universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

p-hacked homeopathy? Conduct and comparison of p-value based effect size estimation and conventional meta-analyses on individualized homeopathic treatment studies

verfasst von / submitted by

Nicole Prochaska BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Psychologie
Betreut von / Supervisor:	Mag. Dr. Jakob Pietschnig, FHEA
Mitbetreut von / Co-Supervisor:	

Acknowledgments

First of all I want to thank Dr. Jakob Pietschnig for his valuable supervision, competent suggestions and improvements on this work throughout all its stages – and for providing me with such an awesome research topic in the first place.

I also want to thank my fellow students in the master's class for making my last year at university as fantastic as it has been. Thank you all for making me laugh even when – or actually all the more when I was stuck with my thesis. Additionally, thanks to my test consultant and test developer colleagues at work for the input on this thesis, and for the cubicle. My cubicle is awesome.

Another big part of my gratitude belongs to Astrid Pichler for being the single most amazing comrade throughout the master curriculum.

Special thanks go to Ben Winter for his perpetual emotional support in every facet of my life and his imperturbable calm when I launched into my passionate monologues yet again, and of course for proofreading this thesis.

Last – but by no means least – I want to thank Stefan Uttenthaler for sharing his knowledge and expertise with me whenever it was needed and even when it was not, and for being such a wonderful, supportive boyfriend.

Abstract

Studies reporting statistically significant results are more likely to get published while most null-hypothesis-confirming results remain unpublished. This problem is known as publication bias. Because the success of researchers is often measured by their number of publications, various strategies to maximize the chance of publishing a study are regularly applied. Such strategies are subsumed under the term "questionable research practices" (QRPs) and range from harmless behaviours to outright fraud. Both publication bias and QRPs threaten the validity of empirical research in general and meta-analyses in particular. Two recently developed methods, *p*-curve and *p*-uniform, aim to deal with these problems by only considering statistically significant *p*-values. Here, both methods were applied on real data. Randomized placebo-controlled clinical trials on individualized homeopathic treatments were subjected to *p*-curve and *p*-uniform as well as conventional meta-analysis. Conventional meta-analysis suggests a small effect in favour of homeopathy, while no effect was found with *p*-value-based methods. No publication bias was detected, however, questionable research practices seem to distort findings by conventional meta-analysis. Subset analysis based on journal affiliation with homeopathy revealed distorting characteristics of studies published in papers affiliated with homeopathy, leading to significant effect size overestimation when conducting conventional meta-analysis. From these results, two conclusions can be drawn. On the one hand, *p*-value based methods yielded consistent results with each other and are relevant alternatives to conventional meta-analyses. On the other hand, there is no evidence that individualized homeopathic treatments have an effect beyond the placebo effect. Observed symptom reduction in or increased well-being of patients are most likely due to a placebo effect.

Keywords: homeopathy, meta-analysis, *p*-curve, *p*-uniform, questionable research practices, p-hacking, publication bias

Abstract

Studies reporting statistically significant results are more likely to get published while most null-hypothesis-confirming results remain unpublished. This problem is known as publication bias. Because the success of researchers is often measured by their number of publications, various strategies to maximize the chance of publishing a study are regularly applied. Such strategies are subsumed under the term "questionable research practices" (QRPs) and range from harmless behaviours to outright fraud. Both publication bias and QRPs threaten the validity of empirical research in general and meta-analyses in particular. Two recently developed methods, *p*-curve and *p*-uniform, aim to deal with these problems by only considering statistically significant *p*-values. Here, both methods were applied on real data. Randomized placebo-controlled clinical trials on individualized homeopathic treatments were subjected to *p*-curve and *p*-uniform as well as conventional meta-analysis. Conventional meta-analysis suggests a small effect in favour of homeopathy, while no effect was found with *p*-value-based methods. No publication bias was detected, however, questionable research practices seem to distort findings by conventional meta-analysis. Subset analysis based on journal affiliation with homeopathy revealed distorting characteristics of studies published in papers affiliated with homeopathy, leading to significant effect size overestimation when conducting conventional meta-analysis. From these results, two conclusions can be drawn. On the one hand, *p*-value based methods yielded consistent results with each other and are relevant alternatives to conventional meta-analyses. On the other hand, there is no evidence that individualized homeopathic treatments have an effect beyond the placebo effect. Observed symptom reduction in or increased well-being of patients are most likely due to a placebo effect.

Keywords: homeopathy, meta-analysis, *p*-curve, *p*-uniform, questionable research practices, p-hacking, publication bias

Index

1. Introduction

When a research domain offers several studies, a summary of some sort is recommended, especially if the results are inconsistent. One method to do that is the narrative review, where the existing scientific literature body is pooled into a verbal report. However, literature reviews are widely criticized for being incomplete, and due to their unsystematic nature prone to subjective opinions of reviewers (Schulze, 2004). Those problems can be addressed by a meta-analysis, a systematic approach characterized by higher objectivity and accuracy than single studies, or narrative reviews. It is used to condense studies to a summary effect and is associated with the ability of detecting a summary effect reflecting a population effect, as well as constructing theories or hypotheses (Schulze, 2004).

Meta-analyses can only assess data that are published or accessible through other means, such as personal correspondence with the authors. However, retrieving an unpublished study is often not successful. Because of that, meta-analytic results are only valid if the sample of studies is reflective of the examined effect. Therefore, the sample has to meet one of the following conditions: either all relevant studies, or a randomised sample of all existing studies is included in the analysis (Torgerson, 2003). Hence, the validity of meta-analytic results varies depending on how well these conditions are met. Violations of the aforementioned conditions affect a meta-analysis' validity. They mainly occur in two kinds: biases in published literature, the so-called publication bias, and specific strategies to maximize publication success, referred to as "questionable research practices" (QRPs).

There are few to no incentives for scientist and journals alike to publish null-hypothesis-confirming outcomes, which leads to such studies often ending up unpublished. To maximize the chances of publishing a study, researchers regularly apply strategies, such as only reporting statistically significant results. Such strategies are problematic, but widespread (Bakker, van Dijk, & Wicherts, 2012). The studies left unpublished are generally called grey literature or file drawers and show systematically different characteristics than published studies. These distinguishing characteristics mainly are either statistically non-significant results or effect sizes lower than the triviality threshold (Torgerson, 2003; Banks, Kepes & Banks, 2012). The triviality threshold is a numerical value, currently set at 0.20 for standardized mean difference, that an effect size has to exceed to be considered non-trivial by the scientific community. They are usually either not covered by literature databases or inaccessible. While studies reporting results supporting the null-hypothesis are less published,

those reporting statistically significant results consequently have a higher chance for publication, and statistically significant results are more likely to get fully reported (Kicinski, 2013). This underrepresentation of published studies on a certain topic is known as publication bias (Rothstein, 2007).

Publication bias can become a problem when interpretations and conclusions of unpublished studies disagree with those drawn by available studies. Consequently, scientific fields may pursue futile research or dangerous treatments may falsely be considered safe (Rothstein, 2007). In meta-analyses publication bias typically results in the overestimation of the examined population effect (van Assen, van Aert & Wicherts, 2015), leading to false conclusions about its existence or relevance.

Publication bias was a known problem even twenty years before meta-analyses were being used (Sterling, 1959; Glass, 1976). Since then, various methods have been developed as an attempt to deal with the issue, e.g., the fail-safe N (Rosenthal, 1979), selection method approaches (see Hedges & Vevea, 2005), the funnel plot (Light & Pillemer, 1984), the trim and fill method (Duval & Tweedie, 2000), the rank correlation test by Begg and Mazumdar (1994) or Egger's regression (Egger et al., 1997). Taken together, the number of methods trying to deal with publication bias have increased over the past years (Parekh-Bhurke et al., 2011).

Despite publication bias being frequently discussed as validity threatening, its prevalence is still high. In 1993, Dickersin and Min reported that statistically significant results are 2.9 times more likely to get published than statistically non-significant results. A few years later Weber et al. (1998) reported the even higher odds of 4.6. More recent estimates yield consistent prevelance rates; e.g., in 90% of all examined meta-analyses outcomes in favour of the alternative hypothesis had a higher chance of being published (Kicinski, 2013). Therefore, it is essential to apply methods for assessing publication bias appropriately.

Depending on the applied method, different information can be gathered. For example, the fail-safe N returns the number of studies that would be required to achieve a null effect (Rosenthal, 1978). However, this approach is only of historic interest, because large study samples are vulnerable to the alpha error. When dealing with large numbers of studies in meta-analyses, statistical significance is not a good criterion for identifying publication bias with this method. Therefore, it lacks practical relevance.

Traditional selection models have a high statistical power, but require a very large number of studies (Hedges & Vevea, 2005). Vevea and Woods (2005) have introduced an

approach for dealing with small numbers of studies. However, their method cannot adjust the effect size estimate so that it reflects the true effect size better, like selection models do when dealing with large study sets.

Other, widely used methods are based on the funnel plot. A funnel plot is a scatter plot of studies depending on their effect size estimate and some kind of measure of study precision, such as the standard error. According to Light and Pillemer (1984) such a scatter plot should look like a inverted funnel if all depicted studies measure the same effect. When such a scatter plot is asymmetric, publication bias is assumed. There are various statistical methods to detect funnel plot asymmetry, e.g., the rank correlation test and Egger's regression test (Begg & Mazumdar, 1994; Egger et al., 1997). However, on the one hand they struggle with low statistical power to detect publication bias (van Assen et. al, 2015). On the other hand, publication bias is not necessarily the only cause for an asymmetry (Sterne et al., 2011). Study-size-dependent effect sizes or sampling variations could account for an observed asymmetry, too, as well as accompanying biases like selective outcome reporting, selective analysis reporting, significance chasing (e.g., data peeking), or even mere fraud (Ioannidis & Trikalinos, 2007a; 2007b). The latter problems, along with numerous other non-intentional as well as intentional strategic behaviours, are subsumed under the aforementioned QRPs. Alternatively used terms are researcher degrees of freedom (Simmons, Nelson & Simonsohn, 2011) or *p*-hacking. All of them essentially describe the same construct.

Banks et al. (2016) described QRPs as "design, analytic, or reporting practices that have been questioned because of the potential for the practice to be employed with the purpose of presenting biased evidence in favour of an assertion." They come in various forms. Some QRPs are perceived as rather harmless and are widely practiced and accepted by the scientific community, such as "data peeking" (doing analyses before data collection is completed to see if the results show the desired direction) or optional stopping (stopping the data collection when statistical significance is achieved) (Simmons et al., 2011). However, other QRPs are more serious biasing behaviour; e.g. excluding data after examining its impact on the results. The most extreme type is data forging (John, Loewenstein & Prelec, 2012). For an extensive list of QRPs see Banks et al. (2016) or John et al. (2012).

QRPs may be induced by a number of different circumstances. Simmons et al. (2011) argued that it is common practice among researchers to explore the data set until they find statistically significant results, because making all analytic decisions a priori sometimes proves impractical. Furthermore, researchers are susceptible to decision heuristics which can

lead to optional stopping (Yu, Sprenger, Thomas & Dougherty, 2013) and ambiguity in analytic decisions depending on the desired outcome (Babcock & Loewenstein, 1997).

In 2009, Fanelli conducted a meta-analysis of studies surveying researchers whether they have engaged in a form of dubious research behaviour, excluding plagiarism. The admission rate of the harshest scientific misconduct – fabricating, falsifying or modifying data – was nearly 2%, and almost 34% of scientists declared other types of scientific misconduct. A newer study by John et al. (2012) exclusively surveying psychologists found self-admission rates for single QRPs from 1.7% for falsifying data to up to 66.5% for failing to report all dependent measures. 94% of psychologist admitted having been involved in at least one QRP. The study was replicated in Germany by Fiedler and Schwarz (2016), yielding lower yet non-trivial self-admission rates for single QRPs. In line with this finding, the majority of studies (91%) examining the subject of QRP find evidence for researchers being involved in at least some form of QRPs (Banks, 2016b).

Whatever the exact prevalence of QRPs in the scientific world may be, studies consistently show their existence to an extent which should not be ignored. While publication bias may threaten the validity of conclusions on an effect, QRPs threaten the validity of single results and studies by enhancing the probability for statistical significance. This is problematic because, as discussed above, statistical significant studies have a higher probability for getting published. Thus, QRPs do not only affect the studies on which they were applied, but contribute to publication bias by further introducing exaggerated or artificial effects into the literature body.

Both publication bias and QRPs can affect results of and inferences drawn from meta-analyses. Simonsohn, Nelson and Simmons (2013) as well as van Assen et al. (2015) have developed new methods for dealing with publication bias by only considering statistically significant studies, which are unlikely to remain unpublished. This approach circumvents the problem with selective non-reporting, that in turn leads to a literature selection, which is more likely unbiased.

1.1. p-Curve and p-Uniform

p-curve (Simonsohn et al., 2013) and *p*-uniform (van Assen et al., 2015) are meta-analytic methods addressing the aforementioned problems of publication bias and *p*-hacking by only considering statistically significant *p*-values for effect size estimations. It is assumed that all statistically significant results have the same probability of getting published, hence, the study

set is unaffected from publication bias. Though their development was unrelated to each other, their methods incidentally work in a similar fashion and are loosely based on selection model approaches (McShane et al., 2016).

1.1.1. *p*-Curve

p-curve focuses on how reported statistically significant *p*-values are distributed. Simonsohn et al. (2013) suggested that from this distribution, a researcher can infer whether a set of significant findings was likely or unlikely produced by *p*-hacking or selective reporting. When there is no effect, any statistically possible *p*-value has the same probability of occurring, forming a uniform distribution. When there is an effect, statistically significant *p*-values have a higher chance of occurring, forming a right-skewed distribution. Accordingly, a right-skewed *p*-curve indicates an underlying effect while a uniform distribution indicates no effect. A left-skewed curve suggests intense *p*-hacking or selective reporting, because of the presumption that QRPs are conducted to the point where statistical significance is achieved, which is conventionally at $p = .05$. Therefore, when many studies within the study set are intensely *p*-hacked, *p*-values just under $p = .05$ accumulate (Simonsohn et al., 2013), giving the curve a left-skewed shape. Effect size and average power of subjected studies can be estimated as well. The implemented statistical test for right-skewness allows inferences about whether the study set contains evidential value. It is worth mentioning that a detected lack of evidential value does not necessarily mean that the underlying theory is wrong or the assessed effect nonexistent, but that the analyzed set of studies provides no evidence for the theory or assessed effect.

p-curve's statistical tests are rather straightforward. In a first step, the probability for observing the original or a more extreme *p*-value is calculated for each subjected *p*-value. In a second step all of these *p*-values of *p*-values, so-called *pp*-values, are subjected to Fisher's method, which returns a χ^2 test for skewness with twice as many degrees of freedom as there are *p*-values (Simonsohn et al., 2013). The null of no effect is discarded if $p < .05$. The tests for right skewness is implemented in the *p*-curve web application (http://www.p-curve.com/app4/) by default.

The *p*-curve web application provides additional information regarding the "stability" of its results. This comes from a cumulative meta-analysis, in which highest or lowest observed *p*-values are dropped cumulatively to monitor how and to what extent the *p*-values of the statistical tests conducted by *p*-curve change when dropping those observed *p*-values cumulatively. When the interpretation drawn from the results changes with just a few dropped

p-values, the estimate is considered not stable and it should not be interpreted. A diagnostic plot for power estimation is also provided. Similar to the effect size estimation, *pp*-values are calculated for the null hypothesis that all included studies are powered at any level of statistical power and subjected to Stouffer's method which returns a Z-score (Simonsohn, Nelson, & Simmons, 2014). The best fit is displayed by the lowest absolute Z-score and indicates the best guess for the study set's overall power. If the plot is V shaped, the power estimate can be interpreted. If not, it should be regarded with caution.

Effect size estimation can be conducted in specialised computer programs, e.g., R, using a loss function, where the loss is a function of how well an expected *p*-curve matches an observed *p*-curve (Simonsohn et al., 2014). The minimum loss indicates the best fitting effect size.

1.1.2. *p*-Uniform

p-uniform, too, is based on the distribution of *p*-values and premises on its uniform shape in presence of a true null effect. It can estimate population effect sizes more accurately than conventional meta-analytic methods, which, as discussed before, may be cofounded by publication bias. *p*-uniform's statistical tests are straightforward in the same ways as *p*-curve's, but instead of fitting a loss function to match the distribution of *p*-values, *p*-uniform shifts the effect size until the distribution of observed conditional *p*-values fits the distribution of expected conditional *p*-values under uniformity (van Aert, Wicherts, & van Assen, 2016). The best fit is obtained when the sum of observed conditional *p*-values equals the sum of expected conditional *p*-values under uniformity. The larger the true effect size, the higher the sum of conditional *p*-values (van Aert et al., 2016). When no shifting is necessary to match the distributions, a null effect is assumed.

The web application features an effect size estimation based on statistically significant *p*-values, a conventional fixed effect size estimation, a publication bias estimation, and a probability-probability plot of observed conditional *p*-values and expected conditional *p*-values. This plot allows inferences on the frequency of observed *p*-values and can be interpreted in a similar way as a *p*-curve. The test for publication bias is considered more powerful than the conventional trim and fill method (van Aert et al., 2016).

1.1.3. Disadvantages

Beside Fisher's method and dependence on *p*-value distributions, *p*-curve and *p*-uniform have several other aspects in common. Different algebraic signs cannot be included in the same

analysis. This means that those methods cannot be applied to a set of statistically significant studies that contain both positive and negative effect sizes.

Both methods perform well with low numbers of *p*-values if effect sizes are homogeneous or heterogeneity is not too large. However, simulation studies have shown that in case of large heterogeneity *p*-curve performs less accurate, and *p*-uniform yields implausible estimates (van Aert et al., 2016). In this case, conventional meta-analyses outperform *p*-value-based methods by applying a random effects model. This makes *p*-curve and *p*-uniform not universally applicable. Van Aert et al. (2016) also conceded that results from both *p*-curve and *p*-uniform may be biased or less precise depending on the type of QRP. Some kinds of QRPs, such as ghost *p*-hacking (obtaining several dependent measurements but only reporting the subsets with statistically significant results), cannot be detected (Bishop & Thompson, 2016). However, no other method is able to do that either.

Another disadvantage is that *p*-uniform as a whole and *p*-curve's effect size estimation are not yet easily accessible. Still those two methods are sensible approaches to the problems surrounding QRPs, while simultaneously providing an approach to the problem of having no access to unpublished, statistically non-significant studies which come along with publication bias.

Simulation studies have already put *p*-curve and *p*-uniform to the test (Simonsohn et al., 2014). This thesis aims to examine their performance on a real study set. The data come from the homeopathy literature.

1.2. Homeopathy

In 1796, German physician Samuel Hahnemann published a paper about a "new principle for discovering healing abilities of remedies" (Hahnemann, 1796) in a medical journal. His idea builds upon two postulated mechanisms of actions: The principle of equivalence and the principle of potentisation.

The first one is based on a self-experiment in which Hahnemann observed fever-inducing properties of china-bark on himself. Because china-bark was used as an antipyretic natural medicine on malarial patients, he concluded that a substance causing specific symptoms in a healthy person can cure the very same symptoms in a person suffering from them. There is no rigorous clinical experiment replicating his observation, to date. The second principle concerns the concentration of the homeopathic remedy. One part of the allegedly

potent substance is diluted in ninety-nine parts liquid (water or alcohol) and churned in a very specific fashion: The vessel containing the solution is knocked against a leather pillow ten times, so that, according to Hahnemann, the allegedly potent substance can transmit its information to the liquid. The solution is called C1 and one part of it is diluted in another ninety-nine parts of liquid and churned again to obtain C2, and so on. While the actual substance becomes exponentially less with every potentisation, homeopaths argue that it becomes more effective. According to homeopaths, not biological or pharmacological mechanisms are responsible for a cure, but "energetic principles" (Frank, 2015; for a discussion, see Ernst, 2010).

Both principles lack scientific evidence and are regarded as implausible among critics ever since they were established (Ernst, 1998). The idea of homeopathy focuses on the concept of treating symptoms individually, disagreeing with modern medical understanding of diseases.

Nevertheless, Relton et al. (2017) found worldwide prevalence of use of homeopathy. For example, homeopathic remedies prescribed by homeopaths or general practitioners ranged from 0.2 – 0.6% in the USA up to 6.4 – 8.2% in Switzerland, where homeopathy is covered by health insurance. The prevalence in Germany is about 1%. In the context of government policies, homeopathy is a controversial topic, especially in Germany and Austria. Discussions about a possible integration in health insurance in Austria and expansion of homeopathy treatment coverage by health insurances in Germany are frequently being held (see Gartlehner, 2016; Springer, 2017; Weber, 2017; Albrecht & Maier, 2017).

There are little doubts that observable improvements in symptoms are taking place when administering homeopathic remedies (e.g., National Health and Medical Research Council, 2015). However, while homeopaths attribute those improvements to the remedies itself, critics hold placebo effects responsible. Placebo effects can occur within the framework of consulting a homeopath because homeopaths take substantially more time for their patients, listen to them carefully and prescribe seemingly tailored remedies for the individual, raising the feeling of being understood and appreciated (e.g., Gray, 2016). Although scientific studies of homeopathic effects exist, the data seem ambiguous. A majority of published studies is purely observational and does not compare effects between treatment and control groups, therefore conceivably capturing mainly placebo effects (e.g., Goossens et al., 2009; Wadhwani, 2013; Karp et al., 2016). Randomised controlled clinical trials, on the other hand, constantly yield inconsistent results (e.g., Adler et al., 2013; Mousavi et a., 2009).

In attempts of clarification, several meta-analyses have been conducted (e.g. Linde & Melchart, 1998; Ernst, 2011; Hahn, 2013; Mathie et al., 2014; Boehm et al., 2014;). The results, however, contradict one another, too. Meta-analyses finding that homeopathic effects are not superior to placebo effects are often criticized for applying arbitrary exclusion criteria and mixing distinct effects by pooling studies of different homeopathic remedies and medical conditions (Hahn, 2013). Furthermore, homeopathy advocates term meta-analyses inappropriate to describe the type of effect that homeopathy produces (Hahn, 2013). At the same time, meta-analyses finding effects in favour of homeopathy are accused of merely describing a placebo effect, including studies with deficient precision and poor overall quality, and examining an overall inadequate concept without scientific founding (Ernst, 2002).

Notably, certain kinds of journals tend to publish certain kinds of studies. Journals containing terms related to alternative or complementary medicine in their journal titles (e.g., The Journal of Complementary and Alternative Medicine, Homeopathy) seemingly publish articles, studies and meta-analysis favouring homeopathy more often. Journals without such terms in their title feature less publications concerning homeopathy. When the latter kind of journals do, the majority of these publications find no evidence for the effectiveness of homeopathy.

Meta-analyses published in the latter journals often exclude studies featured in homeopathy-affiliated journals because of their insufficient quality. Homeopathy advocates criticise this, claiming that authors of such analyses deliberately exclude results that seem to confirm homeopathic effects (Ernst, 2002).

In this analysis, any study was included, as long as it matches the inclusion criteria fully described in the method section. Inclusion criteria address the points of criticism regarding the mixing of effects by verbalizing the specifics of included studies, e.g., individualized homeopathic treatment to prevent possible mixing of effects. The approach of only including individualized homeopathic treatment was chosen because individualized homeopathic treatment prescribed by trained homeopaths corresponds to the doctrine of homeopathy, which claims that homeopathic remedies cure individual symptoms instead of specific illnesses, as conventional medicine does. Furthermore, only randomised controlled trials are included in the study set. To date, they are the best known way to distinguish an actual effect from a placebo effect by comparing a treatment group with a control group.

2. Method

2.1. Hypotheses

While simulation studies back up the efficency of *p*-curve and *p*-uniform, there are only a few applications on real data, to date. The primary goal of this master's thesis was a comparison of the two methods to the conventional meta-analytic approaches regarding usability, interpretation of effect sizes and accuracy, on the basis of homeopathy studies. Thus, the first thematic block of hypotheses had a methodological focus.

As mentioned before, *p*-curve and *p*-uniform can produce nonsensical effect size estimates if between-samples heterogeneity is large. Hence, it is imperative to look at the amount of heterogeneity of a given data set before taking the results as valid. The I^2- and Q-statistics reported by conventional meta-analysis and *p*-uniform provide information about this variable of interest. When heterogeneity is small to moderate, the shape of *p*-curve's loss function plot and *p*-uniform's effect size direction can be examined for further information about the impact of possible heterogeneity on the precision of the respective effect size estimation. If *p*-curve's loss function plot is erratic and *p*-uniform's effect size direction is implausible, the estimate should not be interpreted.

Effect sizes were estimated with *p*-curve, *p*-uniform, and conventional meta-analysis. Separately, plausible effect size estimates were then examined regarding their consistency, meaning that the effect size estimates should both be statistically significant with the same effect size direction or not statistically significant. It was assumed that *p*-curve and *p*-uniform yield plausible values when applied on real data. In case of publication bias, effect size estimations obtained from *p*-curve and *p*-uniform, and traditional meta-analytic estimates should vary. In the presence of publication bias, conclusions about the efficacy of homeopathic treatment deduced from *p*-curve and *p*-uniform would deviate to a significant extent from conclusions about the efficacy of homeopathic treatment deduced from conventional meta-analyses.

It was also of interest whether the estimated effect size of individualized homeopathic treatments yielded by *p*-curve and *p*-uniform was below the triviality threshold and if and how much they differed between study-subsamples split by journal affiliation. Because homeopathy-affiliated journals seemingly tended to find effects while non homeopathy-

affiliated journals seemingly tended to find no effects, it is possible that QRPs were responsible for the statistical significance of those effects. The presence of QRPs were examined with p-curve's statistical tests regarding the shape of the *p*-value distribution.

2.2. Sampling

2.2.1. Include criteria

To determine which studies should be used for the analyses, inclusion criteria were a priori established. For inclusion a study had to be 1.) a randomised placebo-controlled trial and 2.) using individualized homeopathy as treatment. The subjects had to be 3.) adult patients with quantifiable symtpoms. Additionally, the study had to 4.) report sufficient statistical paramteres, 5.) had to be accessible, and 6.) had to be written in German or English. 7.) The data had to be independent, too.

2.2.2. Literature search

Four scientific data bases (PubMed, ISI Web of Science, Scopus, PsycINFO) were searched, using the following key words: (homeopath* OR homoeopath*) AND (rct OR randomized controlled trial OR controlled clinical trial OR placebo controlled) AND (individual*) NOT (animal). Additional filters such as "not animal" and "clinical trial" were applied on search results, yielding a total amount of 284 studies (see Figure 1). Another 4 studies were added because they were cited in meta-analyses and studies examining individualized homeopathic treatments. 130 were duplicate hits, leaving 158 for detailed examination. Of those, 16 did not cover homeopathy, 67 were no randomised controlled clinical trials or no trials at all, 28 had an inept sample (healthy adults, animals or children), and 16 did not use individualized homeopathic treatment or did not compare the treatment to placebo. Another 7 studies failed to report appropriate statistical values for effect size calculation, and 3 were not retrievable, leaving a final sample of 21 studies.

Relevant information about the studies were put into a coding scheme. The coding scheme was developed in view of the hypotheses and was designed to gather three different types of information, namely study identification, statistical values, and wheter the journal in which the study was published has an affiliation with homeopathy. The study identification was done via unique study identification numbers. Typical statistical values coded were, e.g., sample sizes, means of symptom scores or pain scores, corresponding standard deviations or

standard errors when dealing with continuous data, or cell frequencies in case of dichotomuous data. Also, each study contributed one *p*-value to the analysis. The affiliation of the journal was appraised by wheter the journal title contained terms typically associated with homeopathy or alternative medicine.

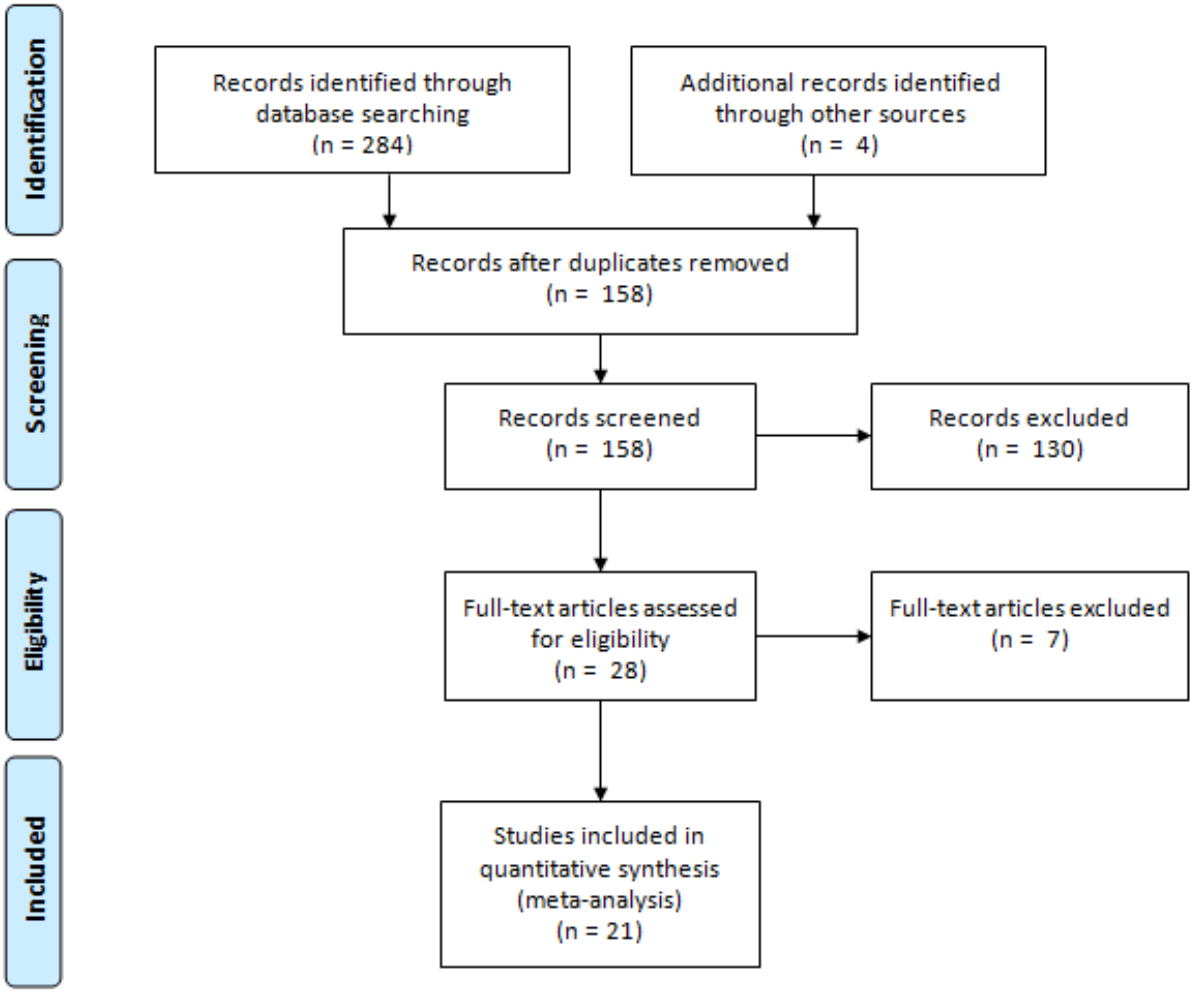

Figure 1: Flow chart diagramm of included and excluded studies.

When more than one *p*-value per study was identified, the most appropriate *p*-value was picked. The choice was based on the deliberation of which *p*-value was most reflective of the effect desired to examine, e.g., post-treatment measures comparing treatment and placebo groups, and proposed main outcome measure. No study reported effect sizes. Therefore, effect sizes were calculated. *F*-values, *t*-values or χ^2-values needed for calculations with *p*-curve and *p*-uniform were either extracted from the studies or calculated.

2.2.3. Final sample

The final sample consisted of 21 studies. The average sample size was 48.19 (*SD* = 27.83)

with the smallest sample being *n*=14 and the largest being *n*=120. The average treatment duration was 26.91 (*SD* = 54.15) weeks, with 6 days being the minimum and 5 years the maximum. The oldest study was from 1991, the most recent from 2015. Most of the studies were conducted in Europe (8), followed by Asia (6), South America (3), North America (3), and Africa (1). Eight studies reported a *p*-value under .05 on the main effect while 13 found no statistical significance. For further details see Table 1.

2.2.4. Effect size calculations

For the conventional meta-analysis, effect sizes were calculated with the escalc-function within the R-package metafor (Viechtbauer, 2010). Two different types of outcome measures were identified. Effect sizes of continuous outcome variables were directly computed as standardized mean difference, effect sizes of discrete outcome variables were computed as Odds Ratio and transformed subsequently into standardized mean differences. Because most studies were designed to examine a potential reduction of symptoms, their effect size was negative if the outcome appeared to be in favour of homeopathy. Effect size directions of studies examining improvements of some sort as outcome measures, as well as transformed effect sizes, were made consistent and integrated in the effect size matrix.

Two studies from the same first author (Bell) and the same year (2004) were coded as Bell.1 and Bell.2. While both studies provided sufficient test statistics for analysis with *p*-curve and *p*-uniform, Bell.1 could not be subjected to conventional meta-analysis due to a lack of reported parameters.

2.2.5. Study subsets

Two study subsets were created for secondary analysis, using the split criterion "observable journal affiliation". Whether there was an affiliation or not was primarily based on the journal title. Journal titles including terms such as "homeopathy", "complementary", or "alternative" were considered to be affiliated with homeopathy. Further criteria were the statements and positions of the journal, and whether they had homeopathy as primary objective. Studies published in journals close to homeopathy were assigned to the homeopathy-affiliated subset, studies published in a journal not related to homeopathy were assigned to the non homeopathy-affiliated subset.

The subset of studies published in homeopathy-affiliated journals consisted of 13 studies, the subset of studies published in non homeopathy-affiliated journals of 8 studies. When running a χ^2-test, affiliation seemed not to be associated with statistical significance (χ^2

= 0.27; p = .60), indicating that in this analysis homeopathy-affiliated journals did not contribute statistically significant more p-values below .05.

Study subsamples did not differ in regard to sample size (t = -0.36; p = .72) or treatment duration (t = -0.99; p = .34).

Table 1: Summary of included studies

ID	First Author	Year	Treatment sample	Placebo sample	main outcome measure	Nature of main outcome measure	Treatment duration	Statistical significance	Effect Size (*d*)	Journal affiliated with Homeopathy
MA01	Adler	2013	16	7	HAM-D	Continuous	6 weeks	No	0.49	No
MA02	Andrade	1991	17	16	SRH-MBTI Functional Assessment	Discrete	6 months	No	-0.37	No
MA03	Bell	2004	23	25	EEG alpha magnitude	Continuous	16 weeks	Yes	-[1]	No
MA04	Bell	2004	26	27	25% improvement in tender point pain	Discrete	16 weeks	Yes	-1.06	No
MA05	Bonne	2003	20	19	HAMA-A	Continuous	10 weeks	No	0.07	No
MA06	Brien	2011	16	16	HAMA-A	Discrete	24 weeks	No	0.00	No
MA07	Cavalcanti	2003	11	9	Percentage of pruritus reduction	Continuous	60 days	No	-0.14	Yes
MA08	Chand	2014	60	60	Symptom score	Continuous	5 years	No	0.05	Yes
MA09	Chapman	1999	27	23	SRHI-MBTI Functional Assessment	Continuous	4 months	Yes	-0.49	No
MA10	Fisher	2006	15	12	100mm visual analogue scale of overall symptom severity	Continuous	12 weeks	No	-0.16	Yes
MA11	Frass	2005	33	34	Survival rate	Discrete	180 days	Yes	-0.69	Yes
MA12	Koley	2015	30	30	Pain Scale	Continuous	2 weeks	No	0.00	Yes
MA13	Macías-Cortés	2015	44	43	HRSD	Continuous	6 weeks	Yes	-1.50	No
MA14	Mousavi	2009	50	50	Moderate improvement rate on ulcer size	Discrete	6 days	No	-0.59	Yes
MA15	Naude	2010	14	16	Sleeping Diary	Continuous	4 weeks	Yes	-[2]	Yes

[1] only *p*-value available

[2] effect size could not be calculated

Table 1: Summary of included studies (*Continued*)

ID	First Author	Year	Treatment sample	Placebo sample	main outcome measure	Nature of main outcome measure	Treatment duration	Statistical significance	Effect Size (d)	Journal affiliated with Homeopathy
MA16	Peckham	2014	16	15	IBSS score	Continuous	26 weeks	No	-0.72	Yes
MA17	Rastogi	1999	20	18	$CD4^{+ve}$ T-lymphocytes	Continuous	6 months	Yes	-0.35	Yes
MA18	Siebenwirth	2009	5	9	Multiparameter-Score	Continuous	32 weeks	No	0.39	Yes
MA19	Straums-heim	2000	35	33	Overall effect assessed by neurologist	Discrete	4 months	No	-0.37	Yes
MA20	Thompson	2005	28	25	MYMOP overall profile score	Continuous	16 weeks	Yes	-0.34	Yes
MA21	Yakir	2001	33	34	MDQ score	Continuous	3 months	Yes	0.80	Yes

3. Results

3.1. Conventional meta-analysis

Included studies varied within their methods, duration and main outcome measures. Hence, a random effects model was applied, yielding an estimated summary effect size of d = -0.46 ([-0.89;-0.03]; p = .03), indicating a moderate effect in favour of homepathy. Study-specific effect size estimations and 95% confidence intervals are shown in Figure 1.

The test for heterogeneity was statistically significant ($Q(18) = 106.39$; p <.001), indicating large heterogeneity. I^2 amounted to 87.71%. This means that 87.71% of the total variability in the estimated effect sizes was due to variations in sample size, method or other moderating variables (Viechtbauer, 2010), opposed to unsystematic variability of observed effects.

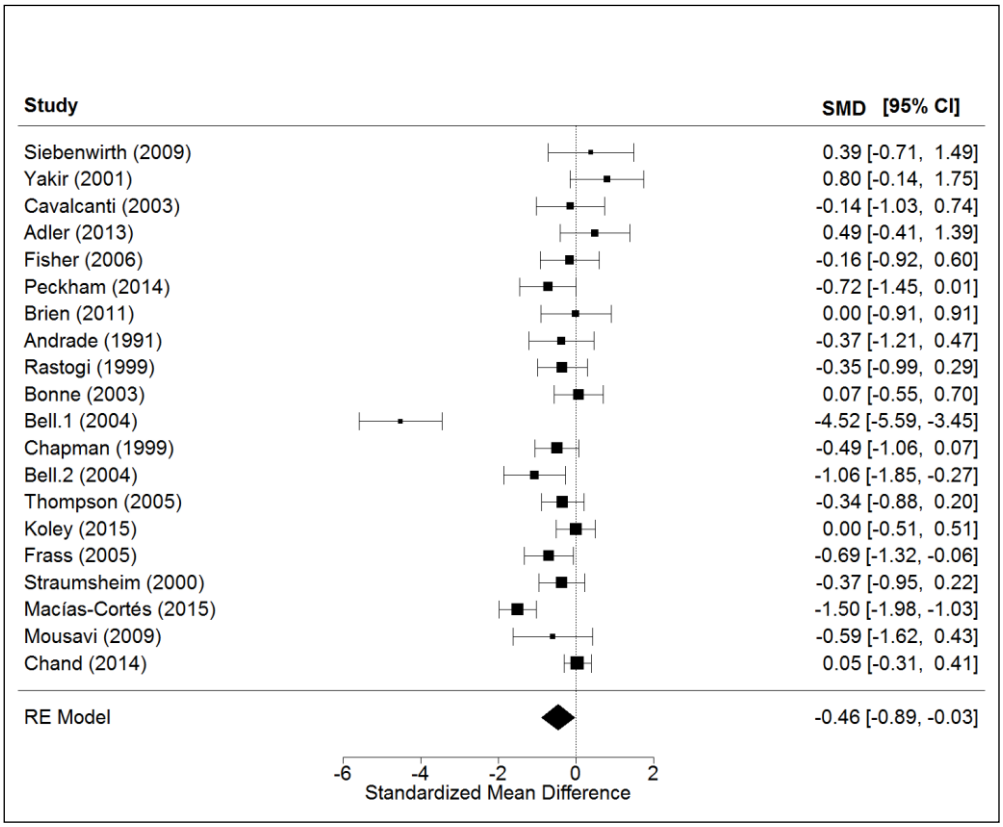

Figure 2: Forest plot of individual effect sizes and 95% confidence intervals of all studies included in the conventional meta-analysis, sorted by study n (ascending).

Sensitivity analysis was conducted via leave-one-out diagnostics. One study was identified as a potential distortion for meta-analytic results. Leaving Bell.1 (2004) out shrank the effect size estimate to d = -0.31 ([-0.55;-0.05]; p = 0.01) and I^2 to 60.33, but since the studies sample size was small, the face value interpretation did not change. Omitting any other study did not change the size of the estimate notably.

Four methods for detecting publication bias were applied. The trim and fill method suggested no missing studies on the right side of the funnel, but identified seven potentially missing studies on the left side of the funnel (Figure 3), implying statistically significant unpublished studies. Both rank correlation test (Kendall's $\tau < .001$; p = 1.00) and Egger's regression test (z = -0.62; p = .54) were not statistically significant, indicating no publication bias.

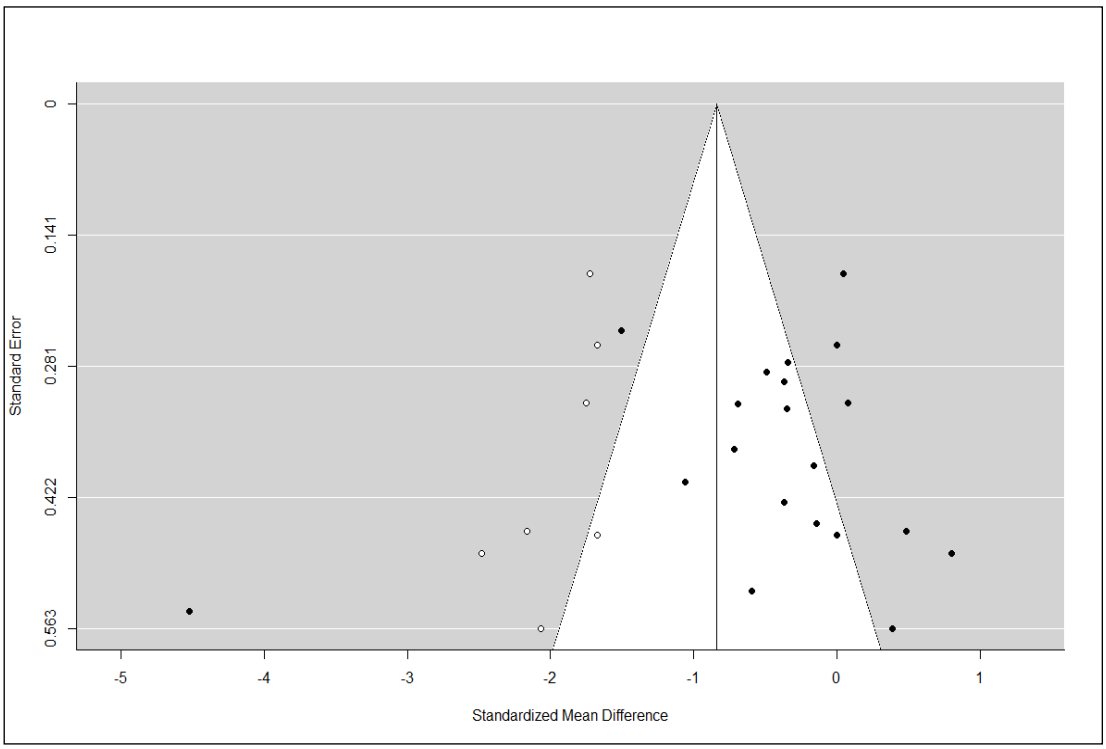

Figure 3: Funnel Plot.

When splitting the sample by journal affiliation and conducting the trim and fill method (Figure 4 and Figure 5), the funnel plots showed different distributions of effect sizes. The non homeopathy-affiliated subset showed missing studies, as well as an outlier with a large effect size but also a high standard error. This may be explained by a massive publication bias, because prestigious journals generally tend to publish studies which found large effect sizes more often than studies that found low effect sizes. The outlier may

contribute to the large heterogeneity observed in this subset (I^2 = 94.20%, Q = 75.92, p < .001). When examining the homeopathy-affiliated subset, no outliers were detected and the heterogeneity was negligible (I^2 = 13.27, Q = 13.47, p = 0.27). When conducting the conventional meta-analysis, the different distributions of the subsets were pooled together and may have caused the funnel plot asymmetry on the wrong side of the funnel.

For the homeopathy-affiliated subset the summary effect estimated by a conventional random effects model amounted to d = -0.19, and for the non homeopathy-affiliated subset to d = -0.90. Both summary effects were not statistically significant. This discrepancy in effect size estimates may be another result of the tendency of prestigious journals to publish studies with large effect sizes more likely than studies with other outcomes.

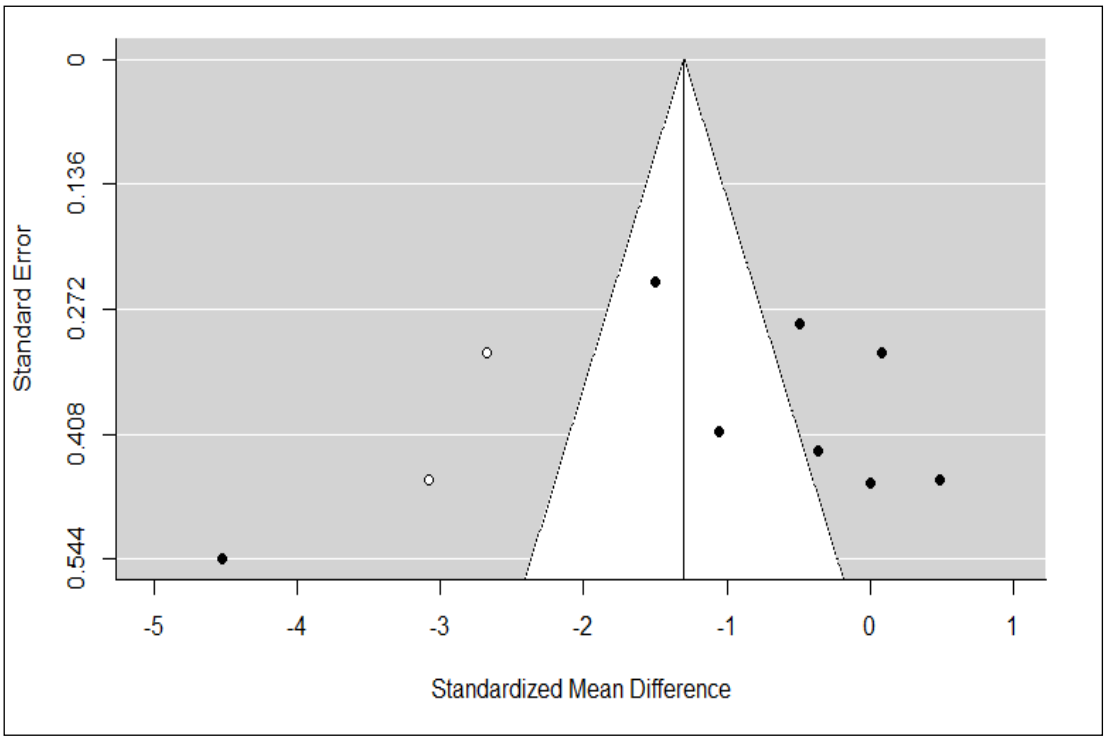

Figure 4: Funnel Plot of the non homeopathy-affiliated subset.

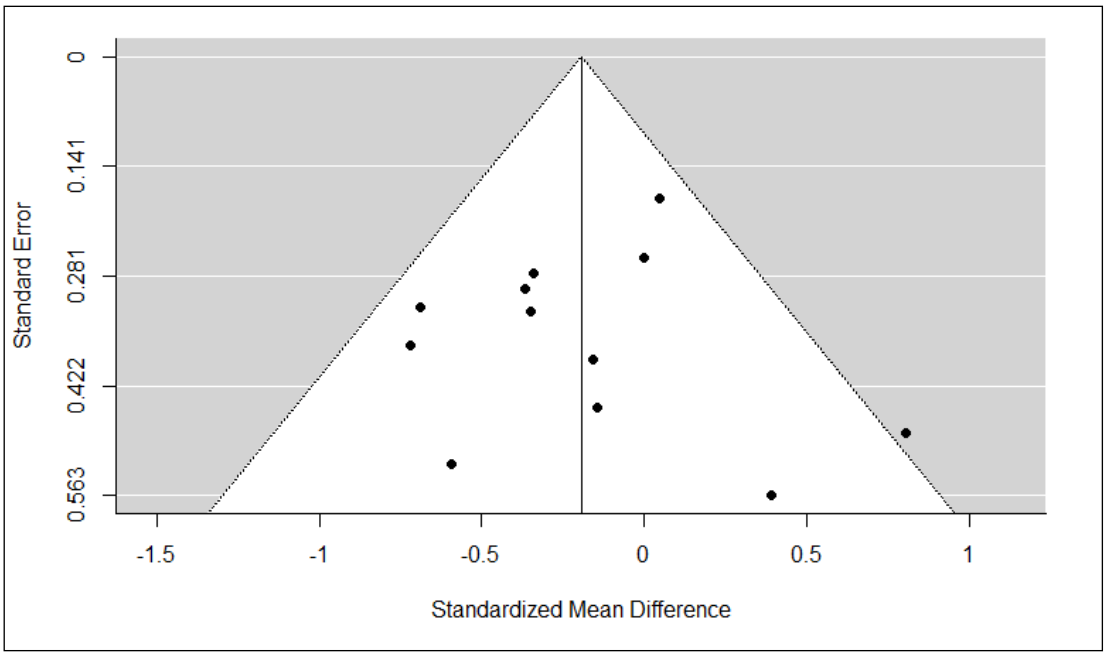

Figure 5: Funnel Plot of the homeopathy-affiliated subset.

3.2. p-Uniform

There was no evidence of heterogeneity within the subset of statistically significant studies when applying the Cochran's *Q* test ($Q = 9.67$; $p = .29$). The Irwin-Hall estimator was selected for estimation. *p*-uniform's effect size estimation based on statistical significant *p*-values yielded a summary effect of $d = -0.28$ [-0.74;0.70] with no statistical significance ($p = .21$). For comparison, a conventional fixed effect size model implemented in the *p*-uniform web application was applied and estimated a summary effect of $d = -0.86$ ([-1.05;-0.66]; $p < .001$), which was larger than both conventional random effect model and fixed effect model effect size estimation ($d = -0.41$ [-0.55;-0.26], $p < .001$).

In their supplementary material to *p*-curve, Simonsohn et al. (2013) pointed out that treating discretely distributed test statistics, like a *p*-value obtained from a χ^2-test, as student distributions may also account for slight deviations. This may add to the observed discrepancy. However, they argue that those deviations do not distort results to a serious extent. Because *p*-uniform is similar in its method, it is assumed that ignoring the actual type of distribution does not lead to significantly different estimates. In fact, leaving those studies out did not change interpretation of effect size estimate ($d = -0.28$ [-0.87;1.95], $p = .28$).

Even though *p*-uniform did not detect publication bias ($p = .28$), probabilities for conditional *p*-values showed two biases when displayed graphically (Figure 6). One was located at the lower end of probabilities, where low *p*-values occured less frequent than

statistically expected. The second one appeared on the upper end of probabilities, where *p*-values near the threshold for statistically significance were observed more frequently than statistically expected.

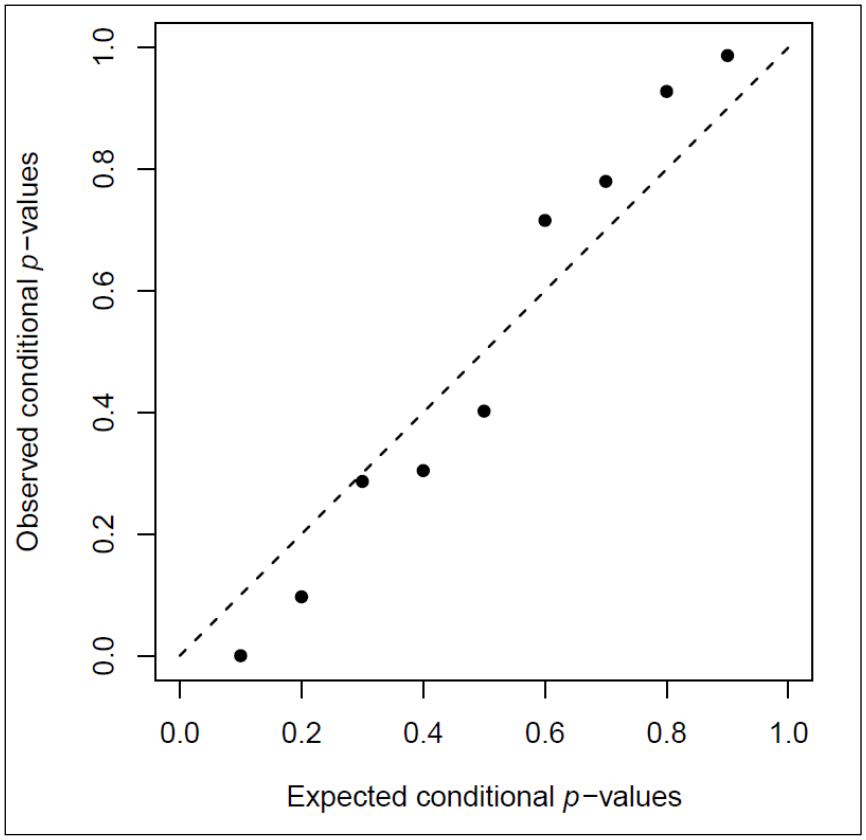

Figure 6: Probability-probability plot for all 9 included studies.

3.3. p-Curve

The nine studies forming the subset of statistically significant studies were subjected to the *p*-curve web application. Power analysis and *p*-value distributions were conducted based on original test statistics.

Effect size estimation yielded a summary effect of $d = 0.71$ (Figure 7) in favour of placebo. Calculations of confidence intervals and *p*-values are not possible. Here, too, *p*-values based on χ^2 tests were excluded for comparison, leaving three studies to analyse. In this case *p*-curve estimated a positive small effect in favour of placebo ($d = 0.22$).

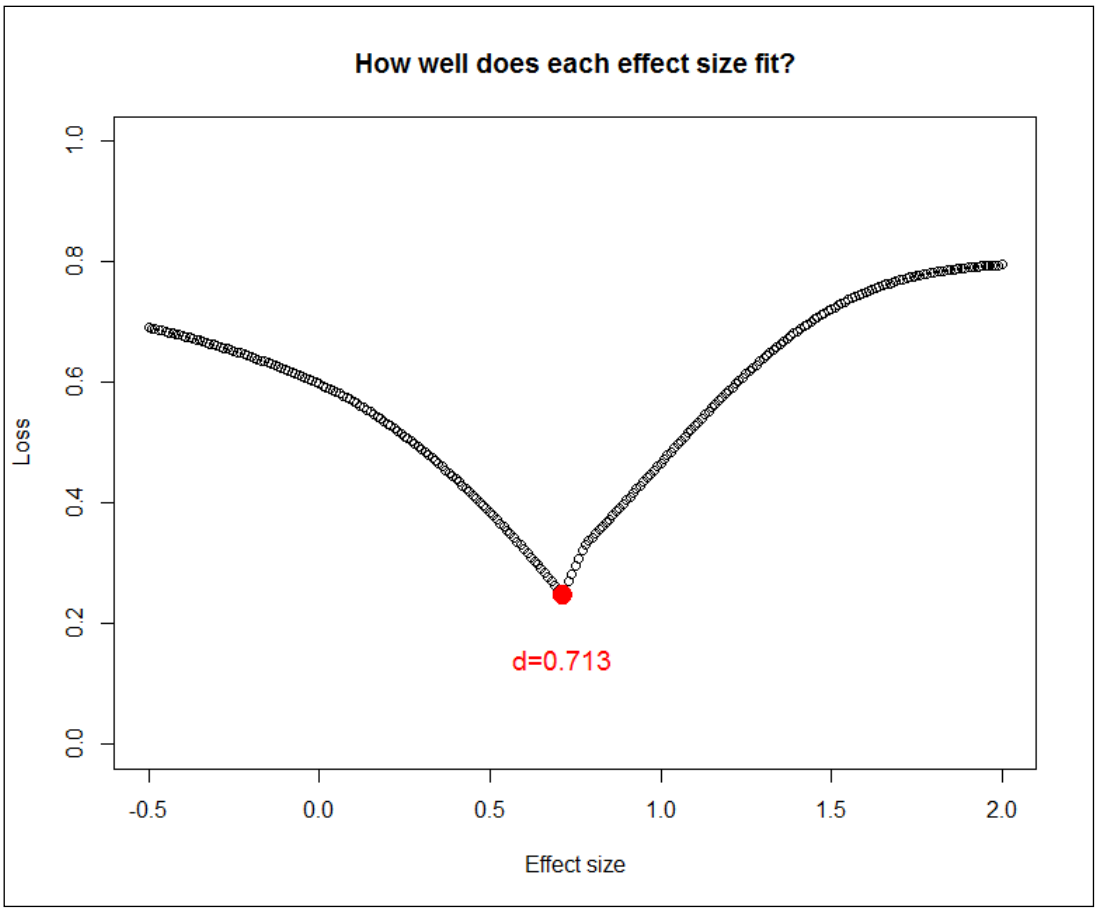

Figure 7: R-Plot of loss function.

Figure 8 depicts the percentual distribution of statistically significant *p*-values (blue), the shape of the distribution if studies were generally underpowered at 33% (dashed green), and the shape of the distribution in case of no effect (dotted red). The actual curve showes that low *p*-values and *p*-values near the threshold of statistical significance were observed more often than expected in case of a population effect. This interpretation was consistent with conclusions drawn from *p*-uniform.

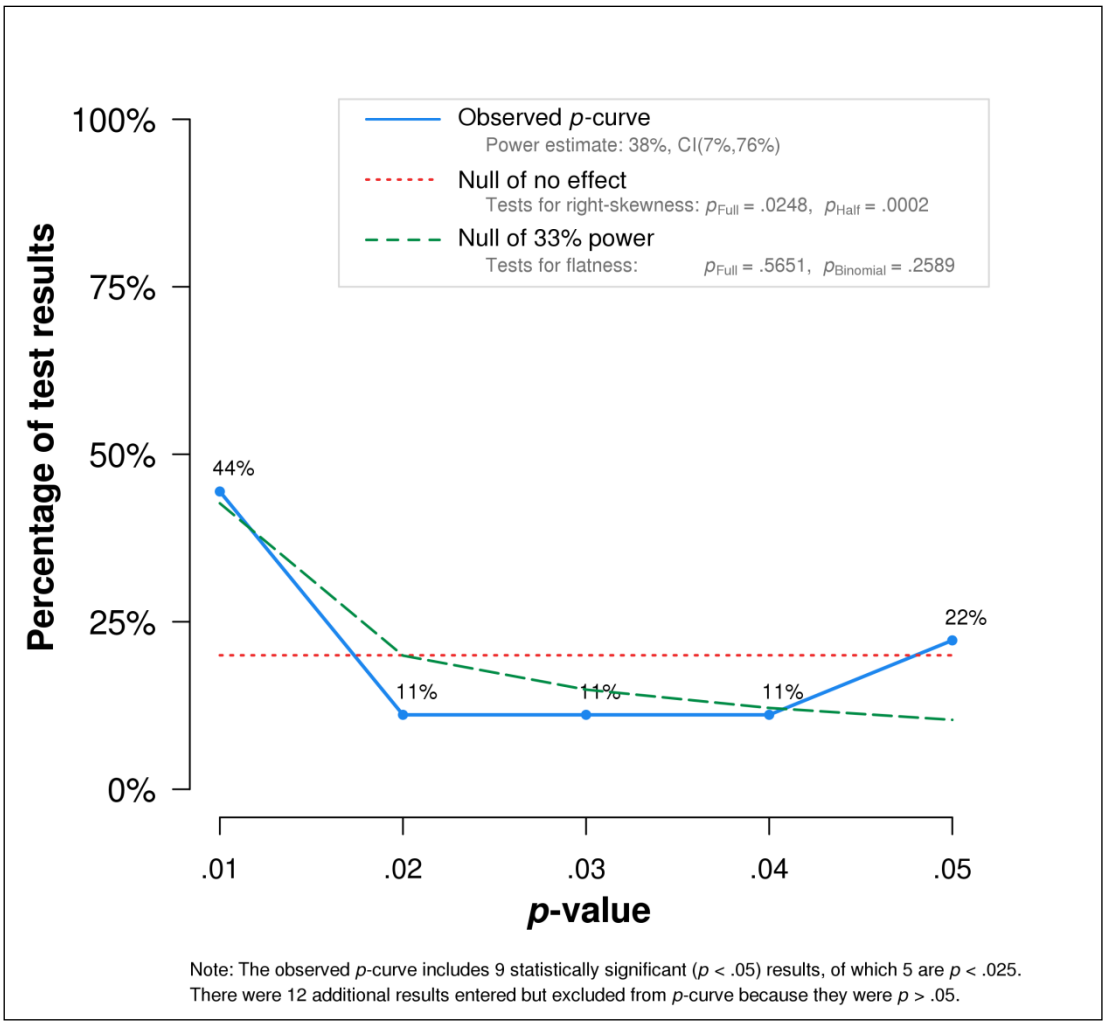

Figure 8: Percentual distribution of observed statistically significant *p*-values. The test statistic is provided on the upper right.

The test for right skewness was not statistically significant for full *p*-curve ($p = .25$), but statistically significant for half *p*-curve ($p = > .001$), indicating evidential value. However, the statistical power of all statistically significant studies was estimated at 38% [7%;76%], which means that the study set was underpowered.

Figure 9 shows what happens to statistical significance of *p*-curve's statistical tests if *k* of the lowest or highest *p*-values are omitted from the subjected data set. The highest *p*-value had to be excluded to achieve a statistically significant test for right skewness (upper graphs). When doing that by deleting Macías-Cortés (2015) from the study set, the percentage of observed *p*-values near the threshold for statistical significance increased while the percentage of low observed *p*-values decreased (Figure 10). More interestingly, power dropped to 5% [5%;39%], with not even the upper bound confidence interval exceeding the threshold for

chance at 50%. The middle graph considers only *p*-values under .25. Dropping the lowest of these *p*-values resulted in the test for right-skweness not being statistically significant anymore. The test for 33% power did not reach statistical significance if any of the lowest or highest original *p*-values were dropped.

Simonsohn et al. (2013) recommended to put more trust in results obtained from study sets that do not change when excluding a few original *p*-values. In this case, excluding the most extreme *p*-values did not change the interpretation regarding the study set's low statistical power. Single studies were not responsible for the misfit between a hypothetical curve of p-values obtained from an underpowered study set, but most studies in the study set contributed to that misfit. This means that the distribution of observed *p*-values could not be fully explained by low power.

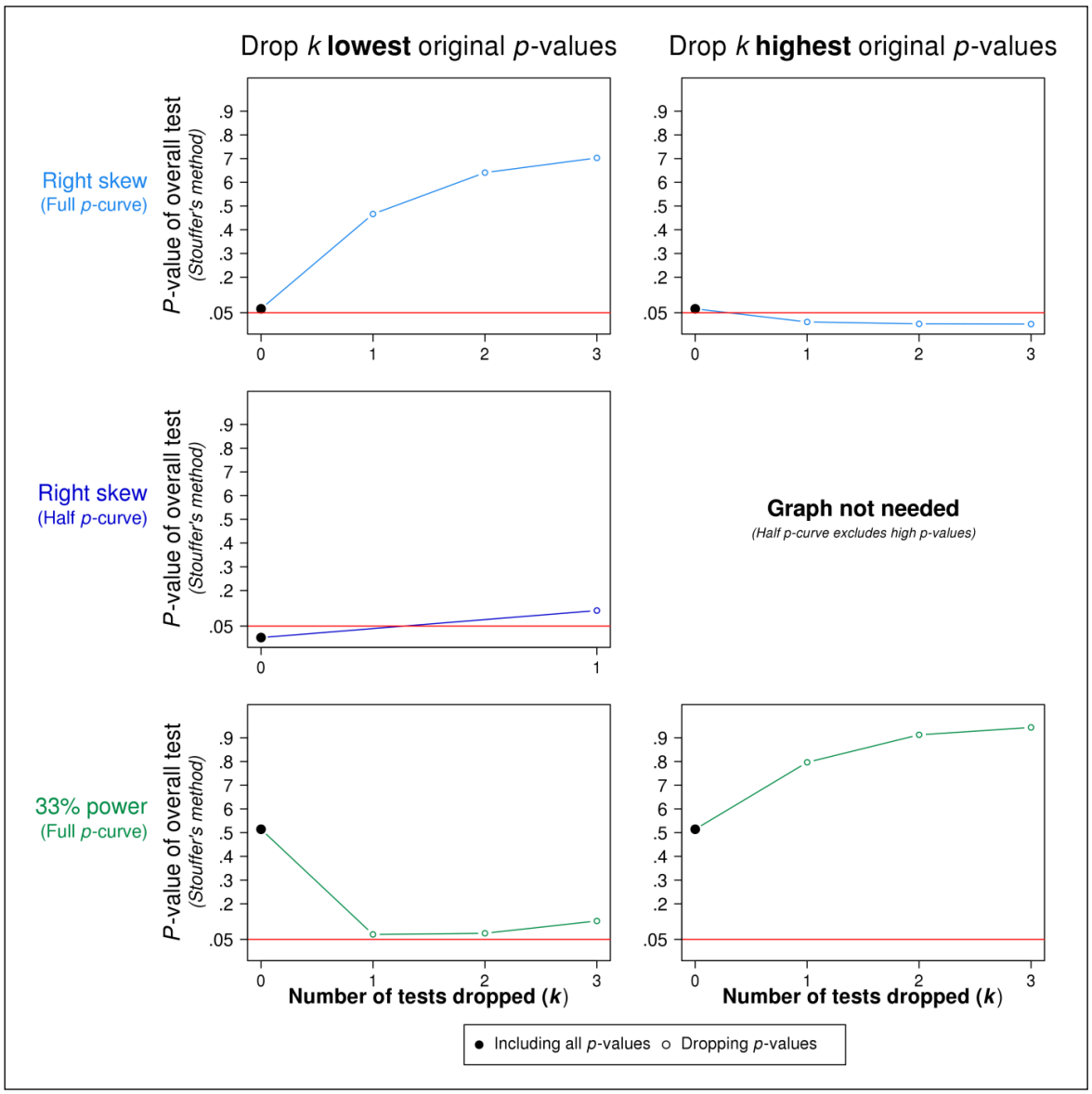

Figure 9: Changes in *p*-values of tests for right skewness and 33% power when dropping k lowest or highest original *p*-values.

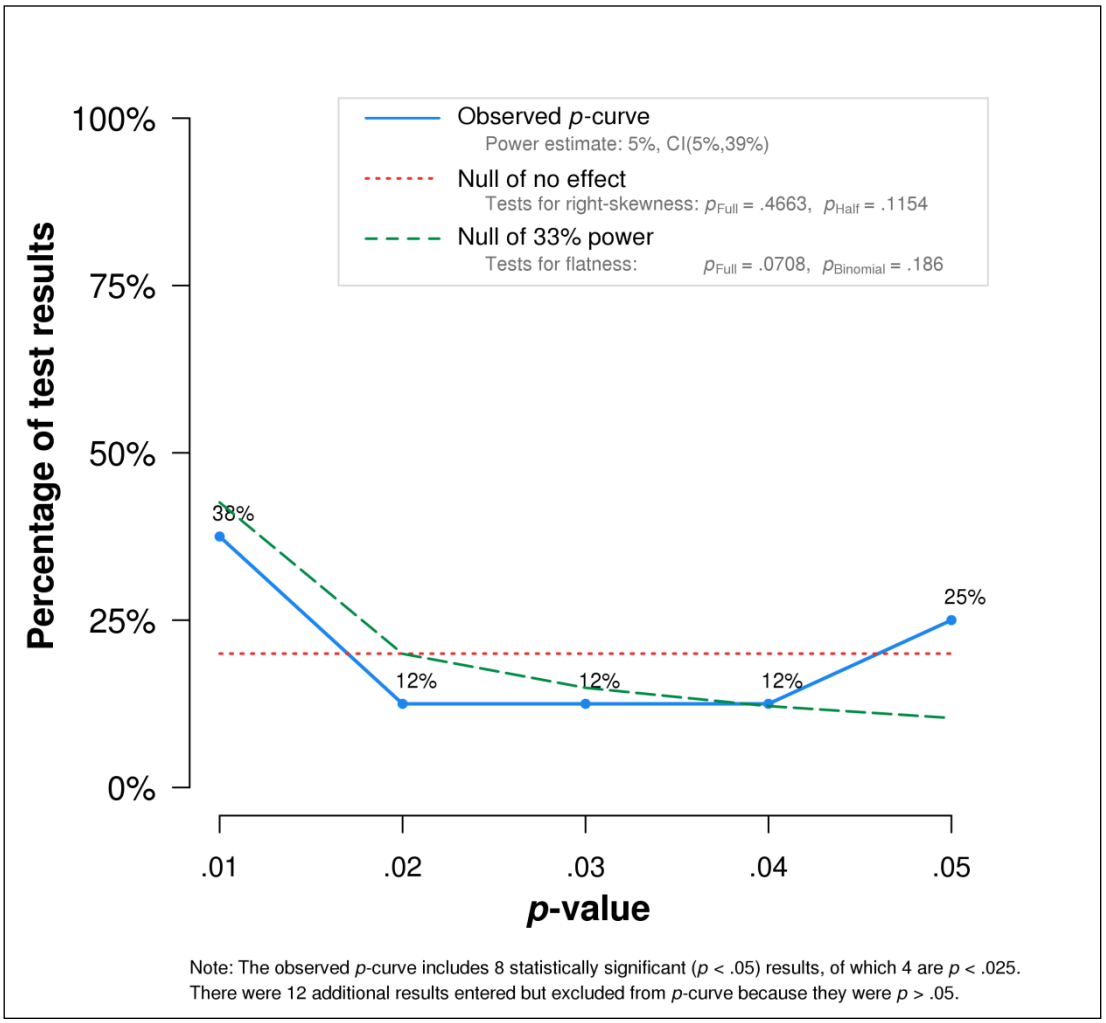

Figure 10: Percentual distribution of observed statistically significant *p*-values when excluding Macías-Cortés (2015). The test statistic is provided on the upper right.

Table 2: Summary of effect size estimates

	Conventional meta-analysis	*p*-uniform	*p*-curve
Effect size (*d*)	-0.46	-0.28	0.71
Confidence interval	[-0.89;-0.04]	[-0.74;0.70]	-
p-value	.03	.21	-

3.4. Secondary analysis

Table 3: Comparison of main findings in subsets split by journal affiliation

	p-uniform		*p*-curve			conventional
	Effect Size	Publication bias	Effect Size	*p*-hacking	Power	Effect Size
Homeopathy-affiliated	1.88[a]	Yes	-0.59	No	5%	-0.19[c]
Non homeopathy-affiliated	-0.86[b]	No	0.87	No	96%	-0.90 [c]

[a] $p = .97$

[b] $p = .002$

[c] $p < .01$

No evidence for heterogeneity was found homeopathy-affiliated subset (I^2= 13.27%, Q = 13.47, p = .26). For this subset *p*-uniform estimated an effect size of d = 1.88 (p = .97), while *p*-curve yielded an effect size estimate of d = -0.59. Conventional fixed effect size estimation amounted to d = -0.19 (p = .06). Both estimates based on *p*-values suggested no evidential value in the data set, leading to the same interpretation of no population effect. However, a clear difference in the distribution of *p*-values is depicted in Figure 11 and Figure 12. While the non homeopathy-affiliated subset was right-skewed (p < .001), the homeopathy-affiliated subset was flatter than it would be if the study set was powered at 33%. Examining the homeopathy-affiliated subset in detail, no *p*-value less than .001 was observed. Most observed *p*-values were near the threshold for statistically significance. While the shape of the curve may be a hint on *p*-hacking, the statistical test indicated flatness (p < .01).

For the non homeopathy-affiliated subset results yielded by *p*-curve and *p*-uniform contradicted each other. *p*-curve suggested a large effect in favour for placebo, while *p*-uniform suggested a large effect in favour of individualized homeopathic treatment. This may be explained by heterogeneity being present in this subset (I^2= 94.20%, Q = 75.92, p < .001), which can lead to unreasonable results.

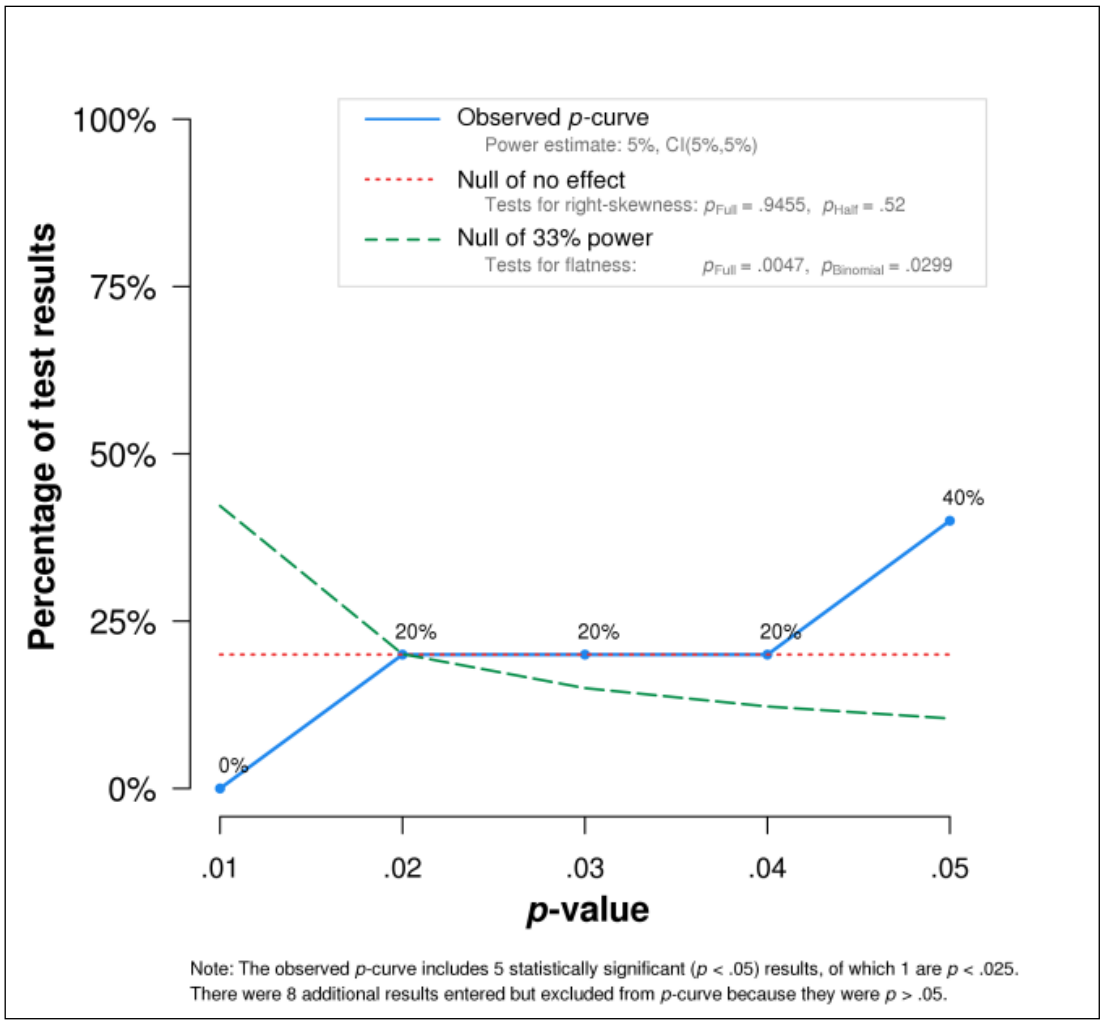

Figure 11: Percentual distribution of observed statistically significant *p*-values in the homeopathy-affiliated subset. The test statistic is provided on the upper right.

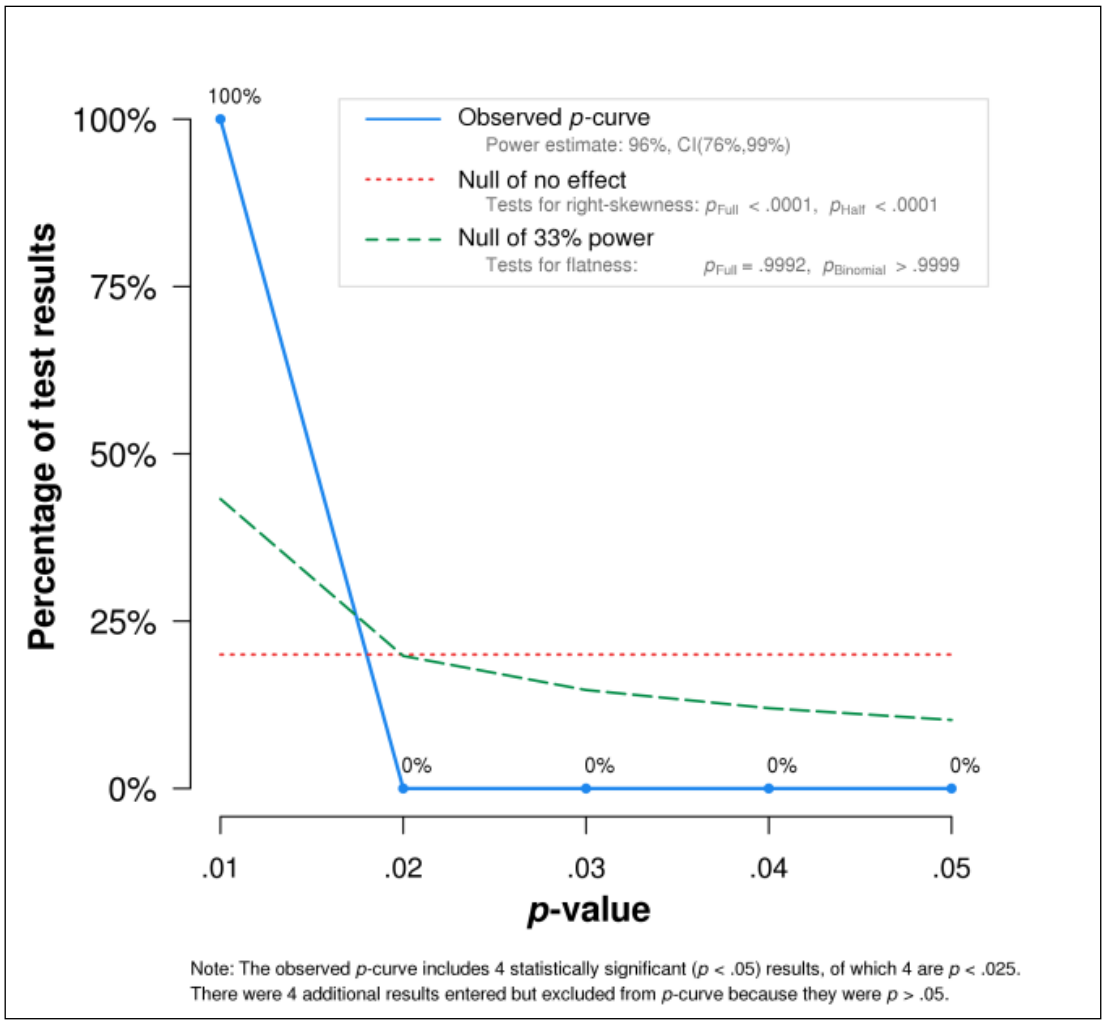

Figure 12: Percentual distribution of observed statistically significant *p*-values in the non homeopathy-affiliated subset. The test statistic is provided on the upper right.

4. Discussion

4.1. Effect sizes

p-curve yielded a large positive effect size and *p*-uniform returned a small negative effect size. Neither effect size was statistically significant. Although the direction and size of effects differed between *p*-curve and *p*-uniform, they both consistently suggested no evidential value in the data set and thus no effect of homeopathy.

As *p*-curve has more power of detecting evidential value than the individual studies from which the *p*-values are extracted, it is "virtually guaranteed to detect evidential value, even when the set of studies is powered at just 50%" (Simonsohn et al., 2013). Simulation studies showed that the risk for a beta error is very low even with just a few *p*-values if a study set is powered at around 80% (Simonsohn et al., 2013). However, the entered study set did not reach any of these numbers. Van Aert et al. (2016) stated that low overall statistical power in the data set and *p*-values just below the significance threshold are signals for *p*-hacking. Since no evidence for publication bias was found in conventional methods, *p*-hacking provides a sensible explanation for the prevalence of statistically significant *p*-values within a study set that is underpowered at 38%. Furthermore, the low statistical power of the study set could not fully explain the misfit between the expected distribution of *p*-values in case of low statistical power and the observed distribution. Here, too, *p*-hacking is a sensible explanation for the observed misfit.

This conclusion differes from the face value interpretation drawn from the conducted conventional meta-analysis, which suggested a moderate effect in favour of homeopathy. The conventional statistical tests for publication bias were not statistically significant. However, the trim and fill method suggested seven missing studies with large effect sizes on the left side of the funnel. It is very unlikely that such unpublished studies exist. Splitting the study set by journal affiliation and conducting the trim and fill method for each subset seperately showed two different distributions of effect sizes. This means that a mixed distribution within the whole study set potentially distorted the results drawn from convential meta-analysis. Furthermore, sensitivity analysis suggested distortion of a single study when conducting conventional meta-analysis. Excluding from the analysis the study with the largest effect size, which had a very small n, lowered the effect size by 32%.

4.2. Subset analysis

Two subsets were created and studies were assigned to them depending on journal affiliation. The subsets were comparable in sample size and duration of treatment.

Within the homeopathy-affiliated subset, no effect was found by *p*-uniform. However, *p*-curve yielded a large effect size in favour of homeopathy. *p*-uniform's test for publication bias was statistically significant. Furthermore, extremely low statistical power (5% [5%;5%]) and an unproportionally high number of *p*-values barely below .05 were observed in this subset, suggesting that the effect size estimate by *p*-curve is not valid.

No such evidence was found in the non homeopathy-affiliated subset, suggesting that intense *p*-hacking mainly takes place within journals that are affiliated with homeopathy. For the non homeopathy-affiliated subset, effect size estimates by *p*-curve and *p*-uniform differed by over one and a half standard deviations. This inconsistency may be explained by heterogeneity, which can lead to unreasonable results in both methods.

Detailed examination of the subjected study set with the conventional trim and fill method showed that there seemed to be a publication bias in a certain subset. This subset contained studies published in journals not affiliated with homeopathy. Two studies with large effect sizes and high standard error were potentially missing within this set, whereas no such bias was found within the homeopathy-affiliated subset. This indicates that journals not affiliated with homeopathy published unproportionally less studies that found large effect sizes in favour of homeopathy. This is in line with the general perception that prestigious journals with a high impact factor require studies to report large effect sizes in order to publish them. Accordingly, journals with a lower impact factor, which are homeopathy-affiliated journals in this case, published studies that tended to find low effect sizes in favour of homeopathy. Additionally, these studies were generally underpowered. The discrepancy between these two types of journals was reflected in both conventional meta-analysis and *p*-value based methods.

4.3. Conclusion

This study showed that homeopathy has no effect superior to a placebo effect. Subgroup analyses using *p*-value-based methods as well as conventional methods suggested that the effect size estimate yielded by conventional meta-analysis is a product of *p*-hacking and publication bias within journals affiliated with homeopathy, whereas statistically significant

results reported by studies published in journals not affiliated with homeopathy are likely chance effects. The interpretation of face value estimates yielded by the conducted conventional meta-analysis is therefore strongly discouraged.

Additionally, interpreting single homeopathy studies would not or would just partially reflect the nature of the examined true effect. Therefore, it is recommended to discuss the results of homeopathy-supporting as well as homeopathy-discrediting studies only in the context of detailed meta-analytic examination. When doing this, the effects of homeopathy, as discussed before, are likely attributable to placebo effects.

Conventional meta-analyses can deal with the presence of publication bias when the face value interpretation is examined in more detail. However, they have a blind spot for QRPs. *P*-value-based approaches adress those problems by default.

Similar distortions may be present in other research domains as well. In future studies, *p*-curve and *p*-uniform should at least be used as additional analysis, as their statistical approach is a helpful and intuitive tool to examine possible *p*-hacking. Simulation studies (van Aert et al., 2016) suggested that their precision on effect size estimation is superior to conventional meta-analysis. The present study put that into context by applying the methods on real data and underpins these claims. Taking all this together, *p*-curve and *p*-uniform are a relevant alternative to conventional methods.

5. Summary

This thesis examined the application, plausibility and interpretation of *p*-value-based effect size estimations compared to conventional meta-analysical estimates. *p*-value-based effect size estimations are needed because publication bias and questionable research practices are highly prevalent and are a possible threat to the validity of conventional meta-analytical results and empirical investigations in general.

For the present analysis the body of homeopathy literature was searched and studies fitting an a priori set of inclusion criteria were coded. Three different meta-analytical methods were used to examine the present set of 21 studies. Two were based on *p*-values, one was based on a random effects model. Compared to conventional meta-analysis, information needed for the analyses were easier to code for *p*-value-based methods. Also, since some studies did not report sufficient test statistics, e.g., means and standard deviations, but provided *p*-values, more studies could be included in the calcuations of *p*-value based methods. Application on these data yielded results that were similar to results obtained by simulation studies and therefore allowed interpretations of the obtained effect sizes.

The effect size estimates obtained from the three applied methods varied and indicated different interpretations of the estimate. While conventional meta-analysis yielded a medium population effect in favour of homeopathic treatment with indication of publication bias, both *p*-value-based methods suggested no evidential value and indicated low power for the data set. Additionally, possible *p*-hacking within the homeopathy-affiliated subset was detected with *p*-value-based methods. While some meta-analyses of homeopathic effects are careful in their conclusion and some studies term their results inconclusive, the present analysis strongly indicates that homeopathy has, in fact, no effect.

The subset analyses showed that the homeopathy-affiliated subset contained no evidential value and was severely underpowered. Studies with statistically significant results may have obtained them by *p*-hacking. On the other hand, studies published in journals not affiliated with homeopathy seemed to contain evidential value and had sufficient statistical power. However, the effects found by studies within this subset were likely obtained by chance.

In essence, the results indicated that statistically significant *p*-values were either obtained by *p*-hacking or by chance.

p-value-based effect size estimations are useful meta-analytical methods. When heterogeneity is low within a data set, they have sufficient precision compared to conventional methods and can deal with publication bias and *p*-hacking. Additionally, this study contributes evidence to the scientific field that homeopathy, as stated by the homeopathic doctrine, has no effect.

5.1. References

Adler, U. C., Krüger, S., Teut, M., Lüdtke, R., Schützler, L., Martins, F., Willich, S. N., Linde, K., & Witt., C. M. (2013). Homeopathy for depression: A randomized, partially double-blind, placebo-controlled, four-armed study (DEP-HOM). *PLOS ONE, 9*. doi:10.1371/journal.pone.0074537

Albrecht, H., & Maier, J. (2017, March 30). Sollen Krankenkassen Homöopathie bezahlen? Zeit Online. Retrieved from: http://www.zeit.de/2017/12/medizin-homoeopathie-krankenkasse-zahlung-pro-contra

Andrade, L. E. C., Ferraz, M. B., Atra, E., Castro, A., & Silva, M. S. M. (1991). A randomized controlled trial to evaluate the effectiveness of homeopathy in rheumatoid arthritis. *Scandinavian Journal of Rheumatology, 20*, 204-208.

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives, 11*, 109-126.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science, *Perspectives on Psychological Science, 7*, 543-554. doi: 10.1177/1745691612459060

Banks, G. C., Kepes, S., & Banks, K. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *American Educational Research Association, 34*, 259-277.

Banks, G. C., O'Boyle, E. H. Jr., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. (2016a). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management, 42*, 5-20.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016b). Editorial: Evidence on questionable practices: The good, the bad and the ugly. *Journal of Business Psychology, 31*, 323-338. doi: 10.1007/s10869-016-9456-7

Becker, B. J. (2005). The failsafe N or file-drawer number. In H. R, Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (111-126). Chichester, UK: Wiley.

Bell, I. R., Lewis, D. A., Lewis, S. E., Schwartz, G. E., Brooks, A. J., Scott, A., & Baldwin, C. M. (2004). EEG alpha sensitization in individualized treatment of fibromyalgia. *International Journal of Neuroscience, 114*, 1195-1220. doi: 10.1080/00207450490475724

Bell, I. R., Lewis, D. A., Brooks, A. J., Schwartz, G. E., Lewis, S. E., Walsj, B. T., & Baldwin, C. M. (2004). Improved clinical status in fibromyalgia patients treated with individualized homeopathic remedies versus placebo. *Rheumatology, 43*, 577-582. doi: 10.1093/rheumatology/keh111

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088-1101.

Bishop, D. V. M., & Thompson, P. A. (2016). Problems in using *p*-curve analysis and text-mining to detect rate of *p*-hacking and evidential value. *PeerJ*. doi: 10.7717/peerj.1715

Boehm, K., Raak, C., Cramer, H., Lauche, R., & Ostermann, T. (2014). Homeopathy in the treatment of fibromyalgia: A comprehensive literature-review and meta-analysis. *Complementary Therapies in Medicine, 22*, 731-742. doi: 10.1016/j.ctim.2014.06.005

Dickersin, K., & Min, Y. I. (1993). Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences, 703*, 135-148.

Bonne, O., Shemer, Y., Gorali, Y., Katz, M., & Shalev, A. (2003). A randomized, double-blind, placebo-controlled study of classical homeopathy in generalized anxiety disorder. *Journal of Clinical Psychiatry, 64*, 282-287.

Brien, S., Lachance, L., Prescott, P., McDermott, C., & Lewith, G. (2011). Homeopathy has clinical benefits in rheumatoid arthritis patients that are attributable to the consultation process but not the homeopathic remedy: A randomized controlled clinical trial. *Rheumatology, 50*, 1070-1082. doi: 10.1093/rheumatology/keq234

Cavalcanti, A. M. S., Rocha, L. M., Carillo, R., Lima, L. U. O., & Lugon, J. R. (2003). Effects of homeopathic treatment on pruritus of haemodialysis patients: A randomised placebo-controlled double-blind trial. *Homeopathy, 92*, 177-181. doi: 0.1016/j.homp.2003.07.002

Chand, K. S., Manchanda, R. K., Mittal, R., Batra, S., Banavaliker, J. N., & De, I. (2014). Homeopathic treatment in addition to standard care in multi drug resistant pulmonary tuberculosis: A randomized, double blind, placebo controlled clinical trial. *Homeopathy, 103*, 97-107. http://dx.doi.org/10.1016/j.homp.2013.12.003

Chapman, E. H., Weintraub, R. J., Milburn, M. A., O'Neil Pirozzi, T., & Woo, W. (1999). Homeopathic treatment of mild traumatic brain injury: A randomized, double-blind, placebo-controlled clinical trial. *The Journal of Head Trauma Rehabilitation, 14*, 521-542.

Duval, S., & 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.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634.

Ernst, E. (1998). The heresy of homoeopathy. A brief history of 200 years of criticism. *British Homeopathy Journal, 87*, 28-32.

Ernst, E. (2002). A systematic review of systematic reviews of homeopathy. *Journal of Clinical Pharmacology, 54*, 557-582.

Ernst, E. (2011). Homeopathic galphimia glauca for hay fever: A systematic review of randomised clinical trials and a critique of a published meta-analysis. *Focus on Alternative and Complementary Therapies, 16*, 200-203. doi: 10.1111/j.2042-7166.2011.01084.x

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE, 5*.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*, 45-52. doi: 10.1177/1948550615612150

Fisher, P., Carney, R., Hasford, C., & Vickers, A. (2006). Evaluation of specific and non-specific effects in homeopathy: Feasibility study for randomised trial. *Homeopathy, 95*, 215-222. doi:10.1016/j.homp.2006.07.006

Frank, R. (2015). Die kulturelle Kontextualisierung der Homöopathie. In Frank, R., *Globalisierung »alternativer« Medizin* (65-156). Bielefeld: Transcript Verlag.

Frass, M., Linkesch, M., Banyai, S., Resch, G., Dielacher, C., Löbl, T., Endler, C., Haidvogl, M., Muchitsch, I., & Schuster, E. (2005). Adjunctive homeopathic treatment in patients with severe sepsis: a randomized, double-blind, placebo-controlled trial in an intensive care unit. *Homeopathy, 94*, 75-80. doi:10.1016/j.homp.2005.01.002

Gartlehner, G. (2016, November 18). Nein zur Homöopathie als Kassenleistung. derStandard.at. Retrieved from: http://derstandard.at/2000047729458/Nein-zur-Homoeopathie-als-Kassenleistung

Goossens, M., Laekeman, G., Aertgeerts, B., & Buntinx, F. (2009). Evaluation of the quality of life after individualized homeopathic treatment for seasonal allergic rhinitis. A prospective, open, non-comparative study. *Homeopathy, 98*, 11-16. doi: 10.1016/j.homp.2008.11.008

Gray, B. (2016). How should we respond to non-dominant healing practices, the example of homeopathy. *Journal of Bioethical Inquiry*, *14*, 87-96.

Hahn, R. G. (2013). Homeopathy: Meta-analyses of pooled clinical data. *Forschende Komplementärmedizin, 20*, 376-381. doi: 10.1159/000355916

Hahnemann, S. (1796). Versuch über ein neues Prinzip zur Auffindung der Heilkräfte der Arzneisubstanzen, nebst einigen Blicken auf die bisherigen. In Hufeland, C. W. (Eds.). *Journal der practischen Arzneykunde und Wunderarzneykunst*. Jena: Acedemische Buchhandlung

Hedges, L. V., & Vevea, J. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (75-98). Chichester, UK: Wiley.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007a). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal, 176*, 1091-1096.

Ioannidis, J. P. A. & Trikalinos, T. A. (2007b). An exploratory test for an excess of significant findings. *Clinical Trials, 4*, 145-253. doi: 10.1177/1740774507079441

Karp, J.-C., Sanchez, C., Guilbert, P., Mina, W., Demonceaux, A., & Curé, H. (2016). Treatment with Ruta graveolens 5CH and Rhus toxicodendron 9CH may reduce joint pain and stiffness linked to aromatase inhibitors in women with early breast cancer: Results of a pilot observational study. *Homeopathy, 105*, 299-308. doi: 10.1016/j.homp.2016.05.004

Kicinski, M. (2013). Publication bias in recent meta-analyses. *PLoS ONE, 11*. doi:10.1371/journal.pone.0081823

Koley, M., Saha, S., & Ghosh, S. (2015). A double-blind, randomized, placebo-controlled feasibility study evaluating individualized homeopathy in managing pain of knee osteoarthritis. *Journal of Evidence-Based Complementary & Alternative Medicine, 20*, 186-191. doi: 10.1177/2156587214568668

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge: Harvard University Press.

Linde, K., & Melchart, D. (1998). Randomzied controlled trials of individualized homeopathy: A state-of-the-art review. *Journal of Alternative and Complementary Medicine, 4*, 371-388.

Lipsey, M. W., & Wilson, D. B. (2001). Problem specification and study retrieval. In: Lipsey, M. W., & Wilson, D. B., *Practical meta-analysis*. London: Sage Publications

Macías-Cortés, E. d. C., Llanes-González, L., Aguilar-Faisal, L., & Asbun-Bojalil, J. (2015). Individualized homeopathic treatment and fluoxetine for moderate to severe depression in pero- and postmenopausal women (HOMDEP-MENOP study): A randomized, double-

dummy, double-blind, placebo-controlled trial. *PLoS ONE, 10*, doi:10.1371/journal.pone.0118440

Mathie, R. T., Lloyd, S. M., Legg, L. A., Clausen, J., Moss, S., Davidson, J. R. T., & Ford, I. (2014). Randomised placebo-controlled trials of individualised homeopathic treatment: Systematic review and meta-analysis. *Systematic Reviews, 3*, 1-16. doi: 10.1186/2046-4053-3-142

McShane, B. B., Böckenholt, U., & 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. doi: 10.1177/174569161666224

Mousavi, F., Mojaver, Y. N., Asadzadeh, M., & Mirzazadeh, M. (2009). Homeopathic treatment of minor aphtous ulcer: A randomized, placebo-controlled clinical trial. *Homeopathy, 98*, 137-141. doi:10.1016/j.homp.2009.05.006

National Health and Medical Research Council (2015). *Administrative Report: NHMRC Advice on the effectiveness of homeopathy for treating health conditions.* Canberra: National Health and Medical Research Council

Naude, D. F., Couchman, I. M. S., & Maharaj, A. (2010). Chronic primary insomnia: Efficacy of homeopathic simillimum. *Homeopathy, 99*, 63-68. doi: :10.1016/j.homp.2009.11.001

Parekh-Bhurke, S., Kwok, C. S., Pang, C., Hooper, L., Loke, Y. K., Ryder, J. J., Sutton, A. J., Hing, C. B., Harvey, I., & Song, F. (2011). Uptake of methods to deal with publication bias in systematic reviews has increased over time, but there is still much scope for improvement. *Journal of Clinical Epidemiology, 64*, 349-357.

Peckham, E. J., Relton, C., Raw, J., Walters, C., Thomas, K., Smith, C., Kapur, K., & Said, E. (2014). Interim results of a randomised controlled trial of homeopathic treatment for irritable bowel syndrome. *Homeopathy, 103*, 172-177. doi: http://dx.doi.org/10.1016/j.homp.2014.05.001

Rastogi, D. P., Singh, V. P., Sing, V., Dey, S. K., & Rao, K. (1999). Homeopathy in HIV infection: A trial report of double-blind placebo controlled study. *British Homeopathic Journal, 88*, 49-57.

Relton, C., Cooper, K., Viksveen, P., Fibert, P., & Thomas, K. (2017). Prevalence of homeopathy use by the general population worldwide: A systematic review. *Homeopathy, 106*, 69-78. http://dx.doi.org/10.1016/j.homp.2017.03.002

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638-664. dx.doi.org/10.1037/0033-2909.86.3.638

Rothstein, H. R. (2007). Publication bias is a threat to the validity of meta-analytic results. *Journal of Experimental Criminology, 4*, 61-81. doi: 10.1007/s11292-007-9046-9

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge: Hogrefe & Huber Publishers

Siebenwirth, J., Lüdtke, R., Remy, W., Rakoski, J., Borelli, S., & Ring, J. (2009). Wirksamkeit einer klassisch-homöopathischen Therapie bei atopischem Ekzem. *Forschende Komplementärmedizin, 16*, 315-323.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359-1366. doi: 10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). *p*-curve: A key to the file drawer. *Journal of Experimental Psychology, 143*, 534-547. doi: 10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666-981. doi: 10.1177/1745691614553988

Springer, G. (2017, January 25). Krankenkassen zahlen zwecks Placeboeffekts für Globuli und Co. derStandard.at. Retrieved from: http://derstandard.at/2000051489910/Krankenkassen-zahlen-zwecks-Placeboeffekts-fuer-Globuli-und-Co

Sterne, J. A., Gavaghan, D., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments*. Chichester: Wiley

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rcker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 343*. doi: 10.1136/bmj.d4002

Straumsheim, P., Borchgrevink, C., Mowinckel, P., Kierulf, H., & Hafslund, O. (2000). Homeopathic treatment of migraine: A double blind, placebo controlled trial of 68 patients. *British Homeopathic Journal, 89*, 4-7.

Thompson, E. A., Oxon, B. A., Montgomery, A., Douglas, D., & Reilly, D. (2005). A pilot, randomized, double-blinded, placebo-controlled trial of individualized homeopathy for symptoms of estrogen withdrawal in breast-cancer survivors. *The Journal of Alternative and Complementary Medicine, 11*, 13-20.

Torgerson, C. (2003). Publication Bias. In *Systematic Reviews*. London: Continuum

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on *p*-values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science, 11*, 713-729. doi: 10.1177/1745691616650874

van Assen, M. A. L. M., Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only significant studies. *Psychological Methods, 20*, 293-309.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*.

Wadhwani, G. G. (2013). Homeopathic drug therapy: Homeopathy in Chikungunya fever and post-chikungunya chronic arthritis: An observational study. *Homeopathy, 102*, 193-198. doi: 10.1016/j.homp.2013.02.001

Weber, E. J., Callaham, M. L., Wears, R. L., Barton, C., & Young, G. (1998). Unpublished research from a medical specialty meeting: Why investigators fail to publish. *Journal of the American Medical Association, 280*, 257-259.

Weber, N. (2017, March 7). Wieso zahlen Krankenkassen Homöopathie? Spiegel Online. Retrieved from: http://www.spiegel.de/gesundheit/diagnose/homoeopathie-warum-zahlt-die-krankenkasse-a-1137637.html

Yakir, M., Kreitler, S., Brzezinski, A., Vithoulkas, G., Oberbaum, M., & Bentwich, Z. (2001). Effects of homeopathic treatment in women witch premenstrual syndrome: A pilot study. *British Homeopathic Journal, 90*, 148-153.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014).When decision heuristics and science collide. *Psychonomic Bulletin and Review Journal, 21*, 268-282. doi: 10.3758/s13423-013-0495-z

5.2. List of Figures

5.2. List of Tables