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 March 21, 2019

Contents

Preface

Howdy!

Max Muster
June 2018

Chapter 1

Introduction

In science, there is not only a need for accumulation of knowledge, but also for concentration. For the same reason as statistical analysis is performed on single experiments, it can also be used for results of multiple experiments or studies: to simplify and summarize the data at hand to a degree that is understandable. The latter procedure goes usually under the name of meta-analysis. Meta analyses are used to summarize results and evidence over multiple studies when they are considered to be similar enough.

In the face of the large amount of research that is done in some fields of empirical science, meta-analysis becomes increasingly important, for looking at all evidence would simply take very long for one person. Meta-analyses are often part of a systematic review, an effort of experts of a field to provide an overlook over the evidence. In the course of a review, all literature and data with respect to a scientific question is collected and a meta-analysis is operated at the end to summarize the findings.

In the case of clinical science, systematic reviews and meta-analyses do not only benefit scientists but also patients and clinicians, for both are provided with up-to-date summaries of current evidence with respect to a certain treatment. Therefore, meta-analysis is at the core of what is called evidence-based medicine.

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

1.0.1 Cochrane and the Cochrane Database of Systematic Reviews

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

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

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

Chapter 2

The Cochrane Dataset

2.0.1 Structure and Content

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

Study	Comparison_type	Outcome	Events	Total	Events_c	Total_c
Bohn 1989	Barbiturate vs no barbiturate	Death at the end of follow-up	11	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 comparisons as given in the dataset. Events denotes the count of events in the treatment group while Events c the count of events in the group compared to. Further descriptive variables have been ommitted

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

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

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

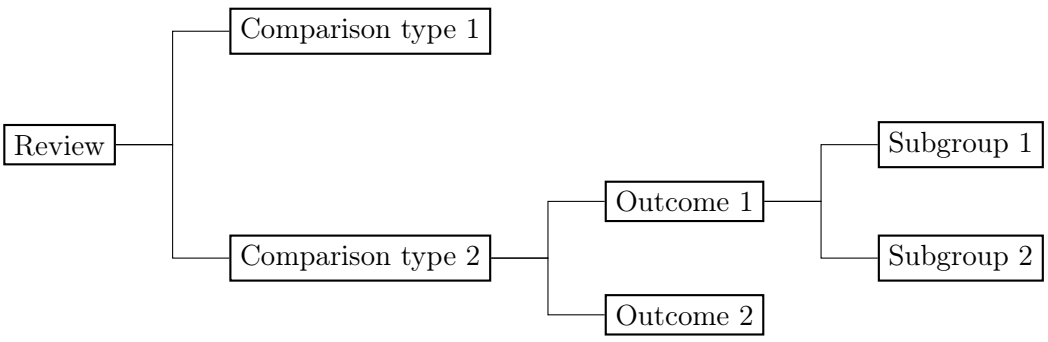

Figure 2.1: Structure of a hypothetical review with two different comparisons

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

Table 2.2: Dataset variable descriptions

Study	Comparison	Outcome
Bohn 1989	Barbiturate vs no barbiturate	Death at the end of follow-up
Bohn 1989	Barbiturate vs no barbiturate	Death or severe disability at the end of follow-up
Eisenberg 1988	Barbiturate vs no barbiturate	Death at the end of follow-up
Eisenberg 1988	Barbiturate vs no barbiturate	Uncontrolled ICP during treatment
Eisenberg 1988	Barbiturate vs no barbiturate	Hypotension during treatment
Perez-Barcena 2008	Pentobarbital vs Thiopental	Death at the end of follow-up (6 months)
Perez-Barcena 2008	Pentobarbital vs Thiopental	Death or severe disability at the end of follow-up (6 months)
Perez-Barcena 2008	Pentobarbital vs Thiopental	Uncontrolled ICP during treatment
Perez-Barcena 2008	Pentobarbital vs Thiopental	Hypotension during treatment
Schwartz 1984	Barbiturate vs Mannitol	Death at the end of follow-up (1 year)
Schwartz 1984	Barbiturate vs Mannitol	Death at the end of follow-up (1 year)
Schwartz 1984	Barbiturate vs Mannitol	Uncontrolled ICP during treatment
Ward 1985	Barbiturate vs no barbiturate	Death at the end of follow-up
Ward 1985	Barbiturate vs no barbiturate	Death or severe disability at the end of follow-up
Ward 1985	Barbiturate vs no barbiturate	Mean ICP during treatment
Ward 1985	Barbiturate vs no barbiturate	Mean arterial pressure during treatment
Ward 1985	Barbiturate vs no barbiturate	Hypotension during treatment
Ward 1985	Barbiturate vs no barbiturate	Mean body temperature during treatment

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

Having provided a rough overview over the dataset, now, some more specific information is provided. The dataset consists of 463820 comparisons and has 25 variables that specify the comparisons. The abundance of some missing values in the dataset is given in Table 2.4. For

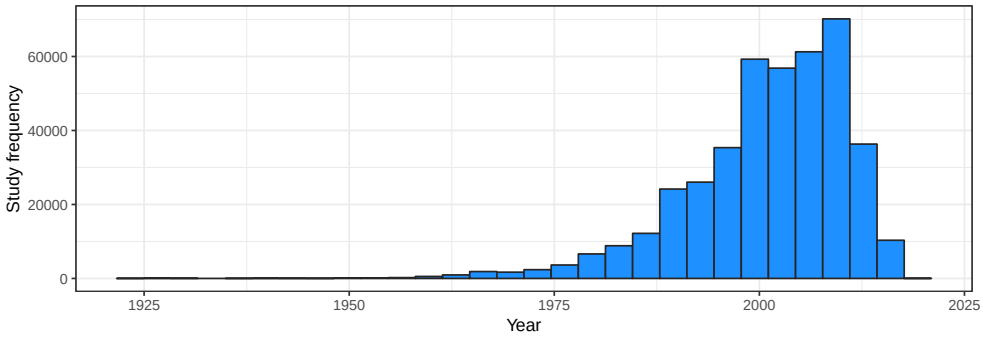

Figure 2.2: Frequencies of study publication years in the dataset. 44655 were excluded due to likely wrong indications

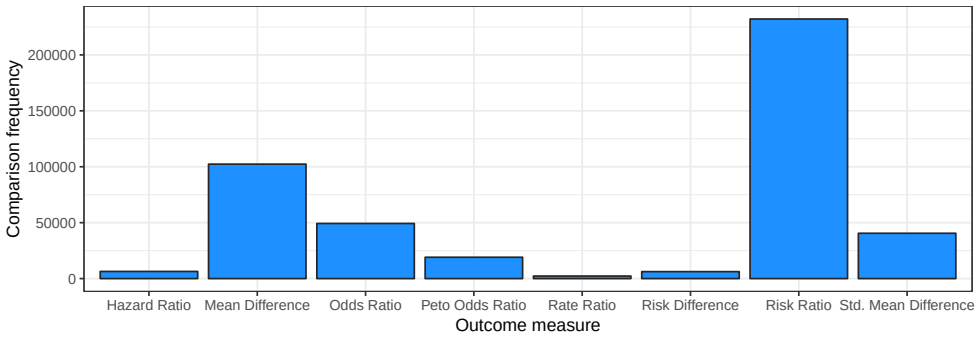

Figure 2.3: Frequencies of some outcome measures for the effects in the dataset. 5593 measures with other outcome measures are excluded

variables as research subject, outcome and subgroup name and event counts there are no missing values. The relative abundance of missing values is very low except for study years.

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

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

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

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

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

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

For comparisons to be suitable for usage in meta-analysis, they have to be somewhat identical

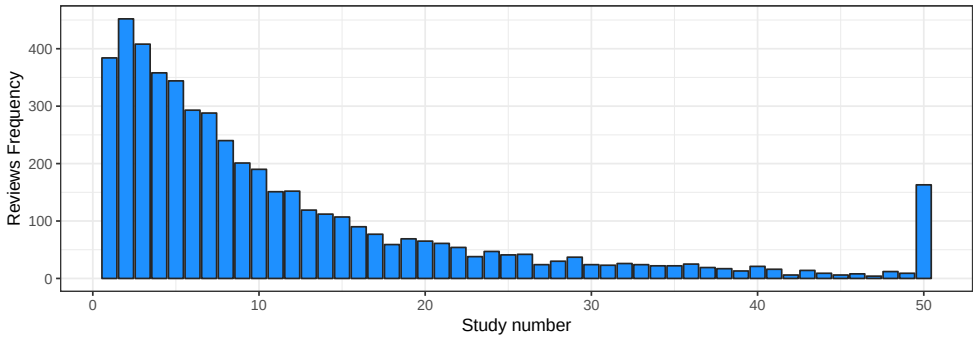

Figure 2.4: Empirical distribution of number of studies per review

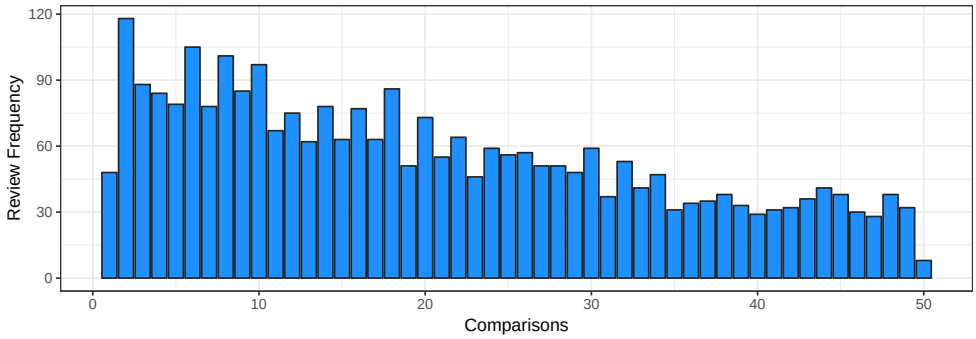

Figure 2.5: Empirical distribution of number of comparisons per review

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

In Table 2.5, two different similarity criteria have been used. One is based on the same comparison type and outcome measure, the other includes additionally subgroup affiliation of comparisons, i.e. only comparisons in the same subgroups are considered to be similar enough.

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

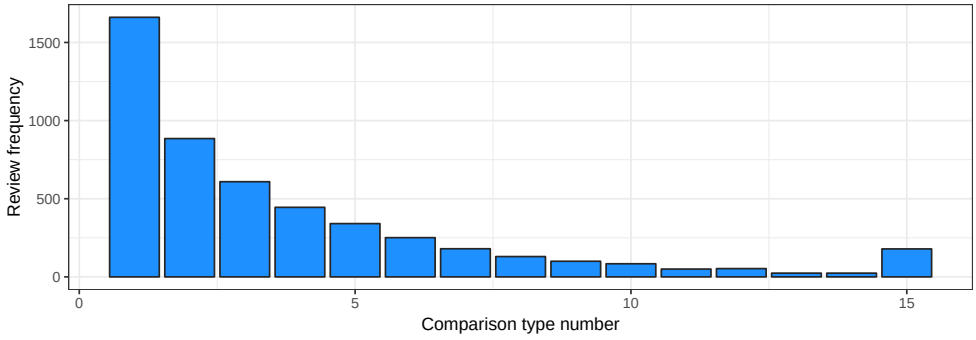

Figure 2.6: Empirical distribution of number of different research subjects per review

n	Cumulative sum (without subgroups)	Cumulative sum (with subgroups)
1	109177	186267
2	67685	83924
3	47786	52245
4	36155	36176
5	28077	26552
6	22689	20113
7	18536	15888
8	15466	12931
9	13000	10817
10	11001	9226
11	9356	7988
12	8052	7067
13	6984	6365
14	6041	5780
15	5325	5325

Table 2.5: Cumulative number of groups with number of reproduction trials >= n

Chapter 3

Results

One crucial assumption in meta analysis is that the availability and publication of studies does not depend on their effect. This is termed reporting bias in the scientific world. If reporting bias is present, the classical approaches to merge single study results in to an overall intervention effect may fail because the underlying studies are biased. There are tests that can be applied to find out if reporting bias is present in the meta analysis.

For continuous outcomes, three tests are available: Eggers (based on linear regression), Thompson and Sharp (weighted linear regression) and Begg and Mazumdar (rank based) test. The following three figures show the distribution of p-values of the corresponding tests. Note that only meta analyses with more than 10 comparisons have been included.

Since each histogram of p-values has 20 bins, the content of the bin with the smallest p-values is equal to the number of meta-analyses whose reporting bias test reports a p-value < 0.05. The fraction of those analyses in which we would reject the null-hypothesis based on the 5 % threshold can therefore be assessed by eye, and would be for example for Eggers test somewhat less than one third of all analyses.

For binary outcomes, Peters and Harbords reporting bias test have been chosen. Also here, only meta-analyses with more than 10 comparisons are included.

Another less conventional way to look for publication bias is to see how many studies are "reflected" by the trim and fill method. As a first impression, a histogram with the fraction of trimmed of all comparisons is shown in figure 3.6.

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

The same is repeated for binary outcomes in figure 3.7.

The mean fraction of trimmed comparisons for binary outcomes is 0.19 and the median 0.17.

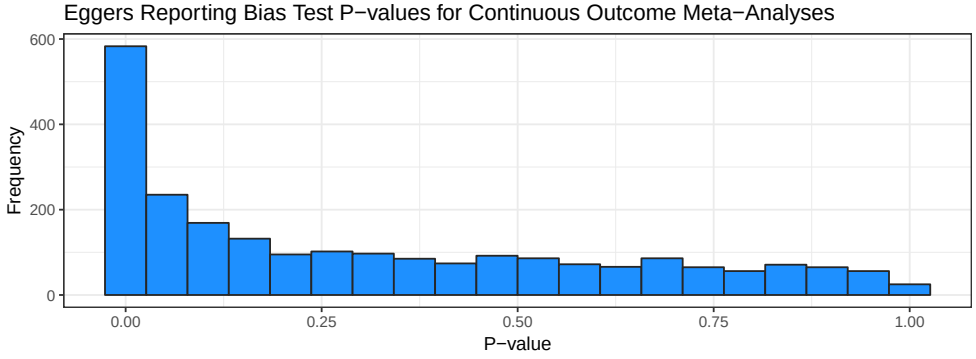

Figure 3.1: Histogram of p-values for Eggers reporting bias test (linear regression based) for continuous outcome meta analysis.

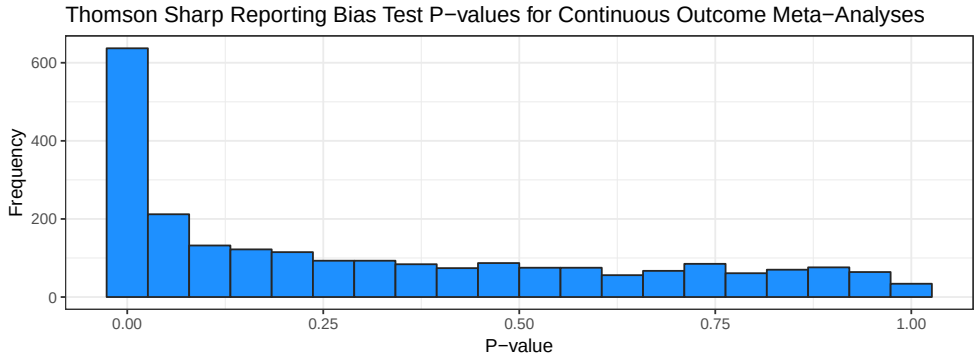

Figure 3.2: Histogram of p-values for Thompsom and Sharp reporting bias test (weighted linear regression based) for continuous outcome meta analysis.

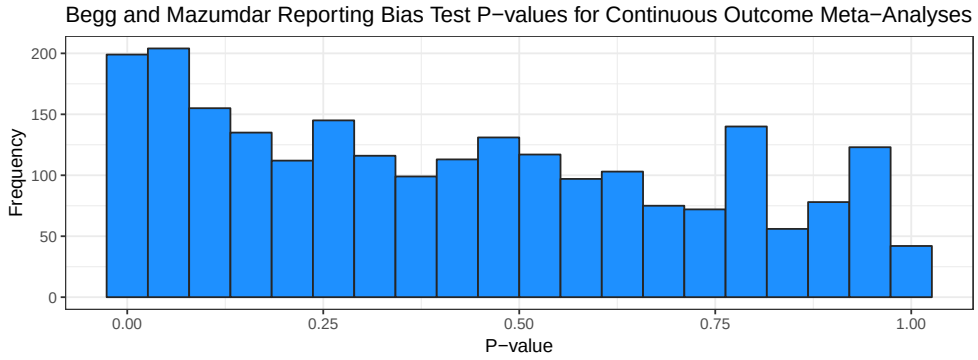

Figure 3.3: Histogram of p-values for Begg and Mazumdar reporting bias test (rank based) for continuous outcome meta analysis.

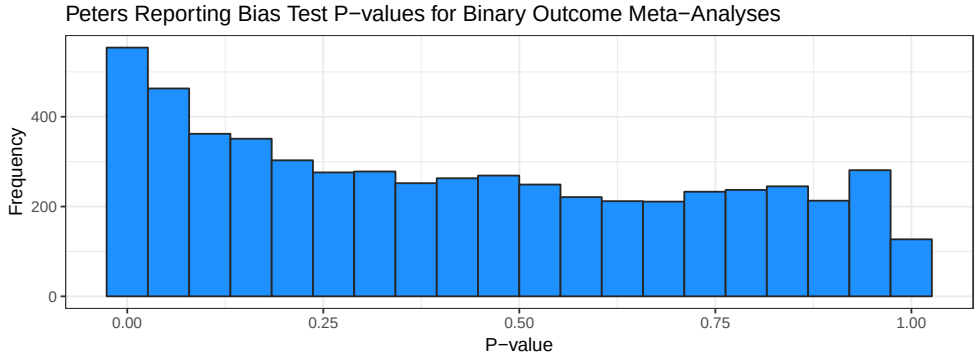

Figure 3.4: Histogram of p-values for Peters reporting bias test (rank based) for continuous outcome meta analysis.

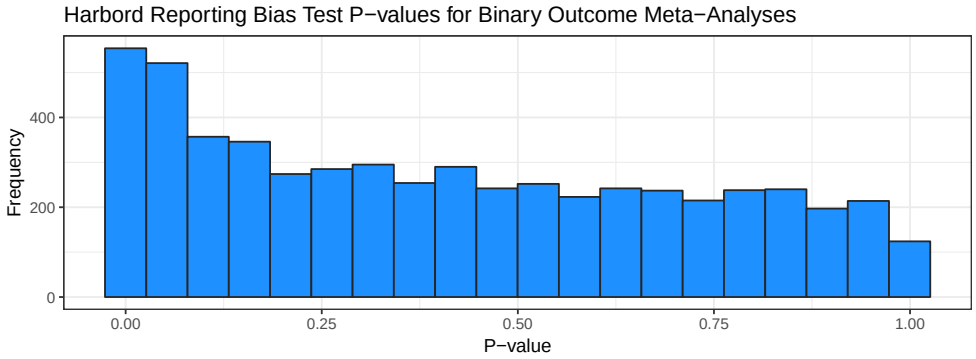

Figure 3.5: Histogram of p-values for Harbord reporting bias test (rank based) for continuous outcome meta analysis.

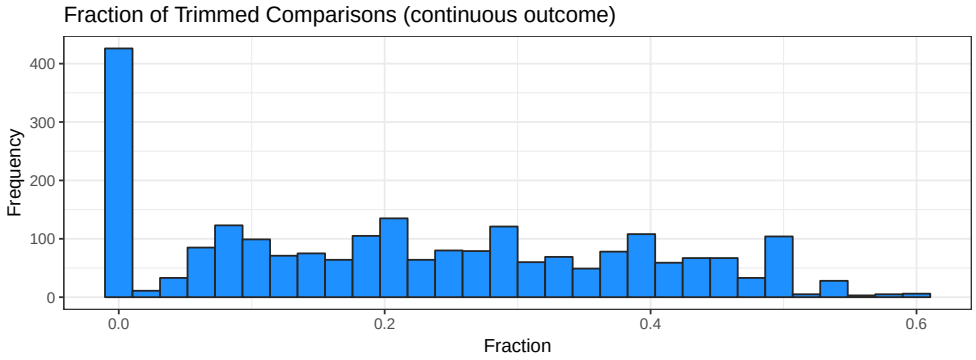

Figure 3.6: Histogram of fractions of trimmed comparisons from meta analyses with continuous outcomes.

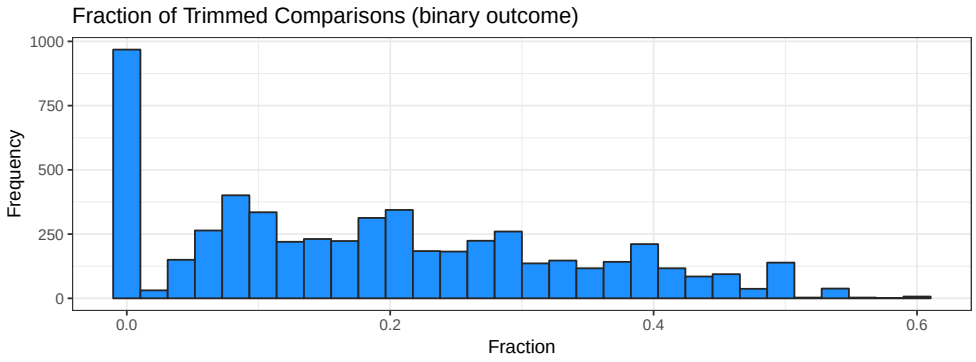

Figure 3.7: Histogram of fractions of trimmed comparisons from meta analyses with binary outcomes.

Chapter 4

Discussion and Outlook

Chapter 5

Conclusions

Appendix A

Appendix

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

Bibliography

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