Replication Report - Peters et al. 2006

Anna Lohmann[1] & Rolf H. H. Groenwold[1,2]

[1,2] Leiden University Medical Center

Abstract

Some summary.

Keywords: replication, replicability, simulation study, publication bias, meta-analysis, Egger's regression test, Peters' regression test
Word count: X

Introduction

This replication report documents the replication attempt of the simulation study Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of Two Methods to Detect Publication Bias in Meta-analysis. *JAMA*, 295(*6*), 676–678. https://doi.org/10.1001/jama.295.6.676 (Peters, Sutton, Jones, Abrams, & Rushton, 2006). Following the definition of (**???**) we understand the replication of a published study as writing and running new code based on the description provided in the original publication with the aim of obtaining the same results.

The method section details the sources of information utilized for the present replication attempt. It furthermore provides an overview of the information that was extracted from those sources as well as the technical implementation of the replication. A separate section covers all researcher degrees of freedom, i.e. decisions that had to be made by the replicators because of insufficient or contradicting information in the original sources. In the results section we present descriptive statistics from the data generating mechanism, i.e. the artificial sample. as well as the replicated results. The discussion section reflects the replication

---

Anna Lohmann, Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherland.

Rolf H.H.Groenwold, Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherland. Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.

The authors made the following contributions. Anna Lohmann: Data Curation, Formal Analysis (lead), Investigation, Software, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Rolf H. H. Groenwold: Conceptualization, Formal Analysis (supporting), Funding Acquisition, Supervision, Writing - Review & Editing.

attempt with regards to general replicability, replicator degrees of freedom and equivalence of results. The appendix contains the README file providing an overview of the accompanying code as well as additional results figures that do not correspond to the original manuscript.

Method

Information basis

The information upon which the replication is based stems from two different sources (1) the published manuscript (Peters et al., 2006) (referred to as * (original) article*) as well as (2) a technical report (Peters, Sutton, Jones, Abrams, & Rushton, 2005). The published article is fairly short and repeatedly refers to details obtainable from a technical report. The technical report is listed in the reference section of the published article, however it was not available from the public domain (i.e. online line supplements or a public online repository). The technical report was hence requested and obtained by email from the Department of Health Sciences at Leister University (hsenquiries@leicester.ac.uk).

Data Generating Mechanism

INformation provided in the above mentioned sources indicated that the following simulation factors were systematically varied in generating the artificial data.

Simulation factor	No. levels	Levels
Varied		
Publication bias	5	none, effect size based moderate (14%), effect size based severe (40%), p-value based moderate, p-value based severe
True effect size (OR)	5	1, 1.2, 1.5, 3, 5
Between Study Heterogeneity (I^2)	4	0, 20, 150, 500 (0, 16.7%, 60%, 83.3%)
Number of primary studies in meta-analysis	4	6, 16, 30, 90
Fixed		
Sample size control group		exponential of the normal distribution with a mean of 5 and variance of 0.3
Ratio treatment:control group		1:1
Probability of event in control group		sampled from unif(0.3, 0.7)

Publication bias. Five different levels of publication bias were implemented: (i) No publication bias, (ii) moderate p-value based publication bias, (iii) severe p-value based publication bias, (iv) moderate effect-size based publication bias as well as (v) severe effect-size based publication bias. Details regarding the implementation of each of these levels was obtained from the technical report (p.15, p. 20). The original article only mentions that funnel plot asymmetry was *"[. . .] induces in 2 ways. First, it was induced on the basis of the value associated with a study's effect size (the larger the P value the more likely that study was excluded from the meta-analysis).[. . .] publication bias was also induced on the*

basis of study effect size with the most extreme negative effect sizes were excluded from the meta-analysis."(p.677)

For p-value based publication bias studies are censored as a result of the one-sided p-value associated with the effect estimate of interest. The p-value based selection probabilities is given in the following table which was taken from the technical report (table 2, report p. 15).

Severity of publication bias	p-value	Selection Probability
Moderate	<0.05	1
	0.05 - 0.2	0.75
	0.2-0.5	0.5
	>0.5	0.25
Severe	<0.05	1
	0.05-0.2	0.75
	>0.2	0.25

For publication bias induced by effect size, a given percentage of studies with the most extreme effect estimates of effect are censored. For moderate publication bias this percentage corresponds to 14% and for severe publication bias 40% of studies.

Primary studies are generated following either a true or a random effects model depending on the heterogeneity of a given scenario.

True effect. Fixed effects model is given by $y_i = \theta + \epsilon_i$ where θ is the true underlying effect lnOR

Random effects model $y_i = \theta_i + \epsilon_i$ with θ_i $N(\mu, \tau^2)$ where θ_i is the true effect in study i μ true underlying effect lnOR τ^2 is the between-study variance

Between Study Heterogeneity. *"[B]etween-study variance is defined to be 20%, 150, and 500% of the average within-study variance for studies from the corresponding simulations. This compares with specifications of I^2, describing the percentage of total variation across studies that is due to between-study heterogeneity rather than chance (ref 25). Here 20%, 150% and 500 % of the within-study variation corresponds to an I^2 of 16.7%, 60% and 83% respectively."* (technical report p. 12)

Size of control group. The number of participants in the control group of a given single study is sampled from the exponential of a normal distribution with a mean of 5 and a variance of 0.3 (article p. 677). The technical report gives $N(5, 0.3)$ as the distribution from which the number of control groups within each primary study is taken (report p.15). As this does not make much sense we interpreted it as a typo and followed the information from the original article.

Repetitions. Each of the 400 unique scenarios were repeated 1000 times.

Data generating process. Figure 1 provides a flow chart with an overview of the data generating process.

Data generation can be summarized with the following pseudo code:

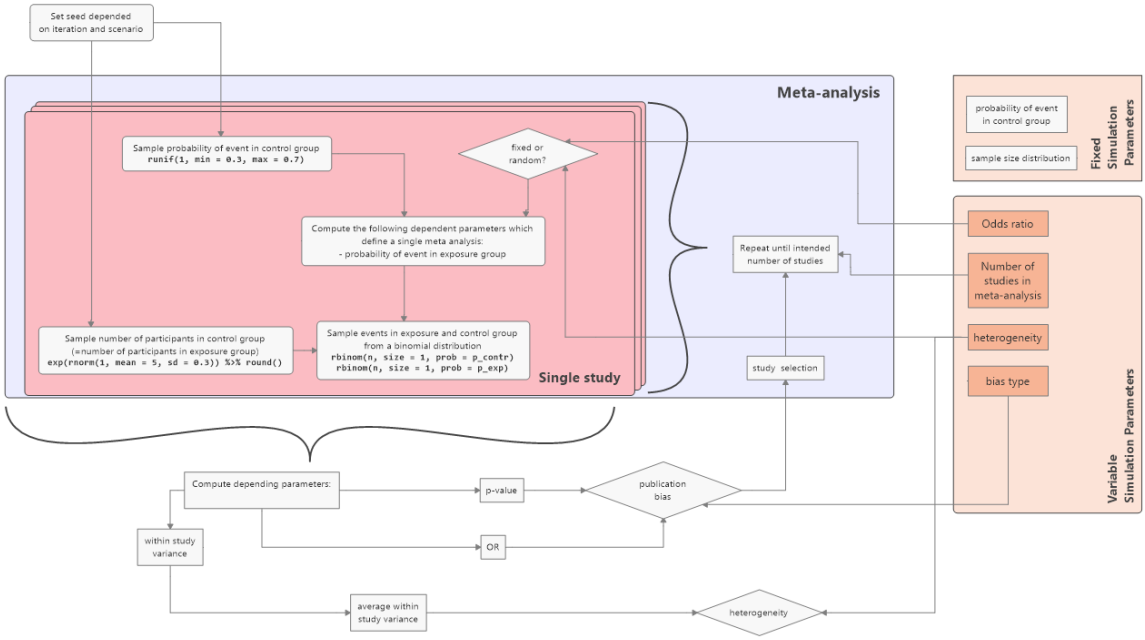


Figure 1. Flow chart of data generating mechanism

For 1000 repetitions of each of 400 unique scenarios:

- Set unique seed based on scenario id and number of repetition.
- Simulate an unbiased study set based on the control group event probability.
- While number of selected studies < number of required studies for given scenario:
  - Sample an event probability for the control group from the given distribution.
  - Sample a sample size from the given distribution.
  - Compute the remaining study characteristics based on these random elements.
  - Determine selection indicator based on publication bias mechanism of current scenario.
- If heterogeneity of present scenario is > 1: If heterogeneity of present scenario is > 1:
  - Determine original between-study variability & resample from corresponding random effects model.
- Apply publication bias.

Compared Methods

The study compares two regression tests for the detection of publication bias in meta-analyses. The first Egger's regression test regresses the standardized effect estimate on a measure of precision (see equation 1). The second (referred to as "alternative regression test in the original article) is a variation of Macaskill's test and regresses the effect size on the inverse of the total sample size (see equation 2). We refer to this test as Peters' regression test in the present report.

Egger's regression test. $\frac{y_i}{se_i} = \beta + \frac{\alpha}{se_i} + \epsilon_i$ which is equivalent to $y_i = \alpha + \beta \cdot se_i + \epsilon_i \cdot se_i$ weighted by $\frac{1}{se_i^2}$ where y_i is the lnOR from study i and se_i is the standard error of y_i (report p. 22)

Peters' regression test. $y_i = \alpha + \frac{\beta}{size_i} + \epsilon_i$ weighted by $(\frac{1}{A+B} + \frac{1}{C+D})^{-1}$ (report p.23)

Performance measures

The original study compares the performance of the two regression tests on the basis of type I error rates and power. The type 1 error rate (proportion of false positives) is defined as statistical significance specified from a 2-tailed test at $p < .10$ (article p. 678).

Technical implementation

While the original simulation study was carried out in STATA 8.2 our replication was implemented using the R programming environment (details regarding software versions can be obtained from the section Reproducibility Information). The corresponding R code can be obtained from https://github.com/replisims/peters-2016.

Replicator degrees of freedom

The following table provides an overview of replicator degrees of freedom, i.e. decisions that had to be made by the replicators because of insufficient or contradicting information. Issues were resolved by discussion among the replicators. Decisions were based on what the replicators perceived to be the most likely implementation with likeliness estimated by common practice and/or guideline recommendations. Whereever feasible multiple interpretations where implemented.

Issue	Replicator decision	Justification
Dealing with empty cells	add 0.5 to every empty cell	Common practice
Which set to compute average within-study-variance on	largest number of studies generated before application of publication bias	Most accurate correspondence to intended I^2
Data dependence	each scenario is implemented in independently generated data	Best practice (Burton, Altman, Royston, & Holder, 2006)
Fixed probability of event in control-group for all studies in one MA	Probability of event in CG assumed as fixed	All replicators tended towards that interpretation

Issue	Replicator decision	Justification
Exact scenario depicted in results	Present multiple result subsetting	Easy to implement alternative interpretations

Empty cells

In the simulation of individual studies it is possible (albeit unlikely) for either the exposed group or the control group to not have any events (or no non-events). This would make it impossible to compute the necessary parameters to continue the simulation. Neither the original article nor the technical report provide any information whether such a case ever occurred and if so how it was dealt with. We implemented the replication such that 0.5 is added to empty cells.

Simulation of between-study heterogeneity

The between-study heterogeneity parameter was calculated as a percentage of the average within-study variance estimate. From the fixed- effects model, the average within study variance was calculated and between-study heterogeneity was then defined to be 20%, 150% and 500% of the within-study estimate.

It is unclear from which fixed-effects model the average within study variance was calculated (e.g. what was the number of studies). We interpreted this passage to refer to the fixed-effects model with the same parameter constellation (i.e. the same OR and number of studies). The random effects model was then redrawn after the average within-study variance was obtained. It is furthermore unclear whether the average within-study variance was obtained before or after the application of publication bias. In our replication we assumed it was before.

Publication bias based on effect size

The technical report describes the effect size based publication bias as follows: *"either 14% or 40 % of the most extreme studies showing a negative effect of the exposure (i.e. OR <1) were censored such that the final number of studies in a meta -analysis was still 6, 16, 30, or 90 i.e. for the 6 studies 10 haven been generated and 4 studies with the most extreme negative estimates have been censored"* The article similarly states *"Studies with the most extreme negative effect sizes were excluded from the meta-analysis"* (p.677).

These statements are contradictory. On the one hand it is suggested, that extreme studies with a negative effect of the exposure should be censored. On the other hand it suggests to censor either 14% or 40% of studies. Especially with large effect sized (e.g. an OR of 5) it is highly unlikely to find 40% of studies with a negative effect of the exposure. We hence ignored the "negative effect" aspect and interpreted the authors' intention to be the exclusion of 14% and 40% of studies irrespective of the sign of the effect.

Data dependence

The number of studies in the meta-analysis is one of the simulation parameters. As in every simulation study with sample size as one of the varied parameters the question arises whether the samples are dependent (i.e. smaller samples being subsamples of larger samples) or independent (i.e. each scenario independently sampled). We could not find information regarding this in neither the original article nor the technical report. We hence implemented each scenario as an independent sampling which is in line with current best practice recommendations (Burton et al., 2006).

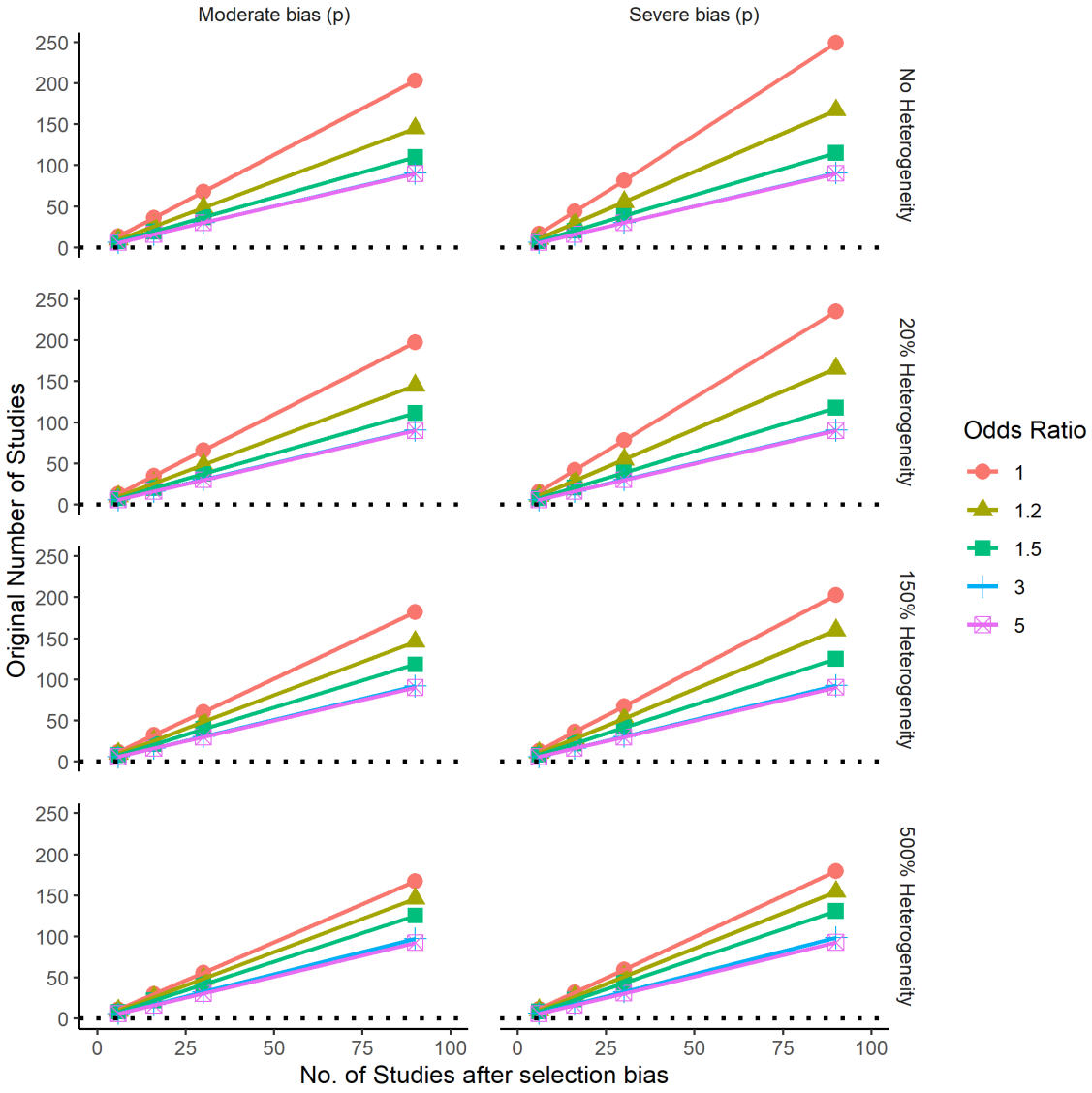
Exact scenario depicted in results

The results are stratified by heterogeneity and method of inducing publication bias. The plots are labeled with *"publication bias induced by p-value"* and *"publication bias induced by effect size"* correspondingly. This suggests that results were collapsed over the severity of bias. However, the following quote makes it seem like the results might only be depicted for severe bias: *"Power to detect "moderate" publication bias is lower than that to detect "severe" publication bias for all models. However, the same general trend in power is seen for "moderate" publication bias as the level of between-study heterogeneity increases, as it is for severe publication bias (results not shown).* It is unclear whether *"(results not shown)"* refers to the entirety of this passage (i.e. the comparison of moderate vs severe bias is not shown) or whether it means to imply that all results depicted pertain to severe publication bias. An additional passage suggests that indeed only severe publication bias might be plotted: *"Model 4c has relatively good power to detect severe publication bias when there is no between-study heterogeneity compared to the other models, regardless of how publication bias is induced (Figures 10 and 11)"* (p.31-32 technical report). The article does not make a distinction between moderate and severe publication bias. We resolve this problem by presenting figures for severe publication bias which seems the most likely choice based on the quotes provided above. Additional figures for both collapsed publication bias as well as moderate publication bias are presented in the Appendix.

Results

Simulation descriptives

The technical report presents a figure showing the mean number of studies generated to obtain the number of studies required under "moderate" and "severe" bias (figure 2, p.17) This closely corresponds to the numbers found in the replication.



As detailed above the exact procedure to simulate the heterogeneity was not sufficiently described in neither the article nor the technical report. The following figure shows the intended I^2 compared to the observed I^2 for a given meta-analysis both before as well as after the application of publication bias.

Replication of result figures

The following table provides an overview of figures related to central outcomes of the original study.

Overview of result figures in the original artical and the technical report

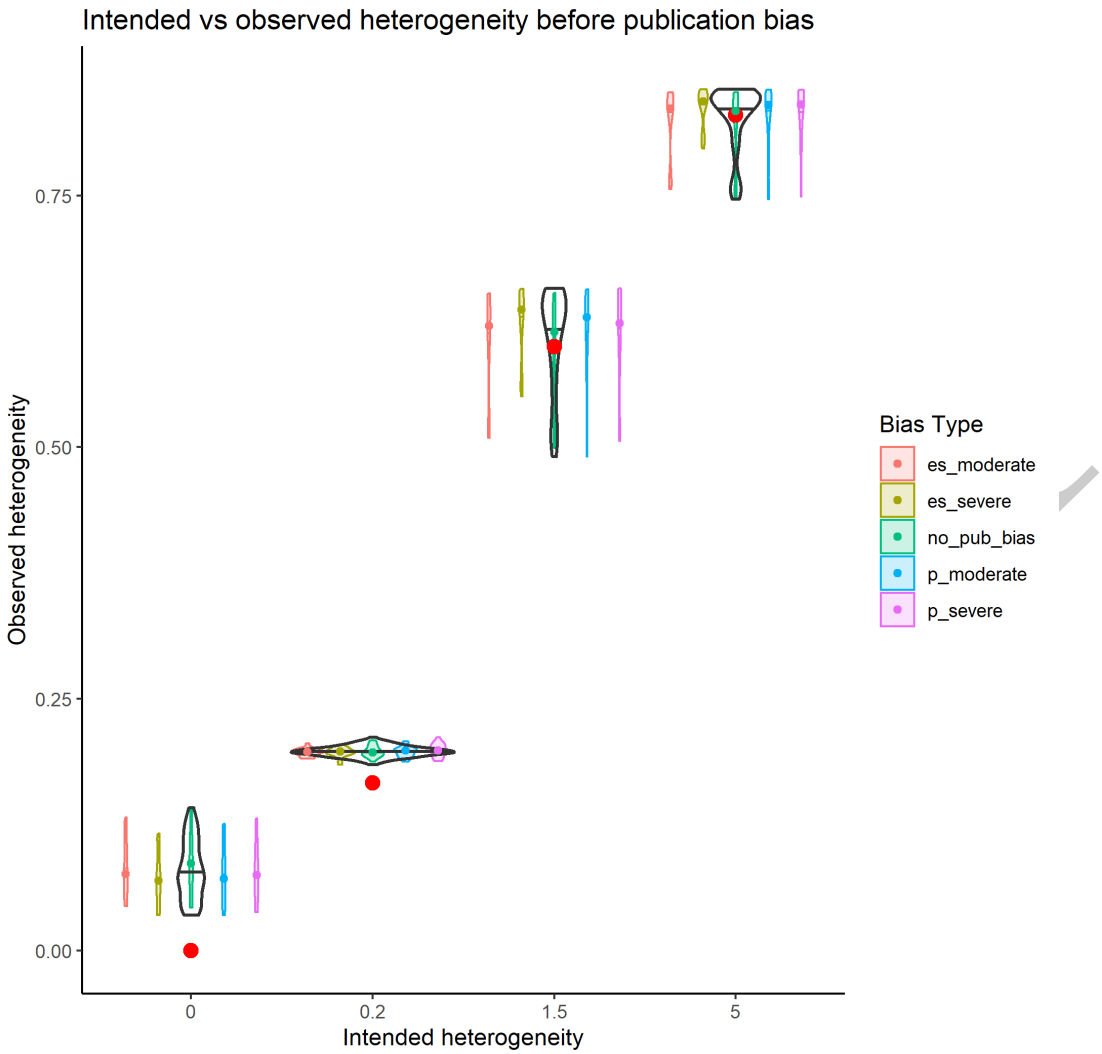


Figure 2. I^2 before publication bias

Performance measure	Test	Scenario
Type I error rate	Egger's regression test	all
Type I error rate	Peter's regression test	no heterogeneity
Type I error rate	Peter's regression test	500% heterogeneity
Power	Both	(severe ?) publication bias induced by effect size, no heterogeneity
Power	Both	(severe ?) publication bias induced by p-value, no heterogeneity

Performance measure	Test	Scenario
Power	Both	(severe ?) publication bias induced by p-value, OR 1,1.5 & 5, all heterogeneity levels
Power	Both	(severe?) publication bias induced by effect size, 500% heterogeneity
Power	Both	(severe?) publication bias induced by p-value, 500% heterogeneity

Replication of results presented in text form

"However, in Model 4c [which corresponds to the alternative test] we have a test that is superior to Egger's test in terms of expected type I error, but which also has good power to detect publication bias (equal to that of Egger's test), regardless of the amount of between-study heterogeneity." (report p. 46)

The superiority in terms of type one error rates can be clearly seen in our replication. While Peter's and colleagues do not specify their definition of "equal", the power of Peter's test by no means seems equal to Egger's regression test.

"When there is little between-study heterogeneity, the alternative regression test and Egger's regression test appear to have moderate power to detect asymmetry when it is induced on the basis of P value (Figure 3) and high power when asymmetry is induced on the magnitude of the effect (data not shown)." (article page 679)

While the authors do not define their definition of "moderate power" and "high power" we would not use these attributes to describe the results of our replication, at least not in respect to Peter's test. The wording furthermore seems to imply that the power of both tests is of a comparable magnitude which clearly does not seem to be the case in our replication.

Discussion

Replicability

The original JAMA article is merely five pages long (including figures and references). Consequently the article repeatedly refers to the technical report for details. A total of 10 references to the technical report illustrate that replication merely based on the article would not have been feasible. Crutial implementation details such as details about how publication bias was induced are missing from the published manuscript. While unfortunately not available in the public domain we were able to obtain the technical report from the corresponding institution. Fourteen years after publication this is commendable on parts of the department. The technical report, which comprises 57 pages, does contain sufficient information for a replication. A few details were, however, described in insufficient detail or using language that leaves room for interpretation on the part of the replicators.

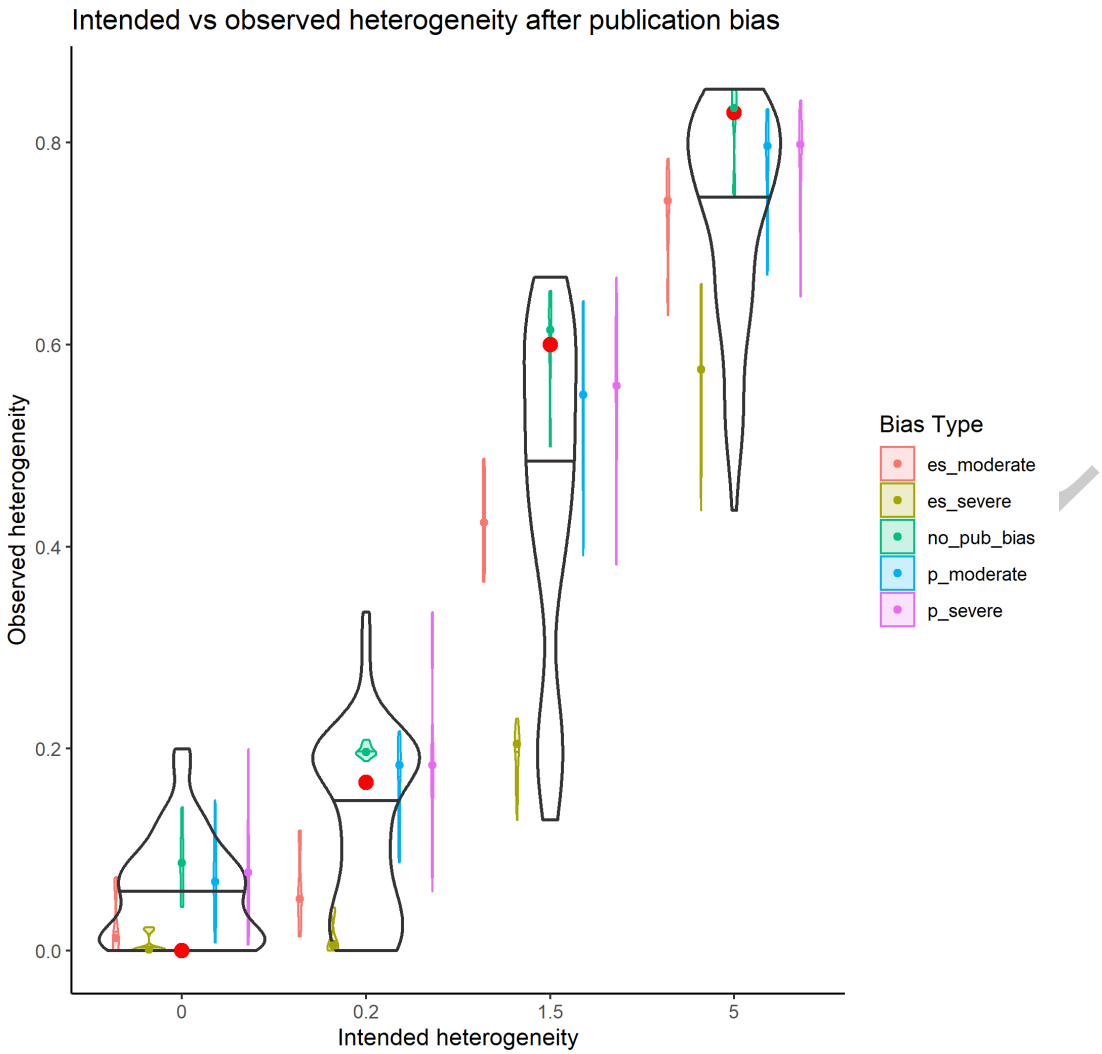


Figure 3. I^2 after publication bias

Replicator degrees of freedom

The information that we found to be ambiguous or insufficient has to be interpreted from our unique perspective. Other researchers with a different background (i.e. different expertise, different language context, different point in history) might have encountered other uncertainties. Furthermore, the fact that we found a certain aspect to be "clear" does not ensure correctness of our interpretation neither does finding something unclear automatically imply that our implementation was wrong. Most replicator degrees of freedom concern the data generation. Only one uncertainty is related to the scenarios the result presentation was based on. The latter can easily be addressed by presenting the results in a multitude of ways. Variation in the data generation is more complex to address. In addition to the complexity of implementing several different interpretations (and their combinations) each full simulation run takes considerable computational resources. The present simulation was

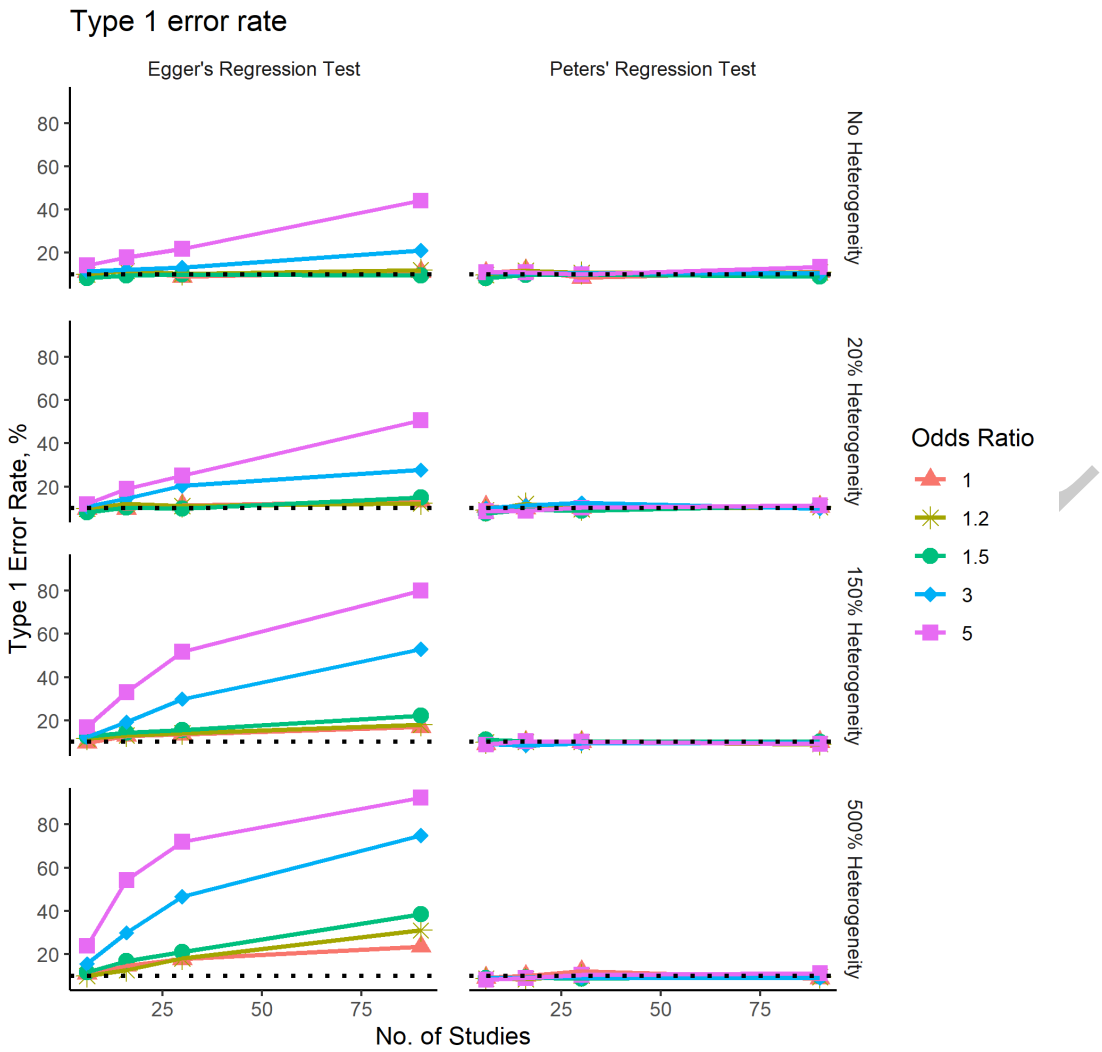


Figure 4. Type I error rate

comparably "cheap" with about eight hours of computation time. The present replication hence only includes our "best guess" of what might have been the original authors intentions.

Equivalence of results

The mean number of studies generated per scenario seems to match that of the original study (compare figure X of this report and figure 2 of the technical report). This indicates that we were able to replicate the data generation based on the p-value based publications bias. Overall results regarding power and type one error match the original article. Due to the above mentioned imprecisions regarding the bias type displayed in certain result figures it is hard to tell which figures should correspond to each other.

Acknowledgments

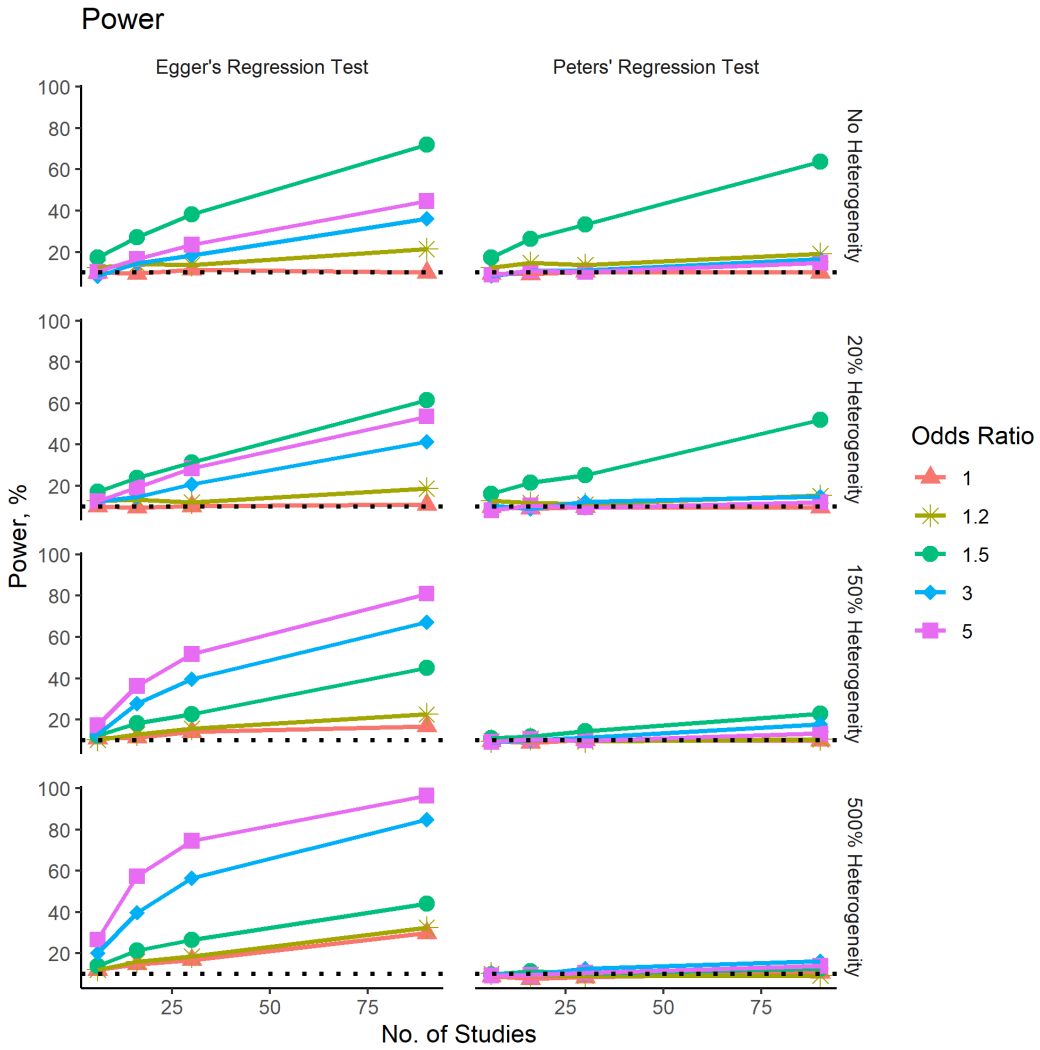


Figure 5. Power to detect severe publication bias induced by p-value

References

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292. https://doi.org/10.1002/sim.2673

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2005). *Performance of tests and adjustments for publication bias in the presence of heterogeneity* (Technical Report No. 05-01) (pp. 1–57). Leicester: Department of Health Sciences, University of Leicester.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of Two Methods to Detect Publication Bias in Meta-analysis. *JAMA*, *295*(6), 676–678. https://doi.org/10.1001/jama.295.6.676

Appendix

Additonal result plots

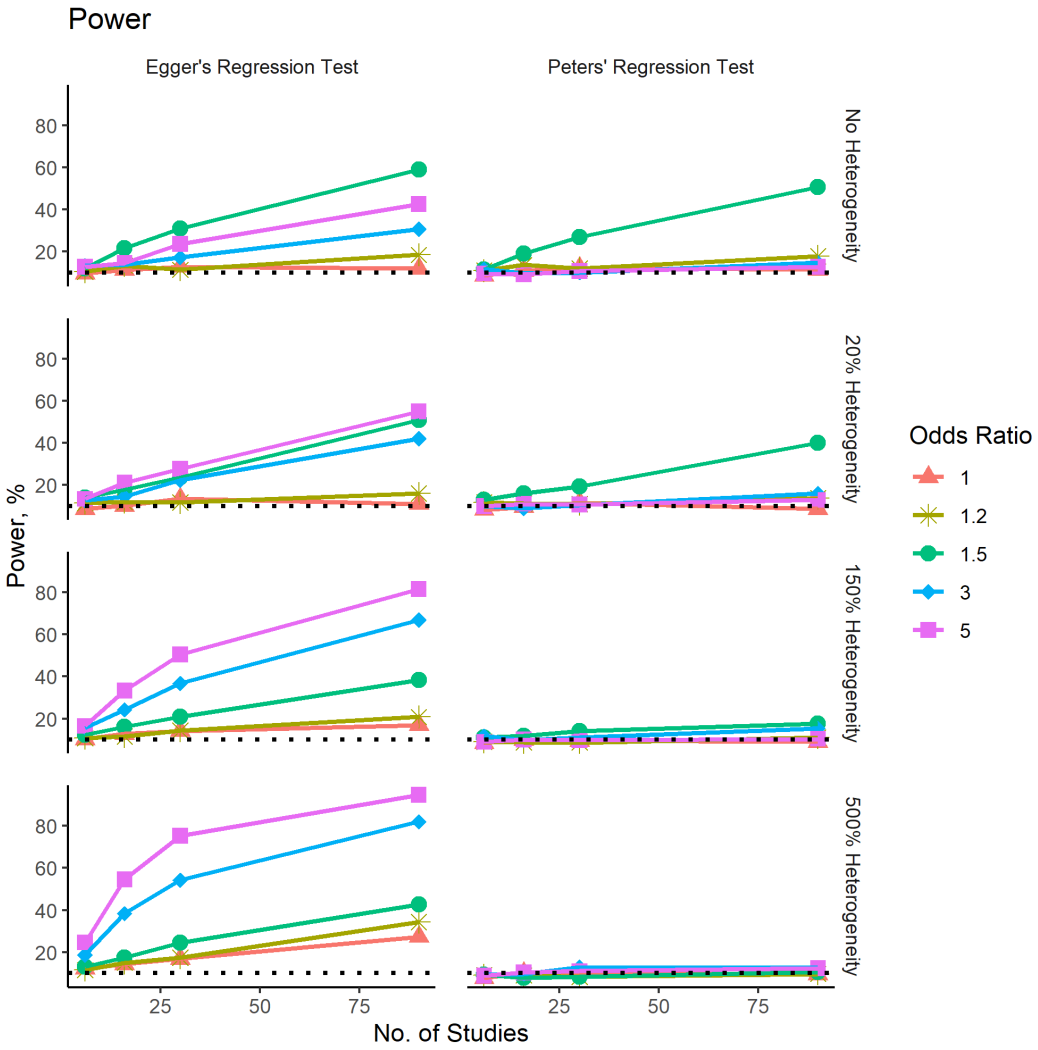


Figure 6. Power to detect severe publication bias induced by p-value

Code organization

The code and the files associated are organized in the form of a reseatch compendium which can be found in the following git repository `https://github.com/replisims/peters-2016`

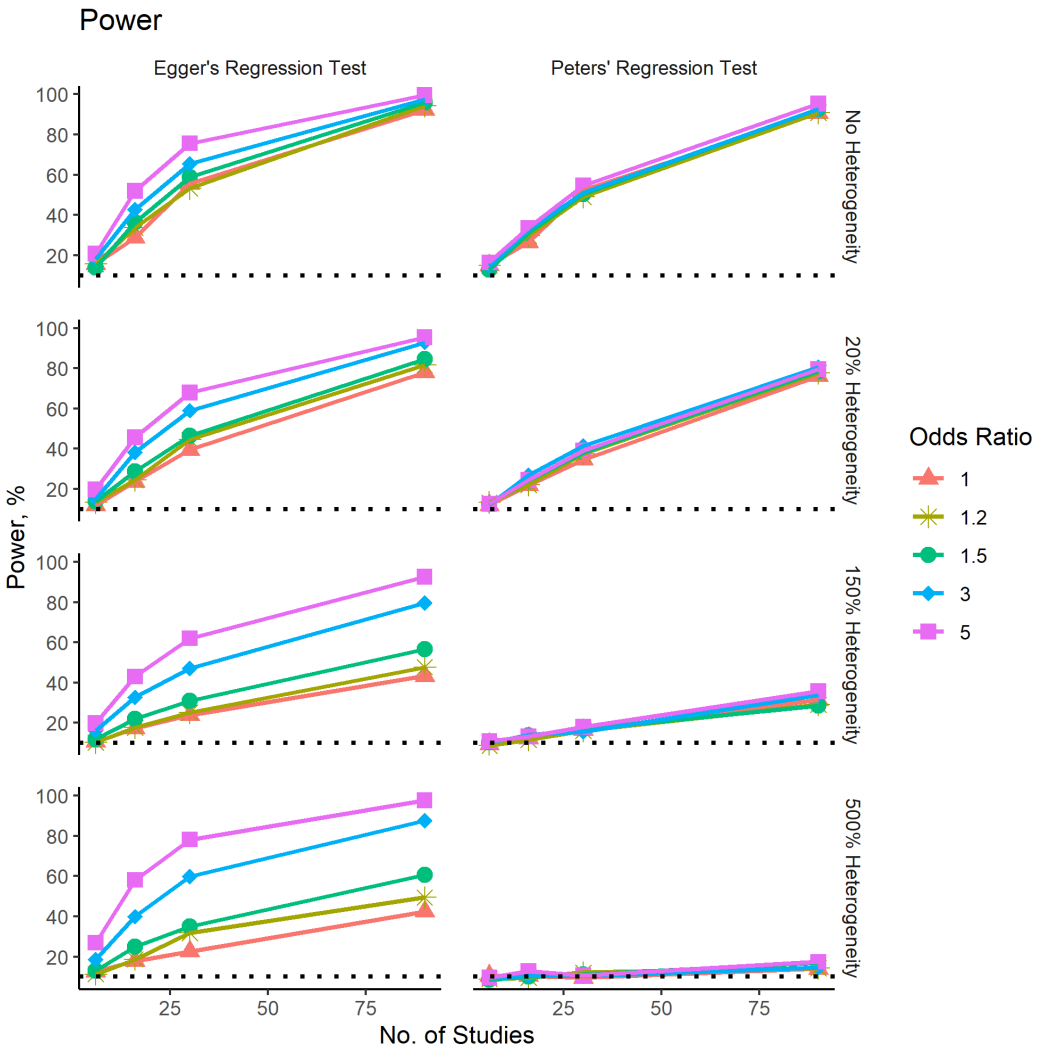


Figure 7. Power to detect severe publication bias induced by p-value

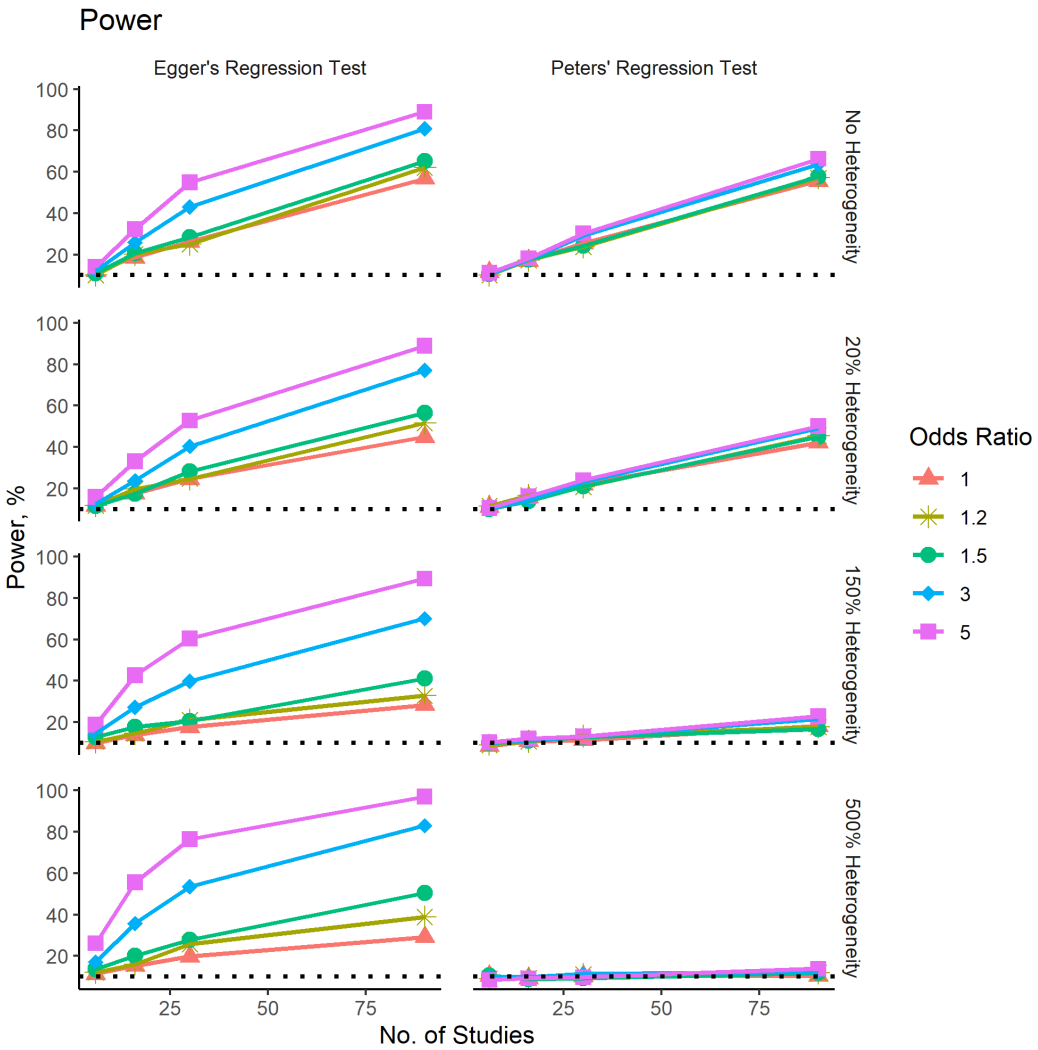


Figure 8. Power to detect severe publication bias induced by p-value

```
peters2006
├── analysis
│   ├── analysis.R
│   ├── data
│   │   ├── derived_data
│   │   ├── DO-NOT-EDIT-ANY-FILES-IN-HERE-BY-HAND
│   │   ├── raw_data
│   │   │   ├── results_data.rds
│   │   │   ├── results_h0_false.rds
│   │   │   ├── results_h0_true.rds
│   │   │   ├── sim_data10.rds
│   │   │   ├── sim_data1.rds
│   │   │   ├── sim_data2.rds
│   │   │   ├── sim_data3.rds
│   │   │   ├── sim_data4.rds
│   │   │   ├── sim_data5.rds
│   │   │   ├── sim_data6.rds
│   │   │   ├── sim_data7.rds
│   │   │   ├── sim_data8.rds
│   │   │   └── sim_data9.rds
│   ├── descriptives.R
│   ├── figures
│   ├── import_data.R
│   ├── main.R
│   ├── nested_loop_plot.R
│   ├── plotting_call.R
│   ├── pretty_plots.R
│   ├── report
│   └── shiny
├── data
│   └── scenarios.rda
├── DESCRIPTION
├── man
├── NAMESPACE
├── peters2006.Rproj
├── R
├── README.md
├── README.Rmd
└── references.bib
```

-
- data: contains the package data in this case a data frame with the original simulation scenarios
- man: contains the documentation for all package functions
- analysis: contains code to run the simulation

- 01_data_generation.R run the simulation with the original simulation parameters
- 02_publication_bias_testing.R apply Egger and Peters regression test to generated data
- 03_performance plots obtain plots for type 1 error rate and power
- 04_primary_study_selection_plot plot the number of primary studies generated before publication bias
- 05_heterogeneity_plots plot intended vs observed heterogeneity under various publication bias scenarios

Reproducibility Information. This report was last updated on 2020-09-17 12:17:45. The simulation replication was conducted using the following computational environment and dependencies:

```
#> - Session info ---------------------------------------------------------------
#>  setting  value
#>  version  R version 3.6.3 (2020-02-29)
#>  os       Windows 10 x64
#>  system   x86_64, mingw32
#>  ui       RTerm
#>  language (EN)
#>  collate  English_United States.1252
#>  ctype    English_United States.1252
#>  tz       Europe/Berlin
#>  date     2020-09-17
#>
#> - Packages -------------------------------------------------------------------
#>  package     * version    date       lib source
#>  assertthat    0.2.1      2019-03-21 [1] CRAN (R 3.6.3)
#>  backports     1.1.6      2020-04-05 [1] CRAN (R 3.6.3)
#>  bookdown      0.20       2020-06-23 [1] CRAN (R 3.6.3)
#>  callr         3.4.4      2020-09-07 [1] CRAN (R 3.6.3)
#>  cli           2.0.2      2020-02-28 [1] CRAN (R 3.6.3)
#>  crayon        1.3.4      2017-09-16 [1] CRAN (R 3.6.3)
#>  desc          1.2.0      2018-05-01 [1] CRAN (R 3.6.3)
#>  devtools      2.3.1      2020-07-21 [1] CRAN (R 3.6.3)
#>  digest        0.6.25     2020-02-23 [1] CRAN (R 3.6.3)
#>  ellipsis      0.3.1      2020-05-15 [1] CRAN (R 3.6.3)
#>  evaluate      0.14       2019-05-28 [1] CRAN (R 3.6.3)
#>  fansi         0.4.1      2020-01-08 [1] CRAN (R 3.6.3)
#>  fs            1.4.1      2020-04-04 [1] CRAN (R 3.6.3)
#>  glue          1.4.1      2020-05-13 [1] CRAN (R 3.6.3)
#>  htmltools     0.5.0      2020-06-16 [1] CRAN (R 3.6.3)
#>  knitr         1.29       2020-06-23 [1] CRAN (R 3.6.3)
#>  magrittr      1.5        2014-11-22 [1] CRAN (R 3.6.3)
#>  memoise       1.1.0      2017-04-21 [1] CRAN (R 3.6.3)
#>  papaja      * 0.1.0.9997 2020-08-25 [1] Github (crsh/papaja@0457653)
```

```
#>  pkgbuild      1.1.0      2020-07-13 [1] CRAN (R 3.6.3)
#>  pkgload       1.0.2      2018-10-29 [1] CRAN (R 3.6.3)
#>  prettyunits   1.1.1      2020-01-24 [1] CRAN (R 3.6.3)
#>  processx      3.4.2      2020-02-09 [1] CRAN (R 3.6.3)
#>  ps            1.3.2      2020-02-13 [1] CRAN (R 3.6.3)
#>  R6            2.4.1      2019-11-12 [1] CRAN (R 3.6.3)
#>  remotes       2.2.0      2020-07-21 [1] CRAN (R 3.6.3)
#>  rlang         0.4.6      2020-05-02 [1] CRAN (R 3.6.3)
#>  rmarkdown     2.3        2020-06-18 [1] CRAN (R 3.6.3)
#>  rprojroot     1.3-2      2018-01-03 [1] CRAN (R 3.6.3)
#>  sessioninfo   1.1.1      2018-11-05 [1] CRAN (R 3.6.3)
#>  stringi       1.4.6      2020-02-17 [1] CRAN (R 3.6.2)
#>  stringr       1.4.0      2019-02-10 [1] CRAN (R 3.6.3)
#>  testthat      2.3.2      2020-03-02 [1] CRAN (R 3.6.3)
#>  usethis       1.6.1      2020-04-29 [1] CRAN (R 3.6.3)
#>  withr         2.2.0      2020-04-20 [1] CRAN (R 3.6.3)
#>  xfun          0.15       2020-06-21 [1] CRAN (R 3.6.3)
#>  yaml          2.2.1      2020-02-01 [1] CRAN (R 3.6.2)
#>
#> [1] C:/Users/Anna/Documents/R/win-library/3.6
#> [2] C:/Program Files/R/R-3.6.3/library
```

The current Git commit details are:

```
#> Local:    wip C:/Users/Anna/Dropbox/anna/projects/replisims/peters2006
#> Remote:   wip @ origin (https://github.com/replisims/peters-2016.git)
#> Head:     [1f8af7d] 2020-09-11: Remove temp files
```