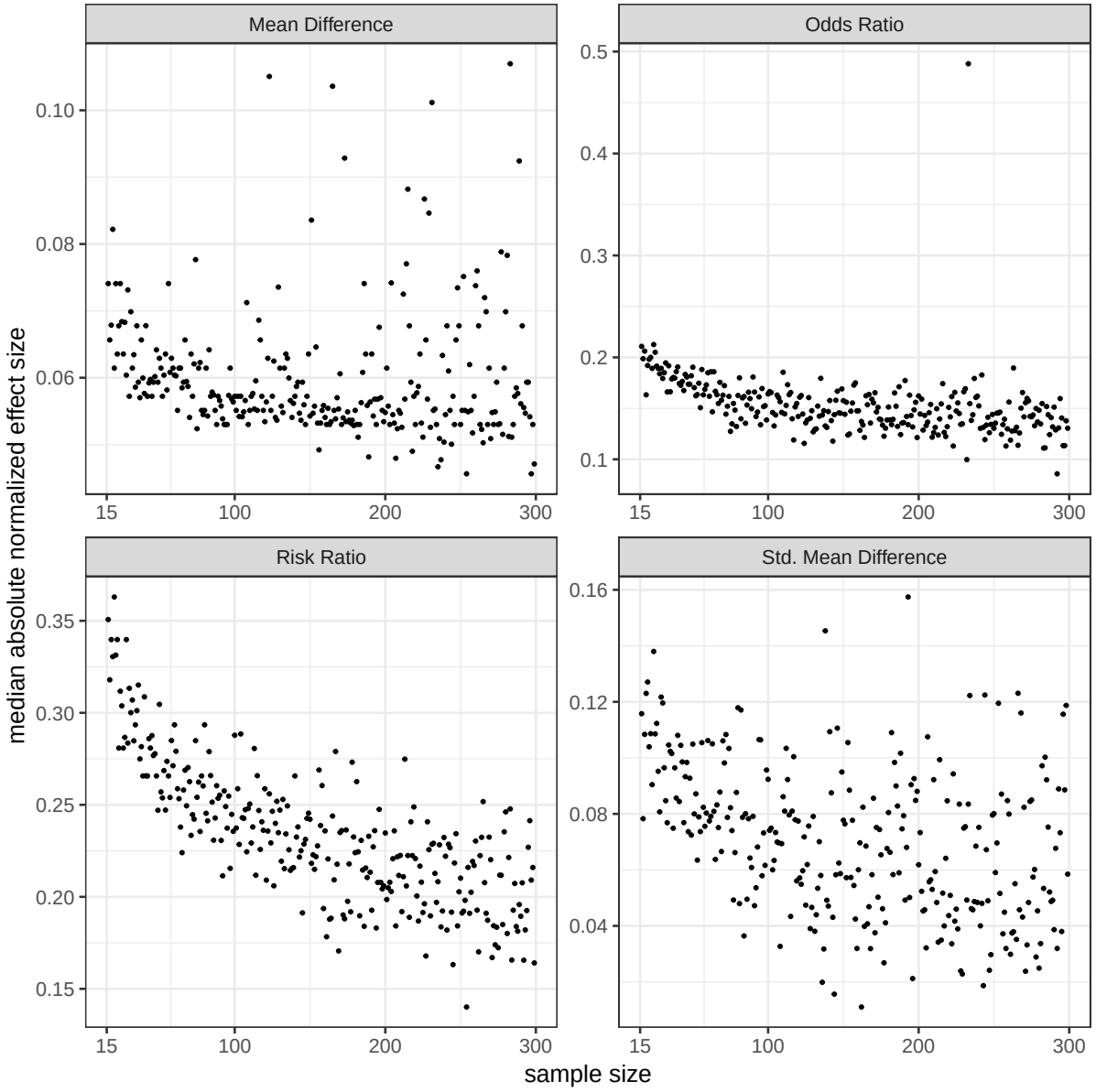

Figure 3.2: Median of the absolute value of the normalized, original effect size plotted against the total sample size.

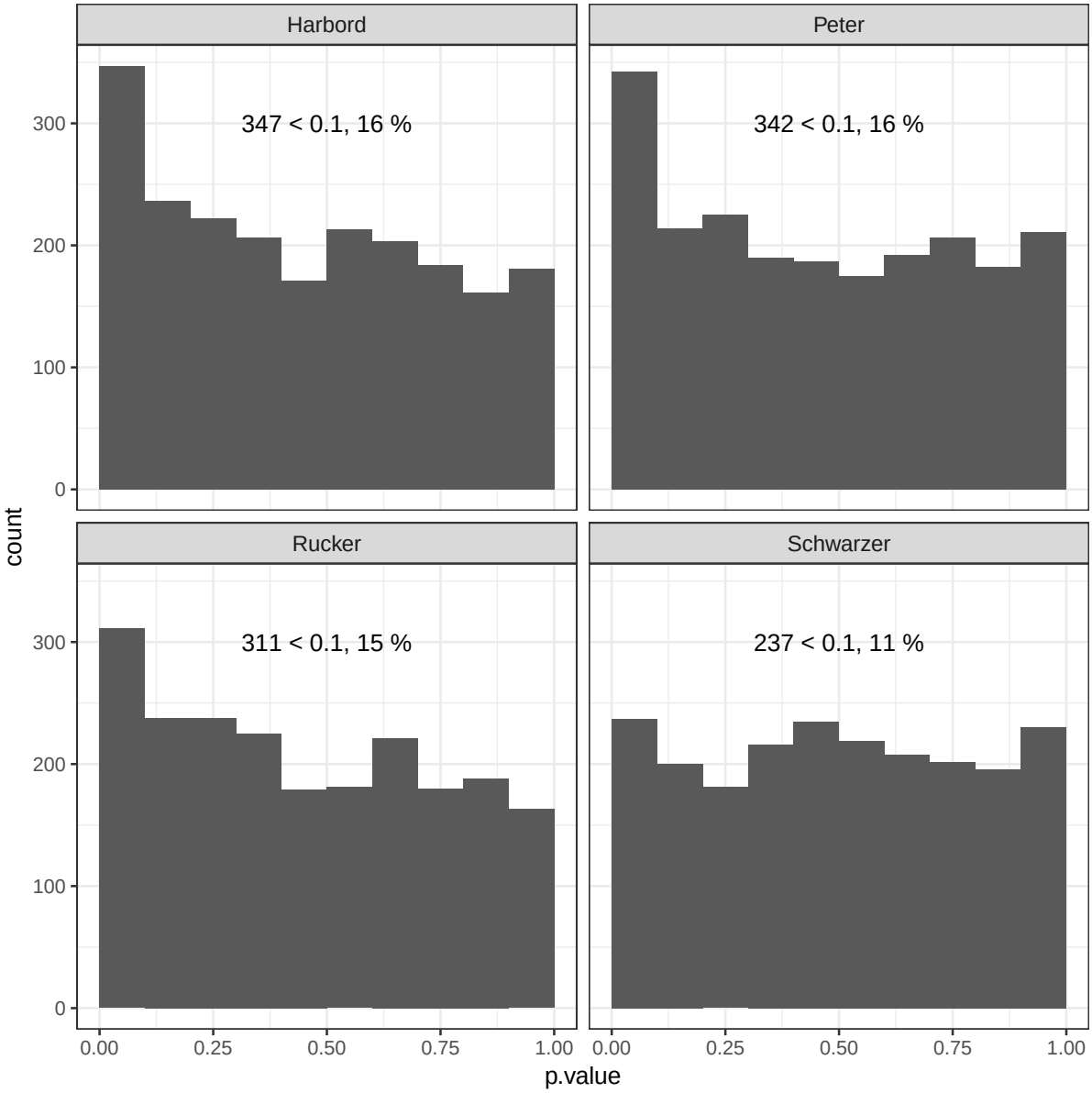

Figure 3.3: Histogram of the p-values for small study effect in binary outcome meta-analyses. The testing method is indicated in the header, binwidth is equal to 0.1. The significant proportion based on the threshold of 0.1 is displayed inside the figures.

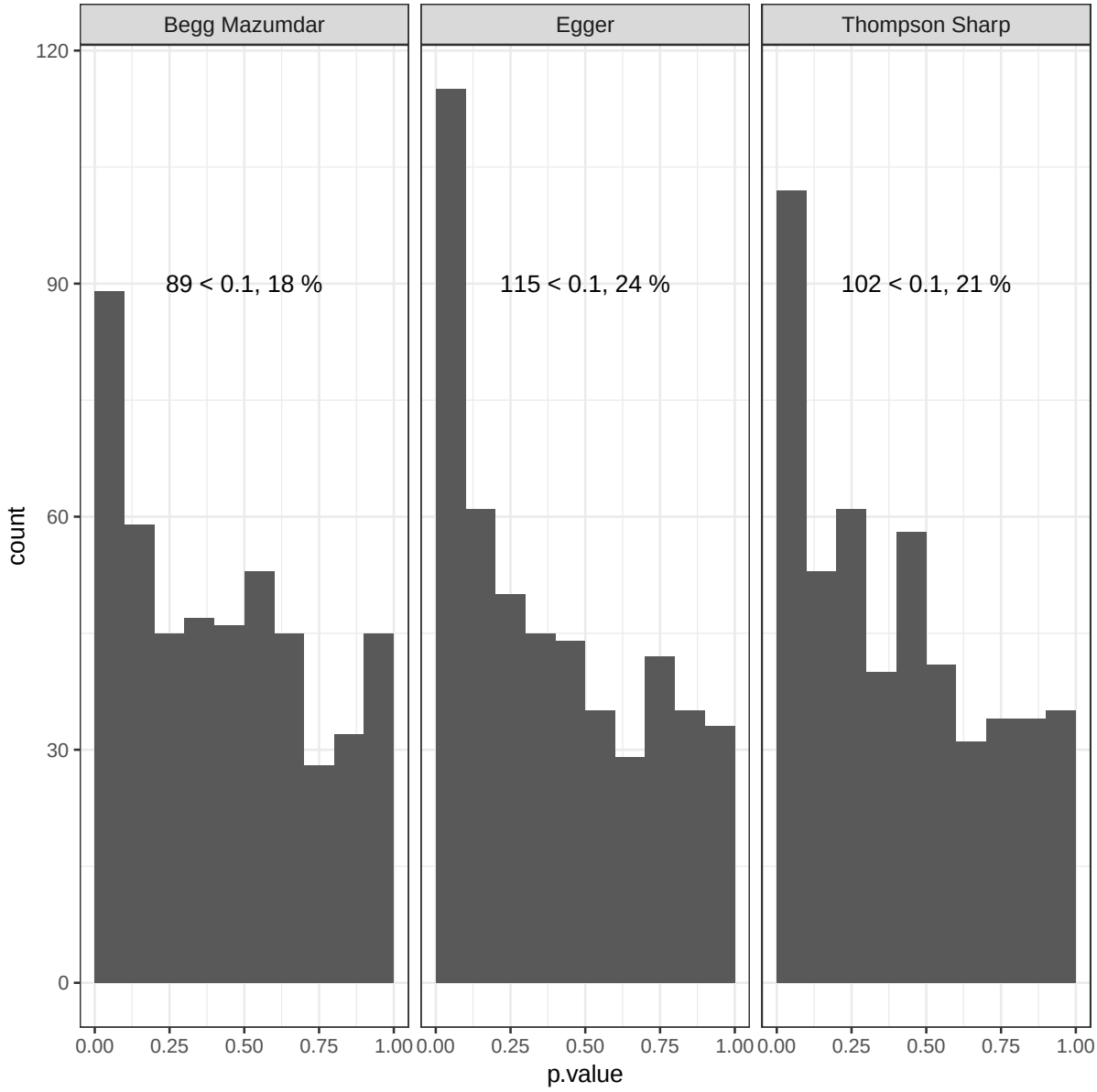

Figure 3.4: Histogram of the p-values for the small study effect in continuous outcome meta-analyses. The testing method is indicated in the header, binwidth is equal to 0.1. The significant proportion based on the threshold of 0.1 is displayed inside the figures.

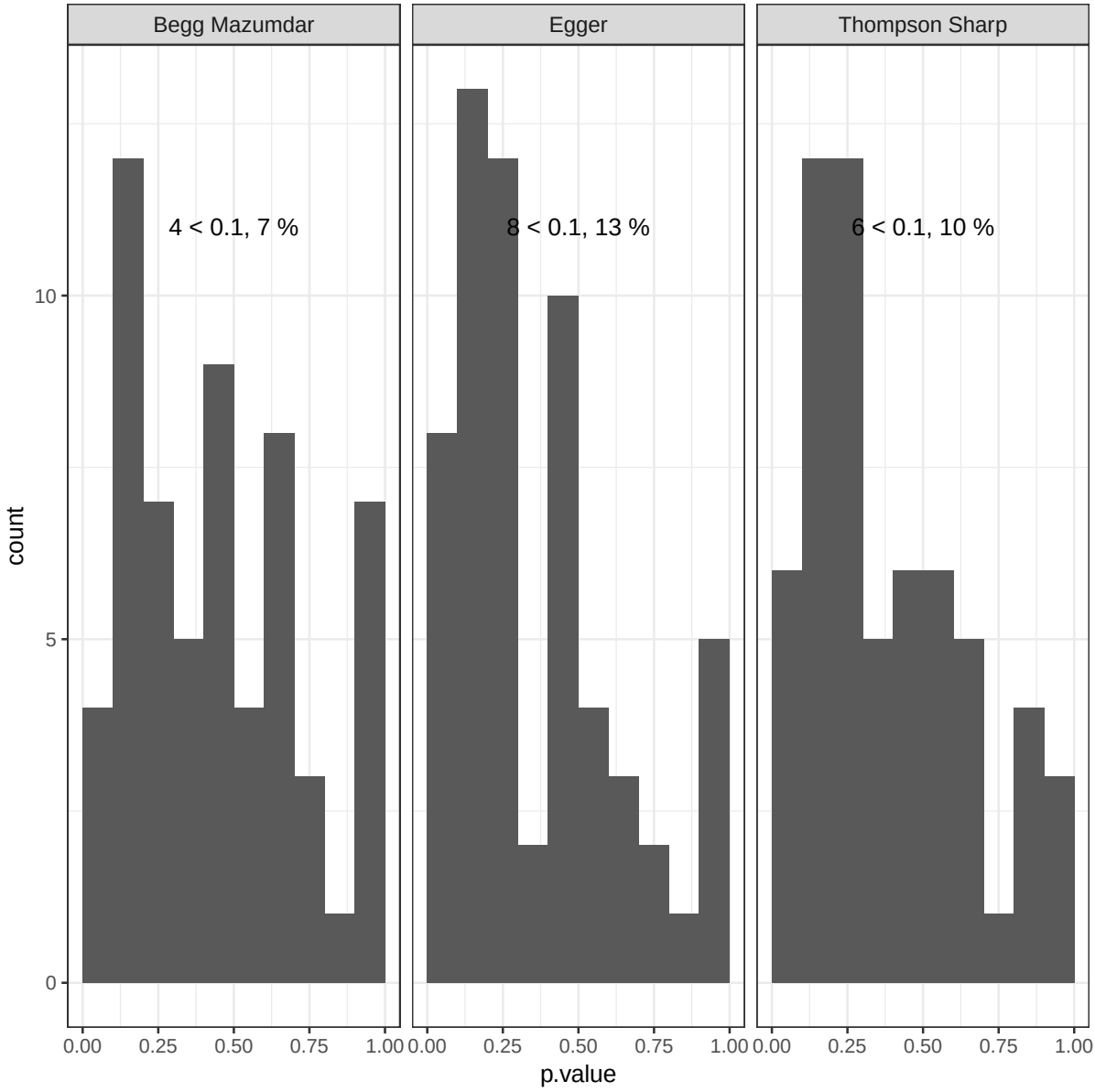

Figure 3.5: Histogram of the p-values for the small study effect in survival outcome meta-analyses. The testing method is indicated in the header, binwidth is equal to 0.1. The significant proportion based on the threshold of 0.1 is displayed inside the figures.

Chapter 4

Methods

4.1 Introduction and Notation

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

4.2 Effect Measures and p-values

4.2.1 Continuous Outcomes

For given $(m_t, m_c), (\mathrm{sd}_t, \mathrm{sd}_c)$ and (n_t, n_c), one can compute mean difference as well as a standardized mean difference (here: Hedges g) and a standard error thereof. Note that the definition of Hedges g and its standard error s varies among the literature, the following applies for this report:

$$s = \sqrt{\frac{(n_t - 1)\mathrm{sd}_t^2 + (n_c - 1)\mathrm{sd}_c^2}{n_t + n_c - 2}} \qquad g = \frac{m_t - m_c}{s} \tag{4.1}$$

The mean difference θ and its standard error s can similarly be obtained by

$$\theta = m_t - m_c \qquad s = \sqrt{\mathrm{sd}_t^2/n_t + \mathrm{sd}_c^2/n_t} \tag{4.2}$$

Both estimators take into account that the two groups might have unequal variances. A p-value to test the null hypothesis that the mean between group is equal is commonly obtained with the students t test. The t statistic is obtained, using s and g from 4.1, by

$$t = g/(s\sqrt{(1/n_t) + (1/n_c)})$$

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

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

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

4.2.2 Binary Outcomes

Two commonly used effect measures for binary outcome data are risk ratios and odds ratios between treatment and control group. The methods presented here can also be found, for example, in (Borenstein *et al.*, 2011, 34). Let θ be the logarithm to base 10 of the odds ratio. θ and its variance s^2 can be obtained by

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

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

$$\theta = \log(\frac{e_t/n_t}{e_c/n_c}) \tag{4.3}$$
$$s^2 = 1/e_t + 1/n_t + 1/e_c + 1/e_t \tag{4.4}$$

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

$$p = 2 * (1 - \Phi(|\hat{\theta}/\hat{s}|)$$

, a p-value for the corresponding estimates, which summarizes the evidence against $\hat{\theta}$ being zero (i.e. the true risk/odds ratio being 1).
Binary and continuous effect measures can be converted into each other, as described in section 4.6.

4.2.3 Survival Outcomes

Time-to-event data with censoring has to be analyzed by special means. One frequently used method to take into account right-censoring is the Cox proportional hazards regression model (Cox, 1972). Because the method itself is not applied in this thesis, but only the resulting estimates of the parameters are used, the reader is referred to the extensive literature covering

this topic (e.g. Cox and Oakes (1984)).
The so-called hazard ratio estimated by cox p.h. regression is the ratio of the instantaneous risk of experiencing the event between two groups. Because it is a maximum likelihood estimator, one can again use its Wald test statistic to test for equal hazards. Let $\hat{\theta}$ be an estimate of the log hazard ratio and $\hat{s}$ an estimate of the standard error of it. As before

$$p = 2 * (1 - \phi(|\hat{\theta}/\hat{s}|)$$

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

4.3 Fixed and Random Effects Meta-Analysis

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

$$\theta_M = \frac{\sum_{i=1}^n w_i \theta_i}{\sum_{i=1}^n w_i} \qquad s_M^2 = \frac{1}{\sum_{i=1}^n w_i} \tag{4.5}$$

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

$$\theta_i \sim N(\mu_i, s_i^2) \qquad \mu_i \sim N(\theta_M, \tau^2) \tag{4.6}$$

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

$$\theta_i | \mu_i \sim N(\theta_M, s_i^2 + \tau^2)$$

τ^2 is often referred to as a population variance or between-study variance, whereas s_i^2 can be interpreted as sampling error. The pooled treatment effect estimate θ_M of the random effects model and its variance is obtained by replacing the weights w_i in equation 4.5 with $w_i = 1/(s_i^2 + \tau^2)$.
The model is superior to the fixed effects model whenever the standard errors of the treatment effects alone are unlikely to fully account for the entire variability observed between studies. Note that as τ^2 increases, each θ_i will eventually get equal weights, irrespective of it's sampling error s_i^2.
The estimation of τ^2 has been subject to some debate in the statistical literature. Oftentimes, the method of moment estimator of DerSimonian and Laird (1986) is used. We use the measure of heterogeneity, Q, and divide by C after having subtracted the degrees of freedom $n - 1$:

$$Q = \sum_{i=1}^n w_i (y_i - \theta_M)^2 \qquad C = \sum_{i=1}^n w_i - \frac{\sum_{i=1}^n w_i^2}{\sum_{i=1}^n w_i} \tag{4.7}$$

$$\tau^2 = \max(0, \frac{Q - (n-1)}{C}) \tag{4.8}$$

Again, $w_i, \theta_i ...$ have to be replaced by their estimates in order to get an estimate $\hat{\tau}^2$.
The Paule-Mandel estimator (Paule and Mandel, 1982) is considered to have most often better

properties than the method of moments estimator (e.g. Veroniki *et al.* (2016)). Since we defined $w_i = 1/(s_i^2 + \tau^2)$, it als holds that

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

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

$$s^2(w_i\theta_i) = \frac{\sum_{i=1}^n w_i(\theta_i - \theta_M)^2}{n-1} \qquad \frac{\sum_{i=1}^n w_i(\theta_i - \theta_M)^2}{n-1} = 1 \tag{4.9}$$

θ_M can be estimated using equation 4.5, the only problem is to estimate τ^2. It can be obtained through an iterative process, using a newly defined function

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

In view of equation 4.9, τ^2 must be such that $F(\tau^2) = 0$. Then, we start with a arbitrary τ^2 and repeatedly add a term τ_0^2 to update τ^2 until $F(\tau^2 + \tau_0^2)$ is close to zero (using $\tau^2 + \tau_0^2$ for $\hat{w}_i, \hat{\theta}_M$). Using a truncated Taylor series expansion, one can obtain the partial derivative after τ^2, which is a reasonable choice for τ_0^2.
The estimation of τ^2 is accompanied by uncertainty. A common procedure is to test if it is there is significant heterogeneity between the studies (Borenstein *et al.*, 2011, 109). Compute Q, as given in 4.7, based on one of the estimators of τ^2. It is assumed that Q follows a central Chi-squared distribution with $n-1$ degrees of freedom under the null hypothesis of equally distributed effect sizes. Thus, the expected value of Q is $n-1$, and the excess dispersion is $Q - n + 1$. The p-value against the null hypothesis of equally distributed effect sizes is $1 - F(Q)$, using F as the cumulative distribution function of the Chi-squared distribution with d.f. $= n-1$.
An advantage of the τ^2 is that it is directly linked to the variability in the data. Additionally, one can use the I^2 statistic to see what portion the between study variance has of the overall variance. It is computed as

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

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

4.4 Small Study Effects Tests

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

4.4.1 Continuous Outcome Tests

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

Begg and Mazumdar: Rank Correlation Test

Begg (1988) proposed a rank based test to test the null hypothesis of no correlation between effect size and variance. A standardized effect size $\theta_i^\star$ can be computed as in 4.10. $s_i^{2\star}$ is the variance of $\theta_i - \theta_M$ as defined in 4.11, θ_M being the fixed effects pooled treatment effect (4.5.

$$\theta_i^\star = (\theta_i - \theta_M)/s_i^{2\star} \tag{4.10}$$

$$s_i^{2\star} = s_i^2 - 1/\sum_{i=1}^{n} \frac{1}{s_i^2} 1 \tag{4.11}$$

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

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

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

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

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

Egger's Test: Weighted Linear Regression Test

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

$$y = \beta_0 + \beta_1 x + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{4.12}$$

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

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}}(\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i) \tag{4.13}$$

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

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

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

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{4.14}$$

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

$$\hat{\sigma}^2 = \frac{1}{n-2} \mathbf{r}^\top \hat{\mathbf{r}} \qquad\qquad \hat{\mathbf{s}}^2_{\hat{\beta}} = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} \tag{4.15}$$

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

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

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

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

where F is the cumulative t distribution with $n - 2$ degrees of freedom. p will give the evidence against the null hypothesis $\beta_0 = \beta_{H0}$.
The concept is extendable to weighted linear regression. Weighted linear regression may be used if the residuals $\mathbf{r}$ have unequal variances, which is equivalent to ascribe different precision to the observed y. The least squares equation 4.13 is extended to

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

with positive weights w_i that penalize large squared residuals for some i more if w_i is larger compared to other w_i.
Let $\mathbf{W}$ be a $n \times n$ matrix with $\mathbf{W}_{ii} = w_i$, the weights on the diagonal and zeros on the off-diagonals. The estimates in 4.14 and 4.15 can still be used if $\mathbf{X}$ is exchanged with $\mathbf{X}^\star = \mathbf{W}\mathbf{X}$ and $\mathbf{y}$ with $\mathbf{y}^\star = \mathbf{W}\mathbf{y}$.
Now it is shown how linear regression can be used to test dependency of effect sizes on study sizes. The simplest application was introduced by Egger *et al.* (1997). Let θ/s be the dependent variable y and $1/s$ the explanatory variable x. If plotted, this corresponds to a radial or Galbraith plot Galbraith (1988). The linear regression equation as introduced before in 4.12 can be written in two ways:

$$\theta/s = \beta_0 + \beta_1/s + \epsilon, \qquad\qquad \epsilon \sim N(0, \sigma^2) \tag{4.16}$$

4.16 is often provided due to the correspondence to the radial plot. However, it is equivalent to

$$\theta = \beta_0 + \beta_1 s + \epsilon, \qquad\qquad \epsilon \sim N(0, w^{-1}\sigma^2) \tag{4.17}$$

with weights $w = 1/s^2$. Thus testing β_0 of 4.16 or β_1 is equivalent. The corresponding p-value is then used as evidence for a small study effect. Plugging in $\theta_i/s_i, 1/s_i$ as y_i, x_i into equation 4.14 and 4.15 will give the estimates for $\hat{\beta}_0, \hat{\beta}_1, \hat{s}_{\beta_0}$ and $\hat{s}_{\beta_1}$.

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

A method proposed in Thompson and Sharp (1999) allows for between study variance τ^2, as introduced before in section 4.3. It extends the previously seen linear regression approach with $x = 1/s$ and $y = \theta/s$ by introducing weights. The effect size θ_i is assumed to be distributed as

$$\theta_i \sim N(\beta_0 + \beta_1 s_i, s_i^2 + \tau^2) \tag{4.18}$$

τ^2 is estimated as in equation 4.8 (method of moments). The weights are set as $w_i = 1/\sqrt{s_i^2 + \tau^2})$. After adjusting for the weights as described in 4.4.1, we can proceed analogous to Egger's test. The p-value for $\beta_0 \neq 0$ reflects the evidence for a small study effect.

4.4.2 Dichotomous Outcomes Tests

The issue with dichotomous outcomes is that effect size and variance of effect size are correlated, which can readily be seen in 4.3 and 4.4. For example, a small number of event counts in one or group will inflate the variance and the effect size. Consequently, the tests above will tend to reject the null-hypothesis too often, i.e. report false positives. A number of solutions to this problem are provided in the literature.

Peters Test: Weighted Linear Regression Test

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

Harbord's Test: Score based Test

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

$$\varphi = e_t - (e_t - e_c)(e_t + (n_t - e_t))/(n_t + n_c) \tag{4.19}$$

$$s^2 = \frac{(e_t + e_c)(e_t + (n_t - e_t))(e_c + (n_c - e_c))((n_t - e_t) + (n_c - e_c))}{(n_t + n_c)^2(n_t + n_c - 1)} \tag{4.20}$$

It can be shown that they are both good approximations of the log odds ratio and its variance if the real θ is not too far from zero. The standardized estimator r_i/v_i is also known as Pet0 odds ratio. The obtained scores and variances can be used in Egger's test as treatment effects and variances.

Schwarzer's Test: Rank Correlation Test

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

$$(e_t - \mathbb{E}(E_t))/\sqrt{(s_i^2)} \tag{4.21}$$

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

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

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

$$\theta_i = \arcsin e_t/n_t - \arcsin e_c/n_c s_i^2 = 1/4n_t + +/4n_c$$

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

4.5 Small study effect and Publication Bias Adjustment

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

4.5.1 Adjustment by Regression

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

$$\theta_i = \beta_0 + \beta_1 \sqrt{s_i^2 + \tau^2} + \epsilon_i \sqrt{v_i + \tau^2}, \epsilon_i \overset{\text{iid}}{\sim} N(0,1) \tag{4.22}$$

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

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

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

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

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

4.5.2 Copas Selection Model

A method proposed in Copas and Shi (2001, 2000); Copas and Malley (2008) assumes that the given sample of treatment effects and standard errors is a selected part of a larger random sample. Selection of studies depends on their effect size and variance. Smaller variance is always

accompanied by larger selection probability.
Let θ_i be the effect size estimate of study i. Then

$$\theta_i \sim N(\mu_i, \sigma_i^2)\mu_i \sim N(\theta, \tau^2) \tag{4.23}$$

which is similar to the random-effects meta-analysis setting. θ is the population mean effect, σ_i^2 the within study variance and τ^2 the between study variance. Equations in 4.23 are termed the *population model*.
The *selection model* is defined as follows. Suppose a selection of studies with reported ($\neq$ estimated) standard errors s (likely different from σ). Only a proportion of the selection will be published, with a defining the overall proportion of published studies and b (assumed to be positive) defining how fast this proportion increases with s becoming smaller. Formally,

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

The equation can be rewritten as

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

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

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

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

$$P(\text{select}|s_i, \theta_i) = \Phi(\frac{u_i + \tilde{\rho_i}((\theta_i - \mu)/\sqrt{(\tau^2 + \sigma_i^2)})}{\sqrt{1 - \tilde{\rho_i^2}}})$$

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

$$\mathbb{E}(\theta_i|s_i, \text{select}) = \mu + \rho_i\sigma_i\lambda(u_i) \tag{4.24}$$

which shows that the expected value for a study is larger for larger σ.
A likelihood for θ_i, conditional on $z > 0$ can be formulated to estimate the parameters of the model. Regarding a and b, there is no way that they can be estimated because the number of

missing studies and their effect sizes is not known. Instead, fixed values for a and b have to be chosen. The nuisance parameter σ_i can be replaced by an estimate if the sample size in the studies is large enough. Since

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

with $c^2 = \lambda(u_i)(u_i + \lambda(u_i))$, we can replace σ_i^2 by $\hat{\sigma}_i^2 = \frac{1}{1-c_i^2\rho_i^2}$. Although one has to evaluate the likelihood for fixed pairs (a, b), one can compare the fit of the model to evaluate which one is more suitable: With equation 4.24, one can obtain fitted values of θ_i based on s_i. Also, for two different pairs (a, b), $(a^\star, b^\star)$,

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

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

$$\theta_i = \theta + \beta s_i + \sigma_i\epsilon_i$$

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

$$\chi^2 = 2 * (\max_{\theta,\tau,\beta} 1\, \tilde{L}(\theta, \tau, \beta) - \max_{\theta,\tau} \tilde{L}(\theta, \tau, 0)) \tag{4.25}$$

with

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

χ^2 can be used with a χ^2 distribution with one degree of freedom to obtain a p-value. Note that the test is almost equivalent to Egger's small study effect test with $\tau^2 = 0$. Thus, although the copas selection model models publication bias, it is dependent on the small study effects to find the most suitable pair (a, b).
If one applies the model to a single meta-analysis, a sensitivity analysis is suggested. One can observe how θ and it's confidence intervals change dependent on the underlying selection process. Selection models are in general not recommended for inference (e.g. McShane *et al.* (2016)). Rücker *et al.* (2011) have shown how the method can be implemented in a simulation for inference purposes.
A range of values of (a, b) are applied, and the test for residual small study effect as described in equation 4.25 is applied. If all obtained p-values from the test are above a threshold 0.1, this is interpreted as no evidence, and no need for modelling, and the standard, classical random effects meta-analysis is retained. If none of the p-values is above the threshold, a wider range of values for (a, b) is used. When some p-values are above, and some below the threshold, the pair (a, b) with the smallest number of missing studies is retained (that is, the least intense underlying selection model is chosen).
Currently, there is no test to detect miss-specifications in the model itself, the authors themselves have argued that a non-parametric test of the residuals would lack power.

4.6 Transformation between effect sizes

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

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

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

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

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

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

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

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

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. 23

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, 22, 29

Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis . *Biostatistics*, **1**, 247–262. 26

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. 26

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. 26

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall. 21

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187–202. 20

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177 – 188. 21

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. 22

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. 24

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons. 22

Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, **7**, 889–894. 24

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. 25

Hedges, L. V. and Olkin, I. (1985). Chapter 11 - combining estimates of correlation coefficients. In Hedges, L. V. and Olkin, I., editors, *Statistical Methods for Meta-Analysis*, 223 – 246. Academic Press, San Diego. 29

Held, L. and Sabanés Bové, D. (2014). Applied statistical inference. *Springer, Berlin Heidelberg, doi*, **10**, 978–3. 20

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

Ioannidis, J. P. and Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *CMAJ*, **176**, 1091–1096. 10

Kasuya, E. (2001). Mann-whitney u test when variances are unequal. *Animal Behaviour*, **6**, 1247–1249. 20

McShane, B. B., Böckenholt, U., and Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, **11**, 730–749. 28

Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, **87**, 377–385. 21

Rosenthal, R. and Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500–504. 21

Rücker, G., Carpenter, J. R., and Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal*, **53**, 351–368. 26, 28

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. 25

Schwarzer, G., Carpenter, J. R., and Rücker, G. (2015). *Meta-analysis with R*, volume 4724. Springer. 23

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, **18**, 2693–2708. 25

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, **7**, 55–79. 22

Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, **10**, 1665–1677. 21