university of groningen / faculty of behavioural and social sciences / research master

Master’s thesis

Therapy or Medication: Comparing Strength of Evidence of Psychological and Pharmacological Depression Treatments

Name and initials:	Keano Bruns (KB)
Student number:	S4372220
E-mail address:	k.bruns.1@student.rug.nl
First assessor:	Prof. Dr. Don van Ravenzwaaij
Second assessor:	Prof. Dr. Laura Bringmann
Programme:	Research Master Behavioural and Social Sciences
Theme:	Psychometrics & Statistics
ECs:	30
Date:	June 27, 2025
Word count:	6881

Are there deviations of the Master’s thesis from the proposed plan?
☐No
xYes, please explain below the deviations
Information of individual studies was not combined using Bayesian model-averaged meta-analysis. The individual study Bayes Factors were used to answer the research questions.

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.

Abstract

Depression is one of the most common mental disorders and requires effective, evidence-based treatments to reduce its burden. Two main kinds of approaches exist for treatment of mental disorders: 1) pharmacological medications and 2) psychological therapies. In this thesis, the two approaches have been explored and compared in terms of their evidential strength. To this end, data on the effects of 113 trials for 16 therapies and 128 trials for 21 antidepressants was collected and analysed. For each clinical trial a Bayes Factor was calculated to quantify how well treatment efficacy is supported by the research evidence. Results showed that therapies generally displayed more consistent evidential strength than medications against control groups. However, this finding was potentially biased by selectivity effects, as therapy trials, contrary to medication trials, are not required to be pre-registered or directly appeal to a governing body for approval. Within psychotherapies, the average strength of evidence was not consistent with the current evaluation labels: therapies labeled as having ‘modest’ research support surpassed therapies labeled as ‘strong’ in their evidential strength. Ultimately, a measure of strength of evidence such as the BF may aid clinical decision making by providing additional information about the evidence for, and the efficacy of mental health treatments.

Keywords: Strength of Evidence, Bayes Factor, Depression, Psychotherapy, Antidepressants

Therapy or Medication: Comparing Strength of Evidence of Psychological and Pharmacological Depression Treatments

Over the past years, there has been a steady increase in the prevalence of mental disorders. This trend was observed across many countries, including the Netherlands (ten Have et al., 2023) and in Germany (Thom et al., 2024), and involved many different diagnoses. Among these was also one of the most common disorders, depression (Moreno-Agostino et al., 2021), which was previously described by the World Health Organization as “the leading cause of ill health and disability worldwide” (World Health Organization, 2017).

This substantial impact of depression on global health necessitates effective treatment options for those affected. Traditionally, treatment for depression follows one of two broad approaches: 1) pharmacological interventions (i.e. medication) or 2) psychotherapy. The two approaches can also be combined or used sequentially, often resulting in better outcomes than either treatment approach alone (Karyotaki et al., 2016). Both kinds of intervention aim to reduce depressive symptoms and improve general functioning, but historically pharmacological treatments were often perceived as superior to psychological treatments (American Psychological Association [APA] Presidential Task Force on Evidence-Based Practice, 2006). This included the perception of psychotherapy as “a livelier experience of the placebo effect than is available in medicine” (Justman, 2010).

Pharmacological medications were first to introduce a set of guidelines to assess the efficacy of treatments and control approval or marketing processes. To date, for medications to be marketed, regulatory standards, such as those employed by the U.S. Food and Drug Administration (FDA), must be met. Among other requirements, the FDA requires that a drug’s efficacy must be established through at least two independent, well-controlled clinical trials demonstrating statistically significant results

(FDA, 1998). In cases where conducting multiple trials is not feasible, evidence from a single clinical trial may also be sufficient for approval (Food and Drug Administration, 2022). These results are then considered to be substantial evidence for the efficacy of the drug.

Psychological therapies later followed this sentiment, expressing a “fundamental commitment” described by the APA (APA Presidential Task Force on Evidence-Based Practice, 2006). This commitment calls for the integration of the best available research evidence with clinical expertise and patient preferences to inform treatment decisions. Treatments based on this combination are called evidence-based treatments. Research evidence for a treatment's efficacy also became central in the evaluation process for the treatment, with the Society of Clinical Psychology (SCP), a division of the APA, further categorizing interventions depending on the available evidence. Initially, these categories have largely been based on the number of studies showing statistically significant effects (Chambless & Hollon, 1998), using the categories “strong” (at least two statistically significant results by independent research teams), “modest” (one statistically significant result or multiple by the same team), or “controversial” (conflicting results). Alternatively, it was possible to reach the thresholds through a series of well-designed single-case studies. The current guidelines by Tolin et al. (2015) focus more on the quality of evidence as derived from systematic reviews, and adapted the categories to “very strong” (high quality evidence), “strong” (moderate to high quality evidence), “weak” (low or very low quality of evidence), or having “insufficient evidence” (no meta-analytic study or it is of too low quality). However, most therapies are still only evaluated using the older guidelines.

Despite the development of similar evaluation criteria for both kinds of intervention, perceptions of differential effectiveness of the two approaches seemed to persist. While the general population seemed to belief that therapy is more effective than medication (Silverman et al., 2021), previous research indicated an increase in pharmacological drug prescriptions at the expense of psychotherapies (Gaudiano & Miller, 2013), possibly still reflecting the perception of clinicians that pharmacological interventions are more effective. To formally assess the differences in efficacy between psychological and pharmacological interventions, the two approaches have often been compared in terms of effect sizes (eg., Leichsenring et al., 2022). However, comparisons of their evidential standards are currently lacking in the literature. While a comparison of effect sizes assesses the question of which treatment is more effective (i.e., the magnitude of the treatment effect), a comparison of evidential standards assesses whether different assessment approaches result in similar outcome standards. Such comparisons may give insight into how well regulatory guidelines reflect evidence-based treatments in practice, and guide clinical decision-making.

One way of assessing and comparing the evidential standards is through the use of Bayes Factors (BFs) as a measure of *strength of evidence* (or evidential load). This measure refers to the degree to which the efficacy of the treatment is supported by the available evidence (Monden et al., 2018), and gives an indication of how likely an effect is to exist (Pittelkow et al., 2021). The BF allows for the quantification of evidence in favour of one hypothesis over another, typically an alternative hypothesis (H_1) versus the null hypothesis (H_0). It represents the ratio between the predictive evidence of the two competing hypotheses given the data by comparing the relative likelihood of the data occurring under each hypothesis (Jeffreys, 1961; van Ravenzwaaij & Ioannidis, 2019). For example, a BF_{10} of 10 (the subscript denotes that we are evaluating the probability of the data given the alternative hypothesis relative to the null hypothesis) indicates that the data are ten times more likely under H_1 than H_0. A BF_{10} of 0.1 on the other hand indicates that the data are ten times more likely under H_0 than H_1. Thus, a BF of 1 indicates that the data are equally likely under either hypothesis.

Mistakenly, *p*-values, on which the operationalizations of evidence-based treatments rely, are often assumed to indicate the strength of evidence of a given (treatment) effect. In the null hypothesis significance testing (NHST) framework (of which *p*-values are a part), a treatment is considered efficacious when the *p*-value lies below a predetermined alpha level (commonly $p < .05$), indicating that the observed effect is unlikely under the null hypothesis (i.e., the treatment and control group are the same). However, indicating the strength of evidence requires the implementation of two competing hypotheses explaining the observed data (Goodman & Royall, 1988). Since NHST only considers the null hypothesis, *p*-values are “logically flawed” as a measure of strength of evidence (Hubbard & Lindsay, 2008). The BF on the other hand can be interpreted bidirectionally, as it enables the differentiation between lack of evidence for an effect and evidence for the absence of an effect (Beard et al., 2016). Additionally, the sensitivity of the *p*-value to sample size and statistical power leads them to give an inconsistent interpretation of the strength of evidence (Lakens, 2022). The same *p*-value can indicate different conclusions about the evidential strength depending on the power of the test. This variability undermines the comparability of findings across studies.

With regard to the criteria of the FDA and the SCP for evidence-based treatments, it has previously been shown that different cases in which two trials achieve a statistically significant *p*-value can have remarkably different BFs (van Ravenzwaaij & Ioannidis, 2017), including cases in which the evidence is actually in favour of the null hypothesis. One reason for obtaining two statistically significant results despite evidence favouring the null hypothesis is a large number of total trials. This is not accounted for in the evaluation criteria of the FDA or the old criteria of the SCP, which require two statistically significant trials regardless of the total number of trials. The current criteria of the SCP aim to address this problem by shifting the focus on systematic combination of trials, but many psychological therapies are not yet evaluated by these newer guidelines.

Partly because of findings like these, the prevailing reliance on NHST and *p*-values to determine the evidence-base has been questioned (Goodman, 1999; Wagenmakers, 2007; Ahmed & Butt, 2025). The NHST framework is suboptimal to determine the evidence-base when considered alone (Sakaluk et al., 2019), as it was highlighted that there may exist considerable differences between metrics of evidence within a treatment. Also, two treatments may differ substantially in terms of evidence for efficacy, despite having the same *p*-value (Monden et al., 2016). These findings, together with the arguments made above, raise questions about the consistency of evaluations made under current guidelines, both psychological and pharmacological. In line with the commitment to evidence-based practice, an exploration and comparison of the strength of evidence via the BF might give insights into the current evaluation standards.

Previous research has started to investigate the strength of evidence across clinical studies, particularly for pharmacological interventions. Monden and colleagues (2016) examined the evidential strength of antidepressants for anxiety disorders and found that even among trials meeting FDA standards for "substantial evidence" many did not show strong support for efficacy in terms of strength of evidence. A follow-up study on antidepressants for depression reported similar heterogeneity, although many drugs did exhibit strong evidential support overall (Monden et al., 2018). Pittelkow et al. (2021) extended this line of research with a Bayesian meta-analysis of several psychotropic drugs, finding generally strong evidential support but also identifying some drugs with only moderate or ambiguous evidence despite approval for clinical use. Investigations of evidential strength across drug subclasses are still currently missing in the literature.

In contrast to pharmacological drugs, much less is known about the evidential strength of psychological therapies despite its usefulness for drawing conclusions about the evidential support of treatment options. While multiple treatments may have similar effect sizes, they might differ in strength of evidence in favor of their efficacy (Monden et al., 2018), consequently being a helpful tool in choosing the treatment with the best level of evidential strength from multiple options with similar effect sizes. It is essential for guiding clinical decision-making to understand not only the size of the treatment effect, but also how well that treatment is supported by the evidence. One example is given by a meta-scientific review of the evidential strength of Acceptance and Commitment Therapy for depression (Williams et al., 2023). Their results show that this therapy was efficacious as a depression treatment when compared to no treatment but lacked evidential support in comparison to other kinds of psychological therapies. Another meta-scientific review assessed selected empirically supported treatments (ESTs) across multiple metrics and found that therapies classified as "strong" under the Chambless and Hollon (1998) criteria failed to continuously surpass therapies classified as "modest" when considering the BF as a measure of strength of evidence (Sakaluk et al., 2019).

Despite their usefulness, evaluations of the strength of evidence for therapies and comparisons to pharmacological interventions are still mostly missing. In this context, the current study seeks to address this gap in the existing literature by exploring the evidential strength of psychological and pharmacological treatments. Specifically, there are three research questions that are investigated:

RQ1) Are current recommendations of psychological treatments for depression consistent with the strength of evidence of results reported in the underlying clinical studies?

RQ2) Are there differences in the strength of evidence between medication subclasses?

RQ3) Are there differences in the strength of evidence for psychological treatments compared to pharmacological treatments of depression?

Importantly, this study was highly data-driven and exploratory in nature, with the research questions guiding the exploration process. By addressing these questions, this study aims to contribute to the understanding of evidence-based depression treatments of both psychological and pharmacological nature, and how comparable the two approaches are in terms of evidential strength.

Method

Data Sources and Extraction

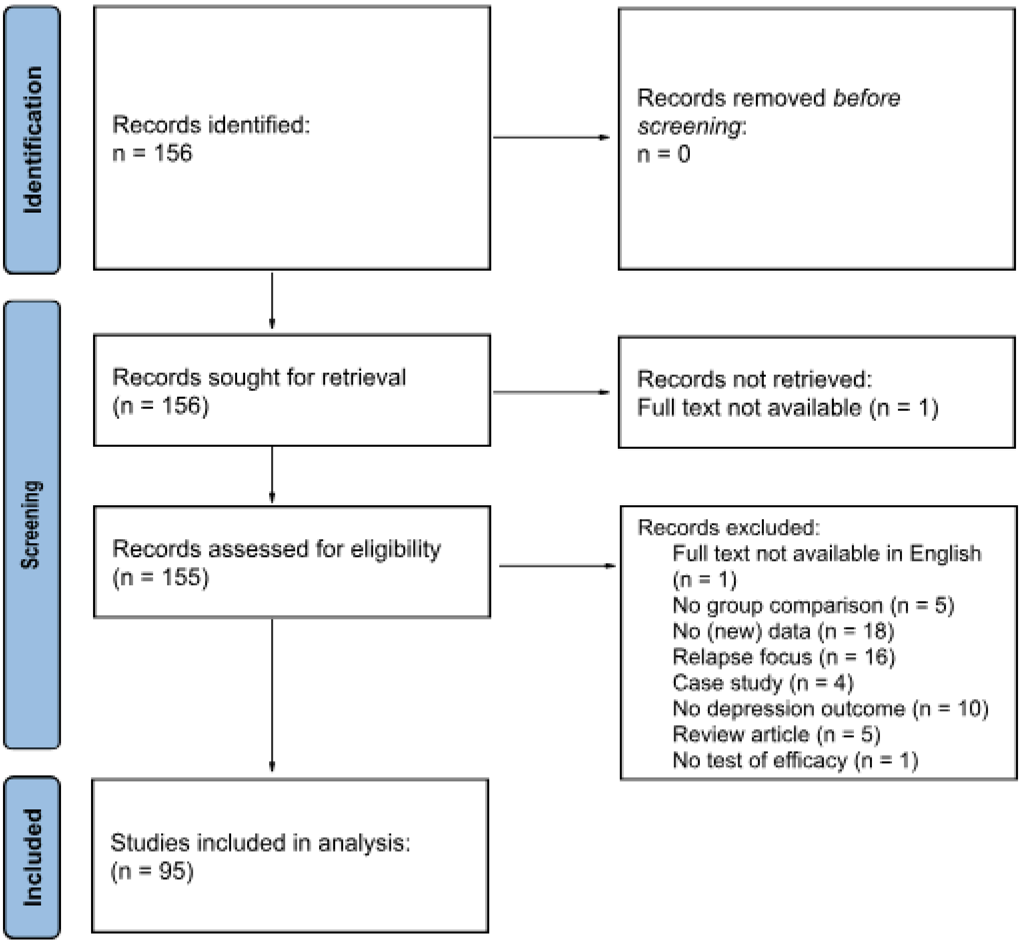
Psychological Therapies

Information on the clinical trials for psychological interventions for depression were extracted from the Division 12 website (https://div12.org/psychological-treatments/), the official website of the SCP. This resource contains a selection of psychological treatment options for 30 psychological disorders or conditions. This selection of treatments is not necessarily fully comprehensive, as some treatments with evidentiary support may not be included. Nevertheless, it provides a good overview of available, research-supported treatments. These therapeutic treatments have been evaluated using either the older (Chambless & Hollon, 1998) or newer (Tolin et al., 2015) set of criteria, and found to be at least modestly efficacious. For each therapy, published evidence of efficacy (i.e., clinical trials) reviewed by the SCP is provided in the form of “Key References” or “Clinical Trials”, from which the relevant data were extracted. Due to the data-driven nature of the project, a pilot extraction was performed at the start of the data extraction period. In this pilot, the first few trials were assessed and the extraction process was adapted for following trials.

At the time of data extraction (May 2025), 17 treatments for depression were listed by the SCP, with a total of 156 references (ranging from 1 to 17 trials per treatment) provided as evidence for efficacy. References were excluded from analysis if they did not collect new empirical data (e.g., overviews or book chapters about treatments, using the same data as another article). Furthermore, severity of depressive symptoms must be assessed as an outcome measure (although not necessarily as the main outcome). Because of the focus on treatment efficacy in terms of relief of depressive symptoms, studies which solely examined relapse rates or the development of symptoms were excluded. Lastly, a group comparison (to a control or active comparator) with a test of difference between the groups must be present. These exclusion criteria were in place to compile the empirical evidence for efficacy (i.e., clinical trials) comparable to the review of pharmacological drugs. In total, 61 references were excluded based on these criteria, leaving 95 trials for 16 different treatments (‘Mom Power’ had no adequate trials) for the analysis. A flow diagram of the screening process with the number of excluded trials per exclusion criteria is presented in figure 1.

Figure 1

Flow Diagram of Screening Process


Identification
Records identified:
n = 156
Records removed *before screening*:
n = 0
Screening
Records sought for retrieval
(n = 156)
Records not retrieved:
Full text not available (n = 1)
Records assessed for eligibility
(n = 155)
Records excluded:
Full text not available in English (n = 1)
No group comparison (n = 5)
No (new) data (n = 18)
Relapse focus (n = 16)
Case study (n = 4)
No depression outcome (n = 10)
Review article (n = 5)
No test of efficacy (n = 1)
Included
Studies included in analysis:
(n = 95)

Trials assessing post-treatment effects as well as trials assessing follow-up effects were included. Trials assessing multiple follow-up timepoints often included a group comparison test over the whole period. If individual tests for multiple follow-up timepoints were reported, the longest follow-up duration was taken. This was done because follow-up measurements assess the long-term effects of the treatment, and the longest follow-up duration provides the best estimate for this long-term effect. For clinical trials comparing a treatment to multiple comparators (e.g., another kind of therapy and a waitlist control group), each comparison was treated as an independent trial in the analysis. Such comparisons in clinical trials are typically conducted using *t*-tests or *F*-tests with degrees of freedom 1,x (which are conceptually equivalent). This *t*- or *F*-value, together with the sample size (total and of each group individually), the *p*-value, and the degrees of freedom, was extracted from the provided clinical trials. Additionally, the kind of comparator (active or control), mean change scores per group (and their SDs), the follow-up period (if present), and the evaluation of the treatment by the SCP was extracted. Furthermore, the outcome measure instrument, analysis method, effect size (and its SE) as well as means and SDs of each group at post-treatment and, if present, all follow-up timepoints were initially extracted but not used for the analysis.

Pharmacological Drugs

Information sources on pharmacological treatments for depression utilized the data provided by the FDA. Data from the clinical trials of antidepressants for depression approved before 2018 were obtained from Pittelkow et al. (2021), who utilized data obtained from previous meta-analyses by Turner et al. (2008) and de Vries et al. (2018). Additionally, data on five novel antidepressants for depression, which have since been approved by the FDA, were extracted from the Drugs@FDA website (https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm) following the approach described in detail by Turner (2013). Specifically, the review files of the Drug Approval Package were sought out. From them, the data were extracted preferably from the statistical review. If a statistical review was not specifically available, data were extracted from the statistical evaluation in the medical or multi-discipline review. Only data from phase II or III trials considered in the FDA review were extracted.

In total, data on 21 antidepressants with a total of 128 trials (ranging from 2 to 18 trials per drug) were collected. Again, trials examining relapse rates as a primary endpoint were excluded. Similar to psychological therapies, multiple drug dosages in fixed-dose trials with multiple drug arms were taken as independent trials. However, flexible-dose trials with one drug arm were taken as one trial. For each trial, the drug dosage, the sample sizes (total and of each group individually), the *p*-value, the mean difference score between the groups (and its SE or CI, whichever was available) as well as the mean change scores per group (and their SDs or SEs, whichever was available) were extracted. Original *t*- or *F*-values were not present in the data obtained from Pittelkow et al. (2021) and were not reported in any FDA review on the five novel antidepressants. Additionally, the subclass of antidepressant (e.g., selective serotonin reuptake inhibitor or N-methyl D-aspartate receptor antagonist) as indicated by the drug label provided by the FDA was obtained.

All relevant study data can be found on OSF (https://osf.io/762hf/?view_only=6fbe0c808972473e83febeb691ce580c). This includes the used for the analysis, the obtained data as well as the complete extracted data for both psychological and pharmacological interventions, and the analysis script.

Statistical Analysis

The analysis will be conducted in RStudio version 4.2.0 (R Core Team, 2022) using the "BayesFactor" package version 0.9.12-4.7 (Morey & Rouder, 2024) to calculate Jeffreys-Zellner-Siow BFs (van Ravenzwaaij & Etz, 2021) for each individual clinical trial. Following previous work (Monden et al., 2016; Monden et al., 2018; Pittelkow et al., 2021), a default Cauchy prior with a location parameter zero and a scale parameter $1/\sqrt{2}$ was used.

For trials with a control group comparison, the prior distribution was truncated below zero to follow the procedure of the FDA, which is to use two-sided tests with a check for directionality. This effectively means that a one-sided test is performed. The truncation, and the calculation of a one-sided BF, follow this reasoning. Therefore, the alternative hypothesis of a positive effect is tested against the null hypothesis of no effect, with negative *t*-values being more consistent with the null hypothesis than with the alternative hypothesis.

For trials with an active comparator, the same default prior was used without truncation. Hence, a two-sided BF is obtained, for which support for the null hypothesis indicates treatment performing similarly to the active comparator and support for the alternative hypothesis indicates treatment performing differently from the active comparator (either better or worse). For reference in interpretation, it was proposed that a BF_{10} between ⅓ and 3 is taken as ambiguous evidence, a $BF_{10} > 3$ as moderate evidence for H_1, a $BF_{10} >10$ as strong evidence, and a $BF_{10} > 30$ as very strong evidence for H_1 (Jeffreys, 1961). A similar interpretation is true for the reverse (i.e., values below 1/3) supporting H_0.

For each trial (both psychotherapies and pharmacotherapies), individual BFs were calculated using the sample sizes of each group and, if available, the reported *t*-statistic. In case an *F*-statistic with degrees of freedom 1,x was reported, the root of the *F*-value was taken to obtain the equivalent *t*-value. If no test statistic was reported, the *t*-values were calculated based on the precise *p*-values. Lastly, if neither a test statistic nor a precise *p*-value was reported, the mean difference between the groups was used to calculate the *t*-value. In the event that only an imprecise *p*-value and no test statistic or mean difference was reported, *t*-values were imputed using multiple imputations. The distribution of possible values was truncated according to the imprecise *p*-value. A BF was calculated for each imputed *t*-value, and the median of these imputed BFs was taken for further analysis.

The effects of therapy trials were split by time point (post-treatment or follow-up) and comparator type (control or active), and the average strength of evidence of SCP evaluations (strong or modest) was compared across time point and comparator type. The effects of drug trials were compared across the types of drugs. All drug trials except for one of the newly collected trials (an active comparator trial for dextromethorphan+bupropion) were post-treatment, placebo-controlled trials. To follow suit with the analysis procedure of therapies, this single trial was analysed separately from the placebo-controlled trials. The differentiation between types of comparators was done because the interpretation of the results is different for actively-controlled trials than for placebo-controlled trials. In contrast to a control group comparison, evidence towards the null hypothesis does not mean that the treatment is not efficacious, but rather that the treatment performs similarly to the active comparator.

To compare the strength of evidence for different kinds of therapies and pharmacological drugs, it was planned to pool the information of the individual studies from each treatment using Bayesian model-averaged meta-analysis. However, trials for therapies frequently use active comparators as well as control groups. The problem in combining these two types of studies is that they ask fundamentally different questions, and thus their results must be interpreted differently. The pooled result from mixing both kinds of studies would be uninterpretable. Using a method to deal with multiple groups (e.g., subgroup analysis) was possible, but would have introduced additional bias, especially considering the limited number of trials in each subgroup. Therefore, only individual trial BFs are calculated (which adequately reflect each trial on the same scale) and were exploratively compared.

Results

The BFs of imputed *t*-values for both pharmacological and psychological interventions can be found in appendix A (tables A1 - A3). Imputed BFs are mostly consistent with each trial, with only a few trials showing considerable variation. Information on the sample sizes, drug dosages for medications, follow-up length, and individual BF of every trial included in the analysis is given in appendix B (tables B1 and B2).

Psychological Interventions

For the psychological therapies, trials were divided into post-treatment and follow-up effects, as well as in active comparator and control comparator trials.

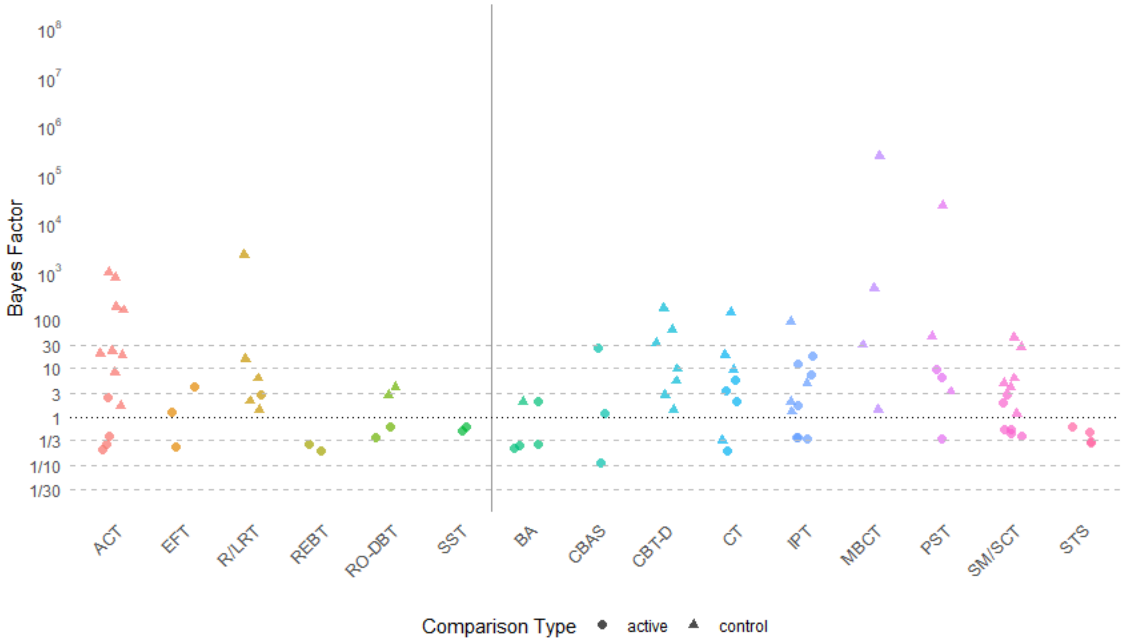
Post-Treatment Effects

Out of the 16 analysed kinds of therapy, 15 included post-treatment data (the exception was short-term psychodynamic therapy). The individual BFs for both kinds of comparator at post-treatment are displayed in figure 2 (see also Table B2). Of these 15 therapies, only 10 provided at least one trial with a control comparator. The median BF_{10} for these therapy trials was 9.72 (Min = 0.32, Max = 225025.5), indicating modest to strong strength of evidence for psychotherapies. Most trials seemed to indicate at least some evidence for efficacy, with only three trials falling below one and no trials indicating evidence towards H0.

13 psychotherapies had at least one trial with a post-treatment effect and an active comparator. As figure 2 shows, BFs in support of a superiority effect were rare and smaller in size. Generally these trials tended to be closer to 1 and the results were much more mixed (see also Table B2). The median BF_{10} was 0.55 (Min = 0.12, Max = 25.94), indicating overall ambiguous evidence. No therapy showed only positive trials, but some (R/LRT, REBT, RO-DBT, SST, STS) had no evidence of superior efficacy compared to an active comparator at post-treatment. Other therapies, like IPT or PST, were more promising. However, as mentioned before, ambiguous evidence or even evidence towards H_0 does not necessarily mean that the treatment is not efficacious when compared to an active comparator.

Figure 2

Distribution of Individual BFs per Therapy at Post-Treatment, divided by Evaluation



Note. Therapies on the left were evaluated as modest, on the right as strong. ACT = Acceptance and Commitment Therapy, EFT = Emotion Focused Therapy, R/LRT = Reminiscence/Life Review Therapy, REBT = Rational Emotive Behavioral Therapy, RO-DBT = Radically Open Dialectical Behavior Therapy, SST = Self-System Therapy, BA = Behavioral Activation, CBAS = Cognitive Behavioral Analysis System, CBT-D = Cognitive Behavioral Therapy for Diabetes, CT = Cognitive Therapy, IPT = Interpersonal Psychotherapy, MBCT = Mindfulness-Based Cognitive Therapy, PST = Problem-Solving Therapy, SM/SCT = Self-Management/Self-Control Therapy, STS = Systematic Treatment Selection.

Follow-up Effects

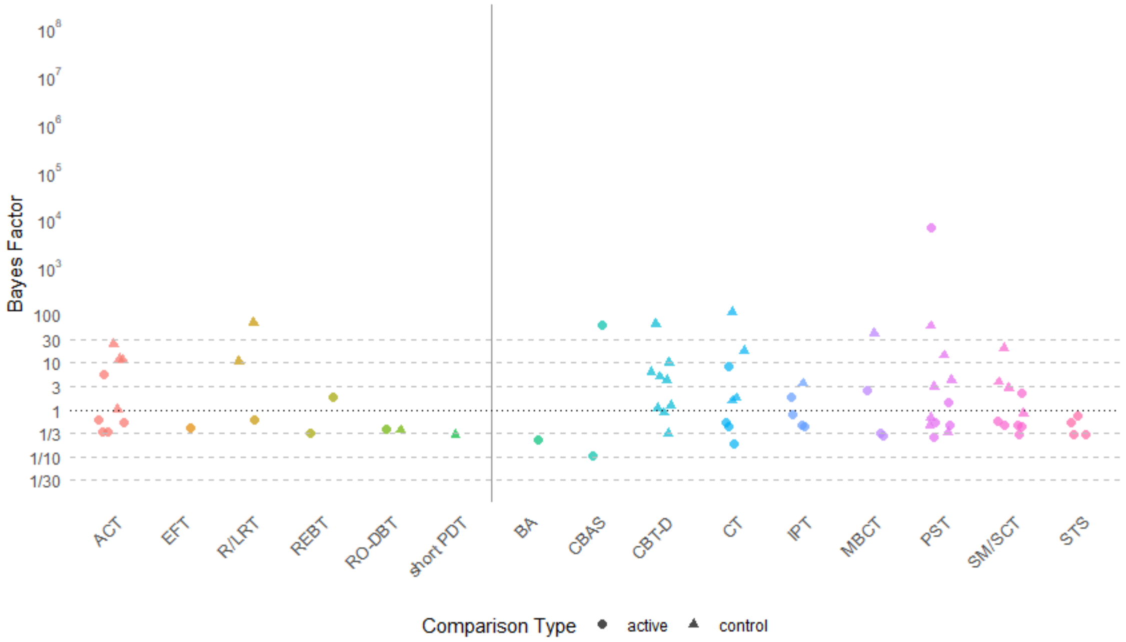
Every kind of therapy had at least one trial reporting follow-up effects, with the exception of self-system therapy. The individual BFs for both kinds of comparator at follow-up are displayed in figure 3 (see also Table B2). Of these 15 therapies, 10 included data on a control comparison. The median BF_{10} was 3.96 (Min = 0.30, Max = 110.51), indicating weak to moderate evidential strength. Again, most trials gave at least some evidence for efficacy, and nearly no trials gave evidence in favor of H_0.

Combining these results with the post-treatment effects against a control comparison, psychological interventions for depression displayed good levels of strength of evidence for efficacy against control groups. However, results should be interpreted cautiously considering the small sample sizes in some groups.

Lastly, 13 therapies included follow-up data with an active comparator. Similarly to the post-treatment effects, results are more ambiguous (figure 3). The median BF_{10} was 0.49 (Min = 0.11, Max = 6539.51). Most trials had a BF_{10} below 1, indicating no or ambiguous evidence for superior efficacy. There is a notable outlier for PST: one of the 5 trials showed extreme strength of evidence, while the other four trials showed ambiguous evidence at best. Generally, the results showed that no therapy consistently outperformed an active comparator in terms of strength of evidence at follow-up.

Figure 3

Distribution of Individual BFs per Therapy at Follow-Up, divided by Evaluation



Note. Therapies on the left were evaluated as modest, on the right as strong. ACT = Acceptance and Commitment Therapy, EFT = Emotion Focused Therapy, R/LRT = Reminiscence/Life Review Therapy, REBT = Rational Emotive Behavioral Therapy, RO-DBT = Radically Open Dialectical Behavior Therapy, short PDT = Short-Term Psychodynamic Therapy, BA = Behavioral Activation, CBAS = Cognitive Behavioral Analysis System, CBT-D = Cognitive Behavioral Therapy for Diabetes, CT = Cognitive Therapy, IPT = Interpersonal Psychotherapy, MBCT = Mindfulness-Based Cognitive Therapy, PST = Problem-Solving Therapy, SM/SCT = Self-Management/Self-Control Therapy, STS = Systematic Treatment Selection.

Some therapies, like CT, CBT-D or PST, stood out, showing good evidential strength both post-treatment and at follow-up, and against an active comparator at post-treatment (as mentioned above, no therapy consistently outperformed an active comparator at follow-up). Other therapies, like STS or BA, generally displayed more ambiguous levels of evidential strength.

RQ1: Comparison of Evidential Strength across SCP Evaluations

Considering the evaluation of the therapies by the SCP, the current recommendations of the SCP were not consistent with the strength of evidence. Therapies labeled as ‘strong’ failed to outperform their ‘modest’ counterparts. Rather, ‘modest’ therapies showed a greater median BF_{10} than ‘strong’ therapies in the post-treatment, control comparator division (17.50 vs 9.20) and in the follow-up, control comparator division (10.40 vs 3.67). When an active comparator was used, both ‘modest’ and ‘strong’ therapies showed similarly ambiguous evidence, with 0.47 and 0.57 respectively at post-treatment, and 0.49 and 0.49 respectively at follow-up. With the exception of acceptance and commitment therapy, ‘strong’ therapies generally seemed to have more trials than ‘modest’ therapies.

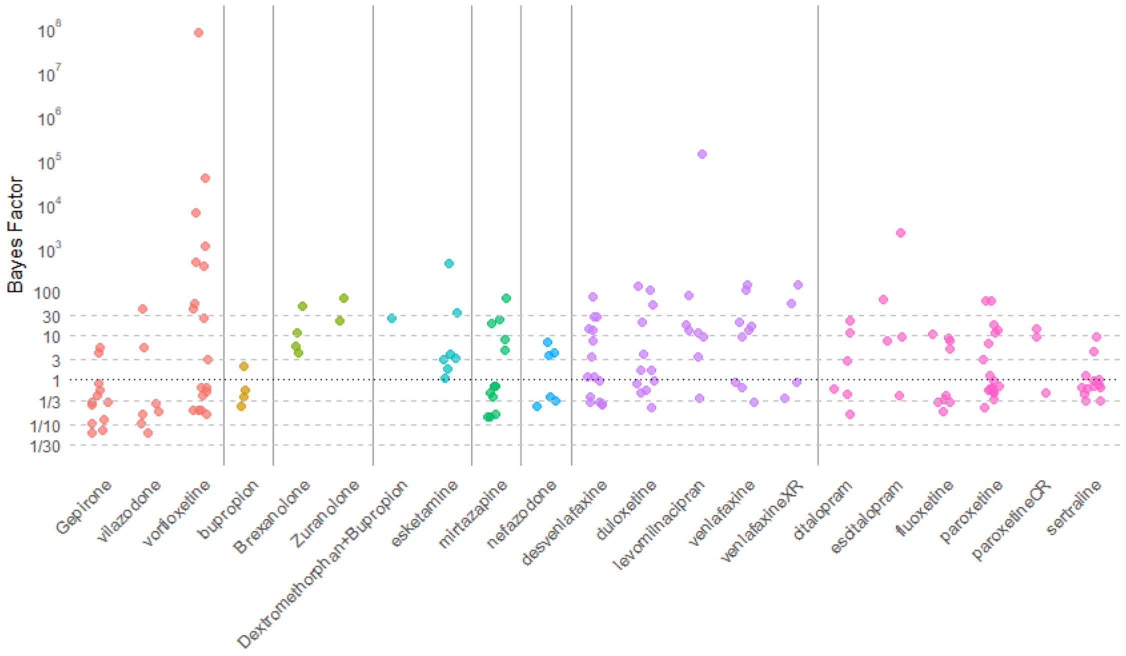
Pharmacological Interventions

As mentioned above, a single drug trial for dextromethorphan+bupropion utilized an active comparator and was therefore considered separately from the other trials. The BF_{10} of this trial was 3.62. The mean difference of the change scores of the treatment and control group was -5.2 in favor of the treatment group. Together, these results indicated moderate evidence towards superiority of the treatment over the active comparator, which in this case was bupropion alone. The rest of the analysis of pharmacological trials refers to placebo-controlled trials.

The individual BFs of the placebo-controlled clinical trials for medications are displayed in figure 4 (see also Table B1). The drugs are ordered and divided by subclass. BFs for most trials fall between either 1/10 and 1, or between 3 and 100, with relatively few trials with a BF_{10} between 1 and 3. The overall median BF_{10} was 1.26 (Min = 0.06, Max = 75060756), indicating ambiguous strength of evidence overall for medications. Most drugs had both a number of ambiguous as well as positive trials, while some drugs (dextromethorphan+bupropion, levomilnacipran, escitalopram, esketamine, brexanolone, and zuranolone) showed mostly, if not exclusively, positive trials. Other drugs, like gepirone or sertraline, showed little strength of evidence in their clinical trials.

Figure 4

Distribution of Individual BFs per Drug, divided by Subclass



Note. The subclasses are (from left to right): 5-HT1A receptor (partial) agonist (5-HT1A), aminoketone antidepressant (AK), GABA-A receptor modulator (GABA), N-methyl D-aspartate receptor antagonist (NMDA), Noradrenergic and specific serotonergic antidepressant (NaSSA), Serotonin antagonists and reuptake inhibitors (SARI), Serotonin-Norepinephrine Reuptake Inhibitors (SNRI), Selective Serotonin Reuptake Inhibitor (SSRI).

RQ2: Comparison of Evidential Strength across Drug Subclasses

The subclasses of antidepressants yielded different evidential strength. Considering the number of trials (*n* = 48) and drugs (*n* = 6) involved, the class of SNRIs generally seemed to provide good strength of evidence (Median BF_{10} = 5.84). The class with the highest BF_{10} was GABA (Median BF_{10} = 17.20). However, seeing that it only includes six trials total, and that zuranolone is only approved for postpartum depression as the trials for major depressive disorder were currently withheld, this value may not reflect the complete picture. The class of 5-HT1A (Median BF_{10} = 0.54) showed some variation, with gepirone and vilazodone displaying low levels of strength of evidence, while vortioxetine showed strong support for efficacy. Similar variation was present in the SSRI class (Median BF_{10} = 0.87), with escitalopram and paroxetineCR performing well, citalopram with ambiguous evidence above 1, and paroxetine, fluoxetine and sertraline with ambiguous evidence below 1. The AK class had both the lowest number of trials (*n* = 4) and the lowest BF_{10} (Median BF_{10} = 0.51). The median BF_{10} of classes NaSSA, SARI, and NMDA were 0.75, 2.00, and 3.44 respectively.

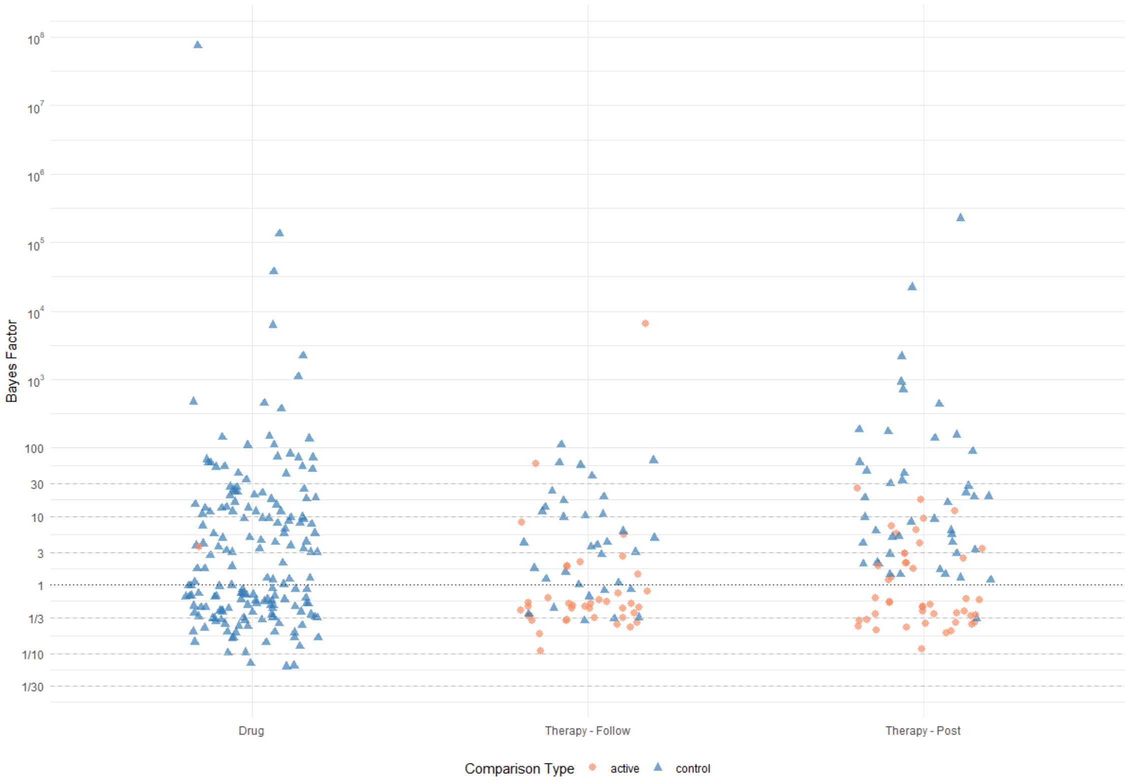
RQ3: Comparison of Pharmacological and Psychological Interventions

For a direct comparison of the BFs of all trials (both pharmacological and psychological) see figure 5. For a more detailed depiction including specific therapies or drugs refer back to the individual sections.

Both kinds of interventions had a similar amount of total trials, with 176 and 163 respectively. The average sample size was higher for drug trials (*M* = 186.72, SD = 103.31) than for therapy trials (*M* = 97.25, SD = 158.05). In terms of trial design, psychological trials comparing post-treatment effects with a control comparator (*n* = 45) are most compatible with the vast majority of drug trials (*n* = 175). Comparing these, psychological therapies presented a substantially greater evidential strength than medications. While the median BF_{10} of medications only indicated ambiguous evidence overall, the median BF_{10} of therapies indicated moderate to strong evidence for an effect. Pharmacological interventions displayed both large and small BFs with a wide spread, producing both the highest and lowest BF_{10}. In contrast, BFs for psychological interventions were less spread out. In other words, while therapies had mostly positive results, medications displayed more ambiguous results next to their positive trials.

Figure 5

Side-by-Side Comparison of All Trials



Discussion

The present study aimed to assess the evidential strength of psychological and pharmacological treatments for depression. To this end, the effects of the clinical trials provided by the SCP for therapies and from the FDA review for drugs were used to calculate a BF for each trial. Specifically, it was explored 1) whether the current recommendations by the SCP for psychological therapies align with the strength of evidence in the clinical studies they provide as references and 2) whether there are differences in strength of evidence between psychological and pharmacological interventions.

Generally, therapies were well supported by the evidence compared to control groups, especially at post-treatment. ACT for example displayed consistently good evidential support against control comparators over a decent number of trials. In contrast, BA had fewer trials and consistently displayed ambiguous evidence. No therapy showed consistent evidence for superiority over other kinds of interventions at either time point, but this is also not needed in order to demonstrate efficacy. As explained above, this is not evidence against treatment efficacy since active comparators are often established treatments. Perhaps trials assessing the comparative effectiveness of treatments should opt for a different design, like a non-inferiority design, which are becoming increasingly popular in medicine but require sophisticated design and larger sample sizes (Rief & Hofmann, 2018; Leon, 2011).

With regard to the first research question, therapies evaluated as ‘strong’ not only failed to outperform those evaluated as ‘modest’, but seemingly showed lower levels of evidential strength when compared to a control group. This was the case at both time points. This means that, based on the references provided by the SCP, the effects of therapies evaluated as ‘modest’ seemed to be more likely to exist than those of therapies evaluated as ‘strong’. In other words, the efficacy of the treatment for ‘modest’ therapies is more strongly supported by the evidence than for ‘strong’ therapies. While these results may be surprising, they are in line with previous findings (Sakaluk et al., 2019). The reason for this pattern was unclear, but its repeated emergence may raise questions about the validity of the evaluation guidelines. However, the guidelines in question have already been replaced with a newer set of guidelines that is more focused on quality instead of quantity of effects, thus potentially taking a step in the right direction.

Regarding the second research question, different subclasses displayed vastly different levels of evidential strength. In the largest subclasses in terms of number of trials included, SNRIs had better moderate evidential strength while SSRIs and 5-HT1As had ambiguous evidence. The other subclasses had generally low numbers of included trials, so interpretation of their results must be more careful. The GABA subclass, consisting of brexanolone and zuranolone, displayed the greatest strength of evidence. However, there was considerable within-class variability for each subclass of antidepressant.

Regarding the third research question, the comparison of psychological and pharmacological treatments, medications had a substantially lower median BF_{10} overall than therapies. While medications showed ambiguous results, the evidential strength of therapies was borderline strong, which would mean that therapies are better supported by the available evidence. This was mostly due to drug trials more commonly having a $BF_{10} < 0$ (ambiguous or even pro-null evidence) against placebo-controls. Therefore, the typical evidential strength was substantially larger for therapies compared to antidepressant medications. Drug trials also showed greater variability than therapy trials (post-treatment, control comparator), reporting both the smallest and largest BF_{10}. A notable proportion of drugs produced ambiguous evidence despite meeting FDA approval for marketing, with gepirone and vilazodone even producing pro-null evidence according to the rule of thumb by Jeffreys (1961).

The finding is further exaggerated by the fact that the average sample size in therapy trials was considerably smaller than in drug trials. From a statistical standpoint, with all other things being equal, the strength of evidence should increase with increasing sample size. Since therapy trials have both smaller sample sizes and larger average strength of evidence, the results seemed to suggest higher effect sizes for therapy trials to compensate. However, a recent meta-analytic review did not find effect sizes to be notably larger for psychotherapies compared to pharmacotherapies (Leichsenring et al., 2022). The results therefore raise questions about how they came to be and how valid they are.

The most plausible reason behind the difference in magnitudes for psychological versus pharmacological treatments lies in the nature of the respective guidelines. New medications must seek direct approval from the FDA, and companies are required to pre-register trials at the national library of medicine (NLM). Failure to disclose all relevant trials to the FDA can lead to regulatory consequences. In contrast, the SCP does not directly approve or endorse treatments, and while the NLM also contains psychotherapy trials there is no requirement to pre-register a trial. Consequently, the seeming superiority of therapies in terms of evidential strength may stem from reference selectivity on the part of the SCP as well as publication bias. Such selective publication based on the study outcome was previously observed for pharmacological trials (Turner et al., 2008) and psychological depression treatments (Cuijpers et al., 2010; Driessen et al., 2015). This effect was found to be stronger in meta-analyses in psychology than in medicine (Bartoš et al., 2024). Perhaps the picture would be different if therapies had to directly appeal to the APA or SCP for approval.

The FDA criteria might also be directly reflected in the distribution of trial BFs (figure 4). While the strength of evidence does not directly reflect a statistically significant result, the two measures are connected. Consequently, and in line with the endorsement criteria of the FDA requiring two statistically significant trials, some drugs (dextromethorphan+bupropion, venlafaxineXR, paroxetineCR, vilazodone, sertraline, zuranolone) displayed only 2 trials with supporting strength of evidence. In the case of dextromethorphan+bupropion (one actively- and one placebo-controlled trial) and zuranolone, the two positive trials were also the only trials, painting an overall positive picture. However, cases like gepirone, vilazodone or sertraline had more ambiguous evidence overall, and did not seem strongly supported by the evidence considering all trials. This pattern may be reflective of the goal of sponsors to market the drug, pursuing approval after failed trials. Such a pattern did not show for the therapy trials.

Overall, the study findings highlight the need for rigorous evidential standards, including quality control and pre-registration of trials. If the goal is to make psychological and pharmacological interventions comparable in terms of evidential standard, they must also underlie the same criteria. That is not to say that the FDA criteria are without flaw. The shortcomings of the NHST framework as the sole instrument to determine the evidence have already been discussed, and the focus on quantity (i.e., two statistically significant results despite other failed trials) is suboptimal (for a critique on the FDA criteria and improvement suggestions see Spielmans & Kirsch, 2014). However, the current comparison leads to most likely incorrect interpretations of superior evidential strength of psychotherapies due to the impact of selectivity, publication bias, and different evidence standards.

Limitations, Considerations and Future Research

While the results and implications of this study are meaningful, there are a few things to consider. First, the evaluation of the SCP for therapies was most likely based on multiple factors, and drawing conclusions about their correctness solely on the basis of their evidential strength would be naive. For example, a number of references by the SCP were case studies or reviews, which may add to the evidence for efficacy of the treatment but were not taken into account in this study due to its quantitative analysis. Furthermore, qualitative information about each trial (e.g., risk of bias or attrition rates) might have contributed to the evaluations. The FDA also rejects trials in case of lacking quality or questionable decisions during the study process, and while there was no certainty that this kind of quality control was present in the SCP evaluations, it seemed likely that such information was incorporated into the evaluation.

Furthermore, differences in the strength of evidence between psychological and pharmacological interventions may also partly stem from the exclusion criteria that were applied to therapy trials. The criteria were chosen to make therapy trials as comparable to the drug trials as possible, but since the majority of data for drug trials were obtained from an external source it cannot be ruled out that different decisions were made in the extraction and screening process.

Another consideration was the general state of therapy trials. There was massive variability in the designs, analysis methods, specific comparators, and populations in therapy trials. While this heterogeneity may improve generalizability somewhat, the differences between studies can complicate the integration and comparison of results. Additionally, a number of references provided by the SCP were inadequate to serve as evidence for a treatments' efficacy. For example, there were articles which did not provide empirical data, clinical trials that did not assess depression in any way as an outcome, and on one occasion just the protocol for a future trial. FDA reviews, at least the newer ones from which data was extracted for this study, seemed to be more coherent in their presentation of the evidence. However, in the case of Zurzuvae information of general use for depression was withheld, and reported effects were only applicable to post-partum depression.

Future research can expand the investigation of evidential strength to other mental disorders. Additionally, selectivity effects and publication bias for therapy trials can be further investigated to gain a better understanding of the validity of these results. While therapies seemed to have superior evidential support, the questions with regard to the validity of this finding give rise to further study possibilities. Finally, the development of a more holistic metric framework for the evaluation of the evidence may improve the evidence-base of clinical treatments, as it has been repeatedly shown that strength of evidence is currently neglected despite the “fundamental commitment” to evidence-based practice.

Conclusion

This study emphasized the importance of assessing the evidential strength when evaluating the evidence for treatment efficacy. Comparing the strength of evidence between psychological and pharmacological interventions, therapies seemed better supported by the evidence. However, questions remained as to the validity of these results, or whether they were biased by selective publication of trials. Moreover, previous findings with regard to inconsistencies between the SCP evaluations and their evidential strength were strengthened, and questions about the criteria requiring two significant trials were raised. Ultimately, a measure of strength of evidence such as the BF may aid clinical decision making by providing additional information about the evidence for, and the efficacy of mental health treatments.

Acknowledgement

I acknowledge the use of AI (ChatGPT, https://chatgpt.com/) in code writing and debugging as well as using it for feedback and suggestions for improvement in text writing.

References

Ahmed, E. S., & Butt, M. N. (2025). The misunderstood P-value: Why statistical significance is not enough in clinical practice. *British Journal of Anaesthesia, 134*(4), 909–913. https://doi.org/10.1016/j.bja.2025.01.008

American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*(4), 271–285.

Bartoš, F., Maier, M., Wagenmakers, E., Nippold, F., Doucouliagos, H., Ioannidis, J. P. A., Otte, W. M., Sladekova, M., Deressa, T. K., Bruns, S. B., Fanelli, D., & Stanley, T. D. (2024). Footprint of publication selection bias on meta‐analyses in medicine, environmental sciences, psychology, and economics. *Research Synthesis Methods, 15*(3), 500–511. https://doi.org/10.1002/jrsm.1703

Beard, E., Dienes, Z., Muirhead, C., & West, R. (2016). Using Bayes factors for testing hypotheses about intervention effectiveness in addictions research. *Addiction, 111*(12), 2230–2247. https://doi.org/10.1111/add.13501

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7–18.

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *British Journal of Psychiatry, 196*(3), 173–178. https://doi.org/10.1192/bjp.bp.109.066001

de Vries, Y. A., Roest, A. M., De Jonge, P., Cuijpers, P., Munafò, M. R., & Bastiaansen, J. A. (2018). The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: The case of depression. *Psychological Medicine, 48*(15), 2453–2455. https://doi.org/10.1017/S0033291718001873

Driessen, E., Hollon, S. D., Bockting, C. L., Cuijpers, P., & Turner, E. H. (2015). Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health-funded trials. *PLOS ONE, 10*(9), e0137864. https://doi.org/10.1371/journal.pone.0137864

Food and Drug Administration. (1998). *Guidance for industry: Providing clinical evidence of effectiveness for human drug and biological products*. U.S. Department of Health and Human Services.

Food and Drug Administration. (2022, August 8). *Development & approval process | Drugs*. https://www.fda.gov/drugs/development-approval-process-drugs

Gaudiano, B. A., & Miller, I. W. (2013). The evidence-based practice of psychotherapy: Facing the challenges that lie ahead. *Clinical Psychology Review, 33*(7), 813–824. https://doi.org/10.1016/j.cpr.2013.04.004

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine, 130*(12), 995–1004.

Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health, 78*(12), 1568–1574. https://doi.org/10.2105/ajph.78.12.1568

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*(1), 69–88. https://doi.org/10.1177/0959354307086923

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Justman, S. (2010). From medicine to psychotherapy: The placebo effect. *History of the Human Sciences, 24*(1), 95–107. https://doi.org/10.1177/0952695110386655

Karyotaki, E., Smit, Y., Holdt Henningsen, K., Huibers, M. J., Robays, J., de Beurs, D., & Cuijpers, P. (2016). Combining pharmacotherapy and psychotherapy or monotherapy for major depression? A meta-analysis on the long-term effects. *Journal of Affective Disorders, 194*, 144–152. https://doi.org/10.1016/j.jad.2016.01.036

Lakens, D. (2022). Why P values are not measures of evidence. *Trends in Ecology & Evolution, 37*(4), 289–290. https://doi.org/10.1016/j.tree.2021.12.006

Leichsenring, F., Steinert, C., Ioannidis, J. P. A., & Jones, R. M. (2022). The efficacy of psychodynamic psychotherapy, cognitive-behavioral therapy, and antidepressant medication in the treatment of major depressive disorder: A systematic overview and network meta-analysis. *World Psychiatry, 21*(1), 133–145.

Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. P. A. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: An umbrella review and meta-analytic evaluation of recent meta-analyses. *World Psychiatry, 21*(1), 133–145. https://doi.org/10.1002/wps.20941

Leon, A. C. (2011). Evolution of psychopharmacology trial design and analysis: Six decades in the making. *The Journal of Clinical Psychiatry, 72*(3), 331–340. https://doi.org/10.4088/JCP.10r06669

Monden, R., de Vos, C. M., Noorthoorn, E. O., Romeijn, J.-W., & van Ravenzwaaij, D. (2016). The strength of evidence for the efficacy of antidepressants in the treatment of anxiety disorders. *PLOS ONE, 11*(8), e0160855.

Monden, R., de Vos, C. M., Noorthoorn, E. O., Romeijn, J.-W., & van Ravenzwaaij, D. (2018). The strength of evidence of antidepressants for depression. *Journal of Psychopharmacology, 32*(10), 1074–1084.

Moreno-Agostino, D., Wu, Y. T., Daskalopoulou, C., Hasan, M. T., Huisman, M., & Prina, M. (2021). Global trends in the prevalence and incidence of depression: A systematic review and meta-analysis. *Journal of Affective Disorders, 281*, 235–243. https://doi.org/10.1016/j.jad.2020.12.035

Morey, R., & Rouder, J. (2024). *BayesFactor: Computation of Bayes factors for common designs* (Version 0.9.12-4.7) [R package]. https://CRAN.R-project.org/package=BayesFactor

Pittelkow, M. M., de Vries, Y. A., Monden, R., Bastiaansen, J. A., & van Ravenzwaaij, D. (2021). Comparing the evidential strength for psychotropic drugs: A Bayesian meta-analysis. *Psychological Medicine, 51*(16), 2752–2761. https://doi.org/10.1017/S0033291721003950

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rief, W., & Hofmann, S. G. (2018). Some problems with non-inferiority tests in psychotherapy research: Psychodynamic therapies as an example. *Psychological Medicine, 48*(8), 1392–1394. https://doi.org/10.1017/S0033291718000247

Sakaluk, J. K., Williams, A. J., & Bierman, A. (2019). Evaluating evidence for empirically supported psychological treatments (ESTs): A meta-scientific review. *Journal of Abnormal Psychology, 128*(6), 500–512.

Silverman, A. L., Werntz, A., Ko, T. M., & Teachman, B. A. (2021). Implicit and explicit beliefs about the effectiveness of psychotherapy vs. medication: A large-scale examination and replication. *The Journal of Nervous and Mental Disease, 209*(11), 783–795. https://doi.org/10.1097/NMD.0000000000001384

Spielmans, G. I., & Kirsch, I. (2014). Drug approval and drug effectiveness. *Annual Review of Clinical Psychology, 10*, 741–766. https://doi.org/10.1146/annurev-clinpsy-050212-185533

Ten Have, M., Tuithof, M., van Dorsselaer, S., Schouten, F., Luik, A. I., & de Graaf, R. (2023). Prevalence and trends of common mental disorders from 2007–2009 to 2019–2022: Results from the Netherlands Mental Health Survey and Incidence Studies (NEMESIS), including comparison of prevalence rates before vs. during the COVID-19 pandemic. *World Psychiatry, 22*(2), 275–285. https://doi.org/10.1002/wps.21087

Thom, J., Jonas, B., Reitzle, L., Mauz, E., Hölling, H., & Schulz, M. (2024). Trends in the diagnostic prevalence of mental disorders, 2012–2022—Using nationwide outpatient claims data for mental health surveillance. *Deutsches Ärzteblatt International, 121*(11), 355–362. https://doi.org/10.3238/arztebl.m2024.0052

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: Recommendations for a new model. *Clinical Psychology: Science and Practice, 22*(4), 317–338.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*(3), 252–260. https://doi.org/10.1056/NEJMsa065779

Turner, E. H. (2013). How to access and process FDA drug approval packages for use in research. *BMJ, 347*, f5992. https://doi.org/10.1136/bmj.f5992

van Ravenzwaaij, D., & Etz, A. (2021). Simulation studies as a tool to understand Bayes factors. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920972624 van Ravenzwaaij, D., & Ioannidis, J. P. A. (2017). A simulation study of the strength of evidence in the recommendation of treatments based on two significant trials. *PLOS ONE, 12*(3), e0173184.

van Ravenzwaaij, D., & Ioannidis, J. P. A. (2019). True and false positive rates for different criteria of evaluating statistical evidence from clinical trials. *BMC Medical Research Methodology, 19*, 218. https://doi.org/10.1186/s12874-019-0865-y

Wagenmakers, E. (2007). A practical solution to the pervasive problems of P values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105

Williams, A. J., Botanov, Y., Giovanetti, A. K., Perko, V. L., Sutherland, C. L., Youngren, W., & Sakaluk, J. K. (2023). A metascientific review of the evidential value of acceptance and commitment therapy for depression. *Behavior Therapy, 54*(6), 989–1005. https://www.sciencedirect.com/science/article/pii/S0005789422000764

World Health Organization. (2017, March 30). “Depression: Let’s talk” says WHO, as depression tops list of causes of ill health. https://www.who.int/news/item/30-03-2017--depression-let-s-talk-says-who-as-depression-tops-list-of-causes-of-ill-health

Appendix A

BFs of Imputed *t*-values

Table A1

BFs of Imputed t-Values for Pharmacological Medications (in Ascending Order)

Drug	BF1	BF2	BF3	BF4	BF5	BF6	BF7	BF8	BF9
Cymbalta	65.66	75.69	102.65	121.69	135.78	144.17	154.72	163.59	201.68
Cymbalta	50.43	53.78	59.89	64.47	109.31	111.91	163.73	165.26	190.63
Lexapro	395.96	1487.60	1657.57	1747.95	2185.39	2436.94	6306.50	11789.48	91552.82
Paxil	0.19	0.34	0.41	0.46	0.59	0.76	1.56	1.77	2.17
Paxil	0.33	0.36	0.67	0.86	1.26	1.52	1.64	1.95	2.00
Paxil	0.19	0.34	0.43	0.49	0.70	0.74	1.00	1.33	1.66
Paxil	0.11	0.17	0.18	0.20	0.36	0.41	0.75	0.84	1.98
Paxil	0.14	0.19	0.19	0.49	0.53	0.56	0.90	1.01	1.31
Paxil	0.22	0.28	0.34	0.39	0.59	1.49	1.70	1.75	2.18
Paxil	0.20	0.26	0.34	0.36	0.50	0.51	0.98	1.00	1.89
Zoloft	0.23	0.23	0.28	0.36	0.75	0.80	0.87	1.15	1.96
Zoloft	0.15	0.15	0.33	0.48	0.77	0.91	1.47	1.54	1.79
Zoloft	0.31	0.36	0.47	0.55	0.69	0.73	1.06	1.07	2.34
Zoloft	0.22	0.25	0.47	0.59	0.98	1.21	1.30	1.45	1.87
Effexor	55.41	56.79	74.99	100.30	111.38	116.82	254.87	272.73	317.82
Effexor	56.47	87.11	100.92	111.24	148.79	154.05	158.52	181.36	187.60
EffexorXR	58.84	61.53	126.37	136.32	143.90	162.91	190.25	209.89	210.28

Table A2

BFs of Imputed t-Values for Psychological Therapies (Post-Treatment; in Ascending Order)

Therapy	Type	BF1	BF2	BF3	BF4	BF5	BF6	BF7	BF8	BF9
RO-DBT	active	0.20	0.23	0.24	0.31	0.72	0.93	1.67	2.53	2.69
CBT-D	control	0.05	60.90	63.52	71.48	73.96	75.17	126.38	148.51	172.53
CBT-D	control	0.09	2.96	3.33	3.99	4.70	4.89	6.00	7.21	8.87
STS	active	0.26	0.27	0.27	0.31	0.31	0.31	0.33	0.34	0.38
STS	active	0.25	0.27	0.30	0.31	0.31	0.32	0.36	0.37	0.38
STS	active	0.16	0.18	0.18	0.19	0.21	0.30	0.40	0.68	1.50
STS	active	0.35	0.44	0.51	0.55	0.78	1.13	1.72	2.41	2.46
CBAS	active	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
REBT	active	0.08	0.08	0.10	0.10	0.15	0.20	0.21	0.28	0.41
REBT	active	0.12	0.12	0.16	0.28	0.65	0.69	0.92	1.48	1.73
R/LRT	active	0.09	0.09	0.09	0.10	0.10	0.10	0.10	0.11	0.11
R/LRT	control	0.06	494.91	556.01	594.24	979.82	1437.39	2734.75	49260.38	54019.83
R/LRT	control	0.10	2.81	3.78	4.69	6.29	6.47	6.80	7.10	8.04
R/LRT	control	10.16	10.67	11.38	17.02	17.87	17.96	23.74	26.50	45.66
R/LRT	control	0.12	1.75	2.02	2.19	2.27	2.32	2.46	2.46	2.51
SM/SCT	active	0.08	3.24	3.52	3.63	4.60	4.72	5.72	8.01	8.04
SM/SCT	active	0.14	0.24	0.30	0.35	0.98	1.06	1.15	1.27	2.44
SM/SCT	active	0.16	0.17	0.18	0.18	0.21	0.42	0.56	1.63	2.55
SM/SCT	control	0.17	0.17	4.62	4.64	4.94	5.90	5.94	6.99	7.11
SM/SCT	control	0.14	10.36	11.86	14.32	16.85	26.67	38.74	49.80	56.77

IPT	active	0.13	0.15	0.17	0.23	0.30	0.35	2.81	3.06	8.36
IPT	active	0.92	1.24	1.45	1.75	1.76	2.35	4.35	5.15	6.52
IPT	active	0.09	0.09	0.09	9.96	10.99	13.99	14.63	17.93	57.55
IPT	active	9.72	9.90	10.89	11.36	14.84	15.50	21.57	38.22	48.22
IPT	control	50.35	64.86	82.96	89.75	148.11	154.67	154.91	163.77	328.09
IPT	control	0.31	0.35	0.37	0.56	0.73	0.80	1.23	1.27	1.36
PST	active	0.09	0.10	0.12	0.13	0.27	0.30	0.53	0.55	0.65
PST	control	2.57	2.70	2.82	3.56	4.83	5.04	6.61	7.05	7.36
CT	control	74.33	94.51	99.58	198.71	224.98	289.12	382.55	440.53	441.37
ACT	active	0.05	0.06	2.09	2.96	3.10	3.67	4.40	5.21	7.52
ACT	active	0.13	0.13	0.15	0.15	0.28	0.67	0.79	2.19	2.22
ACT	control	76.32	79.58	88.21	97.31	120.29	123.00	161.06	195.36	243.96
ACT	control	0.27	0.50	0.54	0.79	0.90	1.58	1.92	2.16	2.18
ACT	control	10.77	11.72	16.09	17.02	17.45	17.81	18.14	20.28	36.51
ACT	control	10.04	13.12	14.50	14.70	24.23	24.88	29.27	37.29	53.94
ACT	control	65.08	65.60	68.53	212.26	268.09	283.89	353.02	413.98	436.25

Table A3

BFs of Imputed t-Values for Psychological Therapies (Follow-up; in Ascending Order)

Therapy	Type	BF1	BF2	BF3	BF4	BF5	NF6	BF7	BF8	BF9
STS	active	0.27	0.27	0.28	0.29	0.32	0.35	0.35	0.35	0.35
STS	active	0.25	0.26	0.27	0.29	0.29	0.30	0.32	0.33	0.34
STS	active	0.20	0.27	0.28	0.30	0.31	0.34	0.36	0.42	0.94
STS	active	0.40	0.41	0.51	0.75	1.14	1.56	1.71	2.46	2.61
CBAS	active	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
REBT	active	2.38	2.70	3.01	3.21	3.89	4.09	4.58	4.76	6.88
REBT	active	0.17	0.31	0.34	0.39	0.68	0.74	0.95	1.08	2.24
R/LRT	control	48.95	59.74	73.66	107.83	138.40	148.52	157.57	202.57	242.86
R/LRT	control	4.12	4.25	4.43	5.25	5.84	7.25	7.86	11.38	15.86
R/LRT	active	0.14	0.16	0.16	0.17	0.17	0.18	0.18	0.21	0.24
SM/SCT	active	0.10	0.10	0.10	0.11	0.14	0.16	0.16	0.19	0.34
SM/SCT	active	0.23	0.35	0.38	0.43	0.60	1.25	1.46	1.95	2.48
SM/SCT	active	0.19	0.29	0.36	0.52	0.53	0.63	1.37	1.62	2.80
SM/SCT	control	2.69	3.25	4.23	4.90	4.94	5.02	5.17	6.26	7.79
SM/SCT	control	0.18	0.26	0.60	0.71	0.72	1.38	1.73	1.94	2.34
SM/SCT	control	10.42	14.13	14.19	14.72	16.19	16.42	18.29	23.7	26.21
IPT	active	0.18	0.28	0.45	0.76	1.39	1.87	2.37	2.47	2.65
IPT	active	0.20	0.42	0.43	0.60	1.04	1.43	1.66	1.90	2.72
IPT	active	0.22	0.38	0.39	0.41	0.48	0.51	0.68	1.61	2.72
PST	control	27.80	29.20	40.67	52.91	54.27	64.59	75.54	120.73	128.12
PST	control	2.60	2.65	3.31	3.35	3.90	5.07	60	8.04	8.09

PST	active	0.21	0.29	0.32	0.44	0.50	1.26	1.67	2.20	2.30
PST	active	0.17	0.21	0.21	0.23	0.37	0.43	1.00	1.71	2.23
CT	active	0.14	0.16	0.16	0.18	0.19	0.25	0.35	0.62	2.05
CT	active	0.41	0.43	0.62	0.72	0.82	0.85	0.86	1.62	2.21
CT	control	10.71	11.22	13.20	13.66	14.93	16.78	25.67	44.88	49.94
CT	control	48.24	56.31	70.66	82.88	86.05	148.20	151.11	202.60	279.53
EFT	active	0.24	0.39	0.45	0.98	1.07	1.23	1.77	2.68	2.90
ACT	active	0.21	0.41	0.52	0.86	1.03	1.35	1.85	2.42	2.82
ACT	active	0.20	0.39	0.46	0.69	1.09	2.03	2.13	2.21	2.75
ACT	active	0.33	0.36	0.43	0.84	0.95	1.31	1.93	2.16	2.19
ACT	active	0.10	0.11	0.25	0.25	0.26	0.34	0.39	0.78	0.89

Appendix B

Information on all Individual Trials

Table B1

Drug Dosage, Sample Sizes and BF of each Drug Trial

Drug (active agent)	Trial	Dose (mg)	N	n (treatment	n (control)	BF
Trintellix	315	20	300	147	153	2.99
(vortioxetine)	316	20	303	148	155	24.88
	13267A	15	307	149	158	37592.29
	13267A	20	309	151	158	75060755.86
	11492A	5	213	108	105	1103.50
	11492A	10	205	100	105	472.92
	305	1	278	139	139	53.43
	305	5	278	139	139	370.11
	305	10	278	139	139	6206.22
	12541	5	300	155	145	42.48
	11984A	2.5	300	155	145	0.46
	11984A	5	300	155	145	0.70
	11984A	10	296	151	145	0.54
	317	10	292	143	149	0.21
	317	15	291	142	149	0.17
	303	5	578	292	286	0.21
	304	2.5	295	146	149	0.68
	304	5	302	153	149	0.21
Viibryd	244	20 - 100	181	86	95	0.10
(vilazodone)	245	40 - 60	196	97	99	0.20
	245	80 - 100	192	93	99	0.07
	246	20	252	123	129	0.30
	248	20	260	132	128	0.17

	7	40	463	232	231	5.66
	4	40	397	198	199	42.24
Effexor	600A-203	75	169	77	92	16.15
(venlafaxine)	600A-203	150 - 225	171	79	92	111.38
	600A-203	300 - 375	167	75	92	20.76
	600A-206	150 - 375	93	46	47	13.11
	600A-301	75 - 225	142	64	78	148.79
	600A-302	75 - 200	140	65	75	9.40
	600A-303	75 - 225	148	69	79	0.33
	600A-313	75	147	72	75	0.69
	600A-313	200	152	77	75	0.87
EffexorXR	208	75 - 150	176	85	91	53.30
(venlafaxine)	209	75 - 225	191	91	100	143.90
	367	75	163	82	81	0.40
	367	150	156	75	81	0.87
Zoloft	104	50 - 200	283	142	141	9.70
(sertraline)	103	50	176	90	86	4.45
	103	100	175	89	86	1.26
	103	200	168	82	86	0.61
	315	50 - 200	148	75	73	0.35
	101	50	45	22	23	0.68
	101	100	42	19	23	0.34
	101	200	40	17	23	1.02
	101	400	35	12	23	0.48
	310	50	61	31	30	0.75
	310	100	58	28	30	0.77
	310	200	57	27	30	0.69
	310	400	60	30	30	0.98
Paxil	02-001	10-50	104	51	53	18.01
(paroxetine)	02-002	10-50	70	36	34	6.57

	02-004	10-50	66	34	32	61.09
	03-001	10-50	76	39	37	13.58
	03-004	10-50	74	37	37	2.99
	03-005	10-50	82	40	42	11.81
	03-006	10-50	76	39	37	60.54
	03-002	10-50	80	40	40	0.71
	03-003	10-50	81	39	42	0.24
	02-003	10-50	66	33	33	0.65
	01-001	10-50	48	24	24	0.99
	7	20	25	13	12	0.59
	9	20	155	104	51	1.26
	9	30	150	99	51	0.70
	9	40	151	100	51	0.36
	UK-06	30	45	22	23	0.53
	UK-09	30	41	20	21	0.59
	UK-12	30	29	19	10	0.50
PaxilCR	487	12.5 - 50	210	103	107	9.36
(paroxetine)	449	20 - 62.5	218	108	110	15.00
	448	20 - 62.5	187	94	93	0.51
Serzone	03AOA-003	100 - 500	89	44	45	3.58
(nefazodone)	03AOA-004B	300 - 600	153	78	75	4.26
	CN104-005	100 - 600	177	86	91	7.28
	CN104-006	100 - 600	158	80	78	0.42
	030A2-007	300	88	41	47	0.35
	03AOA-004A	300 - 600	153	76	77	0.25
Remeron	003-020/3220	5 - 35	80	41	39	18.82
(mirtazapine)	003-002	5 - 35	88	44	44	72.31
	003-022/3220	10 - 35	99	49	50	23.14
	003-023/3220	5 - 35	98	49	49	4.83
	003-024/3220	5 - 35	98	50	48	8.51

	85027	20 -60	125	64	61	0.75
	84023	15 - 50	90	45	45	0.53
	003-021/3220	10 - 35	93	45	48	0.75
	003-003	10 - 35	90	45	45	0.40
	003-008	15	58	30	28	0.15
	003-008	30	56	28	28	0.14
	003-008	60	58	30	28	0.17
Fetzima	F02695 LP2 02	75 - 100	544	267	277	134983.15
(levomilnacipran)	LVM-MD-01	40	351	176	175	3.37
	LVM-MD-01	80	352	177	175	13.40
	LVM-MD-01	120	351	176	175	82.16
	LVM-MD-10	40	370	185	185	17.81
	LVM-MD-10	80	372	187	185	11.78
	LVM-MD-03	40 - 120	429	215	214	9.61
	LVM-MD-02	40 - 120	355	174	181	0.39
Prozac	19	40 - 80	46	22	24	8.94
(fluoxetine)	27	40 - 80	344	181	163	4.96
	62-a	20	159	103	56	0.37
	62-a	40	155	99	56	0.33
	62-a	60	163	107	56	0.32
	62-b	20	145	97	48	10.70
	62-b	40	145	97	48	7.94
	62-b	60	151	103	48	0.46
	25	40 - 80	42	18	24	0.20
Spravato	TRD3001	56	228	115	113	2.99
(esketamine)	TRD3001	84	227	114	113	1.10
	TRD3002	56 or 84	223	114	109	3.76
	TRD3005	28 or 56 or 84	137	72	65	1.85
	TRD2003	28	58	19	39	3.12
	TRD2003	56	59	20	39	34.44

	TRD2003	84	56	17	39	446.31
Lexapro	99001	10	377	188	189	7.64
(escitalopram)	99003	10 - 20	309	155	154	9.40
	SCT-MD-01	10	237	118	119	67.97
	SCT-MD-01	20	242	123	119	2185.39
	SCT-MD-02	10 - 20	249	124	125	0.45
Cymbalta	HMAT-B	40	175	86	89	3.77
(duloxetine)	HMAT-B	80	180	91	89	20.32
	HMAY-A	80	188	95	93	52.48
	HMAY-A	120	186	93	93	135.78
	HMBH-A	60	245	123	122	109.31
	HMBH-B	60	267	128	139	1.72
	HMAQ-A	20 - 60	113	56	57	0.95
	HMAY-B	80	192	93	99	0.50
	HMAY-B	120	202	103	99	1.71
	HMAQ-B	20 - 60	153	81	72	0.25
	HMAT-A	40	179	90	89	0.57
	HMAT-A	80	170	81	89	0.86
Pristiq	332	50	300	150	150	3.37
(desvenlafaxine)	332	100	297	147	150	0.96
	223	200	141	63	78	0.29
	223	400	150	72	78	0.31
	306	100	232	114	118	14.72
	306	200	234	116	118	1.22
	306	400	231	113	118	27.02
	308	200	245	121	124	26.56
	308	400	248	124	124	7.91
	304	100 - 200	234	120	114	0.43
	309	200 - 400	237	117	120	0.33
	317	200 - 400	235	110	125	0.27

	320	200 - 400	235	117	118	1.19
	333	50	325	164	161	13.18
	333	100	319	158	161	74.34
Celexa	85A	20 - 80	160	78	82	2.69
(citalopram)	91206	40	244	120	124	21.86
	91206	60	234	110	124	11.51
	86141	10 - 30	147	97	50	0.49
	89303	40	125	61	64	0.66
	89306	40	185	97	88	0.17
WellbutrinSR	203	300	230	113	117	2.08
(bupropion)	205	300	227	111	116	0.25
	205	400	227	111	116	0.41
	212	300	289	144	145	0.61
Zurzuvae	PPD-301	50	183	93	90	72.34
(zuranolone)	PPD-201B	30	147	74	73	22.67
Exxua	ORG 134001	20 - 80	204	101	103	5.57
(gepirone)	FK-GBE-007	20 - 80	238	116	122	4.01
	ORG 134023	≥ 40	246	123	123	0.13
	FKBE008	≥ 40	195	96	99	0.61
	ORG 134002	≥ 40	205	102	103	0.33
	CN105-078	≥ 40	135	88	47	0.45
	CN105-083	≥ 40	112	73	39	0.27
	ORG 134017	40 - 80	318	159	159	0.07
	ORG 134004	20 - 80	254	124	130	0.06
	CN105-052	10 - 40	72	35	37	0.31
	ORG 134006	20 - 80	283	140	143	0.10
	CN105-053	10 - 60	112	56	56	0.86
Auvelity	AXS-05-MDD-201‡	45 + 105 †	80	43	37	3.62
(dextromethorphan + bupropion)	AXS-05-MDD-301	45 + 105 †	318	156	162	24.48

Zulresso	547-PPD-202A	90 µg/kg/h*	21	10	11	11.73
(brexanolone)	547-PPD-202B	60 µg/kg/h*	81	38	43	48.21
	547-PPD-202B	90 µg/kg/h*	84	41	43	4.17
	547-PPD-202C	90 µg/kg/h*	104	51	53	5.71

Note. *administered as intravenous infusion. † first value indicates the dose of dextromethorphan, the second value indicates the dose of bupropion. Dose ranges or minimal doses indicate a flexible-dose trial design ‡active comparator trial.

Table B2

Sample Sizes, BF and Follow-up Period (if applicable) for each Therapy Trial

Therapy	Reference	N	n (treatment)	n (control)	BF	Follow-up period*
		Post-treatment, control comparator				
RO-DBT	Lynch et al., 2020	183	121	62	4.20	
	Keogh et al., 2016	84	47	37	2.82	
MBCT	van Aalderen et al., 2012	205	102	103	434.18	
	Dimidjian et al., 2014	200	100	100	225025.53	
	Dimidjian et al., 2016	55	24	31	30.10	
	Cladder-Micus et al., 2018	96	44	52	1.42	
CBT-D	Wroe et al., 2018	115	63	52	1.41	
	Newby et al., 2017	77	31	46	73.96	
	Inouye et al., 2015	182	86	96	2.88	
	Safren et al., 2014	78	40	38	33.66	
	Sharif et al., 2014	54	28	26	61.47	
	Penckofer et al., 2012	65	29	36	4.70	
	Lustman et al., 1998	42	20	22	9.72	
R/LRT	Serrano et al., 2004	43	20	23	979.82	
	Haight et al., 2000	104	40	44	1.41	
	Areán et al., 1993	48	28	20	6.29	
	Fry, 1983	162	54	54	17.87	
	Youssef, 1990	60	21	21	2.27	
SM/SCT	Rokke et al., 2000	25	9	16	1.19	
	van den Hout et al., 1995	29	15	14	4.10	

	Stark et al., 1987	18	9	9	4.94
	Reynolds & Coats, 1986	19	9	10	16.85
	Rehm et al., 1979	24	14	10	5.06
	Fuchs & Rehm, 1977	18	8	10	42.77
BA	Kanter et al., 2015	43	21	22	2.03
IPT	Weissman et al., 1974	106	53	53	2.03
	Bolton et al., 2003	341	163	178	148.11
	Sinai & Lipsitz, 2012	17	9	8	0.73
	Sinai & Lipsitz, 2012	26	9	17	4.98
PST	Nezu, 1986	17	11	6	46.45
	Nezu & Perri, 1989	28	15	13	22199.43
	Nezu et al., 2003	89	45	44	4.83
CT	DeRubeis et al., 2005	120	60	60	18.63
	March et al., 2004	223	111	112	0.32
	Wuthrich & Rapee, 2013	62	27	35	224.98
	Troeung et al., 2014	18	11	7	9.20
ACT	Ataie et al., 2015	34	17	17	120.29
	Kohtala et al., 2015	57	28	29	717.07
	Losada et al., 2015	64	33	31	922.66
	Folke et al., 2012	34	18	16	8.30
	Bohlmeijer et al., 2011	93	49	44	22.30
	Petersen & Zettle, 2009	24	12	12	0.90
	Pots et al., 2016	169	82	87	17.45
	Lappalainen et al., 2015	38	18	20	24.23
	Carlbring et al., 2013	80	40	40	268.09

		Post-treatment, active comparator			
RO-DBT	Lynch et al., 2007	32	20	12	0.48
	Lynch et al., 2003	31	15	16	0.72
STS	Beutler et al., 1991	43	22	21	0.31
	Beutler et al., 1991	42	22	20	0.31
	Beutler et al., 2003	25	12	13	0.21
	Beutler et al., 2003	23	12	11	0.78
CBAS	Schatzberg et al., 2005	140	61	79	2.30
	Keller et al., 2000	436	216	220	0.15
	Keller et al., 2000	442	216	226	0.02
REBT	David et al., 2008	114	57	57	0.15
	David et al., 2008	113	57	56	0.65
R/LRT	Areán et al., 1993	47	28	19	0.10
SM/SCT	Dunn et al., 2007	77	33	44	4.60
	Rokke et al., 2000	18	9	9	0.64
	Stark et al., 1987	19	9	10	0.46
	Thomas et al., 1987	30	15	15	0.98
	Reynolds & Coats, 1986	20	9	11	0.21
	Fuchs & Rehm, 1977	18	8	10	5.60
SST	Eddington et al., 2015	49	22	27	1.15
	Strauman et al., 2006	45	24	21	0.89
BA	Dimidjian et al., 2006	43	22	21	4.06
	Dimidjian et al., 2006	60	22	38	0.33
	Hopko et al., 2001	80	42	38	0.27
	Ly et al., 2014	81	40	41	0.36

IPT	Weissman et al., 1979	81	17	23	0.30	
	Weissman et al., 1979	37	17	20	1.76	
	Weissman et al., 1979	38	17	21	10.99	
	Markowitz et al., 2005	47	23	24	3.31	
	Markowitz et al., 2005	44	23	21	14.45	
	Markowitz et al., 2008	26	14	12	0.44	
	DiMascio et al., 1979	81	40	41	14.84	
PST	Nezu, 1986	20	11	9	18.74	
	Nezu & Perri, 1989	30	15	15	12.72	
	Nezu et al., 2003	88	45	43	0.27	
CT	DeRubeis et al., 2005	180	60	120	0.28	
	March et al., 2004	220	111	109	6.80	
	Dimidjian et al., 2006	34	18	16	4.07	
	Dimidjian et al., 2006	45	18	27	11.08	
EFT	Goldman et al., 2006	72	36	36	2.47	
	Watson et al., 2003	85	40	45	0.32	
	Greenberg & Watson, 1998	34	17	17	8.09	
ACT	Losada et al., 2015	63	33	30	0.28	
	Tamannaeifer et al., 2014	19	10	9	0.46	
	Forman et al., 2007	99	55	44	0.25	
	Pots et al., 2016	149	82	67	3.10	
			Follow-up, control comparator			
RO-DBT	Lynch et al., 2020	167	112	55	0.37	11 months
MBCT	Dimidjian et al., 2016	50	21	29	38.62	6 months
CBT-D	Inouye et al., 2015	167	77	90	1.22	12 months

	Safren et al., 2014	68	38	30	1.06	12 months
	Sharif et al., 2014	57	28	29	6.05	2 months
	Penckofer et al., 2012	60	26	34	61.29	3 months
	Lustman et al., 1998	41	20	21	9.74	6 months
	Amsberg et al., 2009	69	32	37	4.14	1 year
	Snoek et al., 2008	86	45	41	0.86	1 year
	van der Ven et al., 2005	68	32	36	0.32	3 months
	Hermanns et al., 2015	181	93	88	4.87	1 year
R/LRT	Areán et al., 2007	356	265	91	138.40	1 year
	Haight et al., 2000	52	29	23	5.84	3 years
SM/SCT	Robinson-Whelen et al., 2007	96	42	54	4.94	3 months
	van den Hout et al., 1995	29	25	24	0.72	13 weeks
	Reynolds & Coats, 1986	19	9	10	16.19	5 weeks
	Rehm et al., 1979	24	14	10	2.77	6 weeks
short PDT	Simpson et al., 2003	145	73	72	0.30	12 months
IPT	Weissman et al., 1974	106	53	53	3.55	4 months
PST	Unützer et al., 2002	1759	889	870	54.27	12 months
	Garand et al., 2013	73	36	37	3.00	12 months
	Rivera et al., 2008	67	33	34	3.90	12 months
	Choi et al., 2014	102	63	39	0.46	36 weeks
	Ell et al., 2008	256	144	114	0.68	12 months
	Katon et al., 2004	288	146	142	13.73	12 months
	Dowrick et al., 2000	218	89	129	0.33	12 months
CT	Ebrahimi et al., 2013	31	16	15	14.93	3 months
	Watkins et al., 2011	299	140	159	86.05	6 months

	Clarke et al., 2005	152	77	75	1.54	52 weeks
	Stice et al., 2010	173	89	84	1.76	2 years
ACT	Losada et al., 2015	47	25	22	1.01	6 months
	Folke et al., 2012	34	18	16	10.70	18 months
	Bohlmeijer et al., 2011	93	49	44	23.46	3 months
	Hayes et al., 2011	12	8	4	11.77	3 months
Follow-up, active comparator						
RO-DBT	Lynch et al., 2007	31	17	14	0.53	6 months
MBCT	Kuyken et al., 2008	118	59	59	5.13	15 months
	Kuyken et al., 2015	336	169	167	0.52	24 months
	Shallcross et al., 2015	92	46	46	0.54	12 months
STS	Beutler et al., 1991	43	22	21	0.32	3 months
	Beutler et al., 1991	42	22	20	0.29	3 months
	Beutler et al., 2003	14	6	8	0.31	6 months
	Beutler et al., 2003	13	6	7	1.14	6 months
CBAS	Keller et al., 2000	436	216	220	0.13	3 months
	Keller et al., 2000	442	216	226	0.02	3 months
REBT	David et al., 2008	97	48	49	3.89	6 months
	David et al., 2008	95	48	47	0.68	6 months
R/LRT	Areán et al., 1993	47	28	19	0.17	3 months
SM/SCT	Dunn et al., 2007	66	29	37	0.14	1 year
	Rokke et al., 2000	20	12	8	0.62	1 year
	Stark et al., 1987	17	8	9	0.70	8 weeks
	Thomas et al., 1987	30	15	15	0.60	6 weeks
	Reynolds & Coats, 1986	20	9	11	0.53	5 weeks

	Fuchs & Rehm, 1977	18	8	10	4.18	6 weeks
BA	Ly et al., 2014	81	40	41	0.29	6 months
IPT	Weissmann et al., 1981	31	13	18	1.39	1 year
	Weissmann et al., 1981	28	13	15	1.04	1 year
	Weissmann et al., 1981	29	13	16	0.48	1 year
	de Mello et al., 2001	24	11	13	3.49	48 weeks
PST	Nezu, 1986	20	11	9	13078.93	6 months
	Hopko et al., 2013	80	38	42	0.36	12 months
	Choi et al., 2014	119	63	56	0.07	36 weeks
	Nezu & Perri, 1989	30	15	15	0.50	6 months
	Nezu et al., 2003	88	45	43	0.37	12 months
CT	Ebrahimi et al., 2013	32	16	16	0.19	3 months
	Ebrahimi et al., 2013	31	16	15	0.82	3 months
	Stice et al., 2010	177	89	88	0.27	2 years
	Stice et al., 2010	169	89	80	16.13	2 years
EFT	Greenberg & Watson, 1998	32	15	17	1.07	6 months
ACT	Losada et al., 2015	44	25	19	1.03	6 months
	Zettle & Rains, 1989	21	11	10	1.09	2 months
	Zettle & Rains, 1989	21	11	10	0.95	2 months
	Pots et al., 2016	149	82	67	0.26	12 months
	Lappalainen et al., 2014	35	19	16	10.69	18 months

Note. *Longest follow-up period for which effects were reported.