The comparative evidence-base for the efficacy of second-generation antidepressants in the treatment of depression: A Bayesian Meta-analysis

Rei Monden (MSc)\textsuperscript{1}, Annelieke M. Roest (PhD)\textsuperscript{1}, Don van Ravenzwaaij (PhD)\textsuperscript{2}, Eric-Jan Wagenmakers (PhD)\textsuperscript{3}, Richard Morey (PhD)\textsuperscript{4}, Klaas J. Wardenaar (PhD)\textsuperscript{1}, Peter de Jonge (PhD)\textsuperscript{1}

\textsuperscript{1} University of Groningen, University Medical Center Groningen, Department of Psychiatry, Interdisciplinary Center Psychopathology and Emotion regulation, Groningen, the Netherlands
\textsuperscript{2} University of Groningen, Faculty of Behavioural and Social Sciences, Groningen, the Netherlands
\textsuperscript{3} University of Amsterdam, Department of Psychology, Amsterdam, the Netherlands
\textsuperscript{4} Cardiff University, School of Psychology, Cognitive Science, Wales, United Kingdom

Funding:

The current study was supported by a VICI grant (no: 91812607) received by Peter de Jonge from the Netherlands organization for Scientific research (NWO-ZonMW).

Competing interests:

All authors have no competing interests to report.

Hanzeplein 1, P.O.Box 30.001 –CC72, 9700RB, Groningen, the Netherlands
+31 50 361 4490
Email address of the corresponding author:
RM: r.tendeiro-monden@umcg.nl
Abstract

Background: Studies showed similar efficacy for antidepressants to treat depression.

Aims: Quantify the comparative evidence-base for the efficacy of Food and Drug Administration (FDA) approved second-generation antidepressants for depression and to estimate pooled effect sizes for antidepressants in the Bayesian framework.

Method: Data were extracted from the FDA reviews, which is less liable to the selective publication, for 12 antidepressants by Turner et al (2008). Meta-analytic Bayes factors, which quantify the strength of evidence for the efficacy, were calculated and compared with the classical estimations.

Results: The evidence load for the efficacy varied strongly: duloxetine had the largest evidence load for its efficacy and venlafaxine had the highest estimated effect size. Bupropion had the lowest estimated effect size and evidence load.

Conclusions: The results illustrating the importance of considering effect sizes together with the evidence load. The Bayesian approach showed added value above traditionally used approaches.

Declaration of interests: None

Abstract: 150/150 words

Main text: 3526/4000 words
Introduction

Depression is one of the most important contributors to the global burden of disease\(^1\-^4\). In 2013, it was estimated that 15.7 million adults in the US had had at least one major depressive episode in the past year\(^5\). Given the large impact of depression on patients\(^6\) and society\(^7\), implementing effective treatment is a key priority. Antidepressant medication is one of the most common treatments for depression with about a third of severely depressed patients using antidepressants in the US\(^8\). The popularity of antidepressants is evident from the fact that their consumption in the US increased by almost 400% between 1988-1994 and 2005-2008\(^9\). The same trend was observed in other high income countries, with an average increase of 10% over the last decade\(^10\).

An important aspect of the pharmacological treatment of depression is the selection of an antidepressant. However, most antidepressants have been shown to have very similar, moderate effects in randomized controlled trials (RCT)\(^11\) and the APA’s revised Practice Guideline for the Treatment of Major Depressive Disorder concludes that many antidepressants are equally effective (P.33)\(^12\). The similar efficacy of antidepressants in RCTs could be due to different antidepressant compounds acting on the brain in a similar way\(^13\), resulting in similar effects when compared to a placebo (i.e. effect sizes around 0.3\(^14\-^16\)). The lack of a clear differentiation between the efficacy of various antidepressants has led to a large variability in prescription behavior among clinicians\(^17\), more reflective of clinicians’ personal preferences and experiences than actual scientific evidence and/or clear cut guidelines. This situation is unsatisfactory, but a better and more evidence-based way to choose between antidepressants has yet to be identified.

One way to gain more insight into differences between antidepressants’ efficacy could be to focus not only on estimated effect size, which is similar between antidepressant and therefore not useful for differentiation, but also on the presence of evidence load. Evidence load is the degree to which each antidepressant’s efficacy is supported by the available evidence. It
is important that evidence load is not confused with effect size: effect size quantifies the estimated effect of an antidepressant (e.g., the antidepressant reduces symptoms of depression by half an SD), whereas evidence load quantifies the strength of the evidence in favor of the estimated efficacy (e.g., strong evidence that the antidepressant reduces symptoms of depression). The results of such an analysis could help clinicians to choose the antidepressant with the highest evidence load for its efficacy from a range of antidepressants with comparable effect sizes. This can be done with Bayes factors (BFs)\textsuperscript{18-21}, which originate from Bayesian statistics and quantify the strength of evidence for an efficacy estimation.

To quantify the available evidence for different antidepressants’ efficacy, it is important to avoid the reporting bias as much as possible. Because the published literature on antidepressant efficacy has shown to over represent positive results\textsuperscript{15,22}, it is better to rely on data provided by the Food and Drug Administration (FDA). Trials are registered with, and results are reported to, the FDA by pharmaceutical companies to receive marketing approval. When the FDA approves a drug, the FDA reviews become publicly available, which are much less liable to the effects of publication bias. The current study aimed to quantify and compare the evidence-base for the efficacy of FDA approved second-generation antidepressants by means of BFs, using data extracted from the FDA reviews.

**Method**

**Data from FDA reviews**

The precise data extraction method is explained elsewhere\textsuperscript{15} and briefly summarized below. Phase 2/3 clinical trials were identified for the FDA approved antidepressants in the treatment of depression from 1987 to 2004. In total, reviews of 74 FDA registered trials were extracted for the following 12 second-generation antidepressants: bupropion, citalopram, escitalopram, fluoxetine, paroxetine, paroxetine controlled release [CR], sertraline, duloxetine, mirtazapine,
nefazodone, venlafaxine and venlafaxine extended release [XR]. In line with Turner et al.\textsuperscript{15}, the data for dosages ultimately approved by the FDA were included in the study, but the data for dosages ultimately disapproved by the FDA were excluded. From the obtained FDA review, the efficacy data on all randomized, double-blind, placebo-controlled studies for the short-term treatment of depression were extracted and analyzed in the study. Ethical approval was not required for the current study as our data came from previously published studies that were all received IRB approval.\textsuperscript{15}.

\textbf{Statistical Analysis}

\textit{Calculation of test statistics, pooled effect sizes and CIs}

BF was calculated for each antidepressant. To do this, the sample sizes and P-values reported in the FDA reviews, as summarized by Turner et al.\textsuperscript{15}, were used to calculate t-statistics. The latter, together with the sample sizes were needed to calculate the BF\textsubscript{s}. When the precise P-value was unavailable, we estimated the t-statistics using the following three approaches, which is consistent with Turner et al.\textsuperscript{15}: (a) the mean difference score was used together with the standard deviations/standard errors/confidence intervals (CIs) to calculate precise P-values, (b) the top of the reported P-value range was used as a precise P-value (e.g. p<0.001 $\rightarrow$ p=0.001), and (c) when the trial was published in agreement with the FDA conclusion (i.e. when the trial published in a journal had the same conclusion as the FDA’s conclusion or when no selective publication was found in Turner et al.\textsuperscript{15}), the precise P-value reported in the journal was used. If (a) was impossible, we applied (b). In case both (a) and (b) were impossible, we applied (c).

The above-mentioned approaches were not possible for 4 not-positive trials for paroxetine (FDA study number: 07, 09, UK-06 and UK-12) and 1 trial for sertraline (FDA study number: 310). We obtained approximations for these five trials by modeling all t-statistics/effect sizes for a given drug as coming from a truncated normal distribution. The truncation point differs for each of
those trials, depending on the sample sizes used in the trial and was the point for which the associated P-value would be .05. Modeling was done using JAGS\textsuperscript{23}, with R-package, rjags (version 4-5)\textsuperscript{24}.

For trials with a fixed-dose design, where drug dosages were set before the trial, the t-statistic was calculated for each of the dose-levels. For a flexible-dose design, where drug dosage could be increased or remained stable over time, one t-statistic was calculated for the whole range of dosages. For trials in which placebo performed better than the study drug, the t-statistics were multiplied by -1. To compare the Bayesian and classical effect size estimations, pooled effect sizes (Hedges’ g) and CIs were calculated by using a random effect-pooling method, performed in STATA version 13.1\textsuperscript{25}.

**Bayes factors, meta-analytic Bayes factors and Posterior effect sizes**

A Bayes factor (BF\textsubscript{10}), which ranges from 0 to infinity, is a ratio that quantifies the extent to which the data supports one hypothesis (H\textsubscript{1}) over another (H\textsubscript{0}). That is, a BF\textsubscript{10} quantifies the strength of evidence for the presence of the effect. It is important to note that BF\textsubscript{10} indicates the extent to which the *existence of the drug effect* is supported by the available evidence and is not a measure of the effect size. Suppose we define H\textsubscript{1} as “the tested antidepressant has a positive effect on treating depression” and H\textsubscript{0} as “the tested antidepressant has no effect on treating depression”, then, BF\textsubscript{10} is the ratio between the evidence that supports the effect of the tested antidepressant and the evidence that supports the effect of the placebo. The first subscription indicates the hypothesis that is used as the numerator, H\textsubscript{1} in our study, and the latter subscription indicates hypothesis that is used as the denominator, H\textsubscript{0} in our study. Therefore, BF\textsubscript{10}=1 means that the data equally support H\textsubscript{1} and H\textsubscript{0}, BF\textsubscript{10}>1 means the data support H\textsubscript{1} over H\textsubscript{0} (existence of antidepressant effect), and BF\textsubscript{10}<1 means the data support H\textsubscript{0} over H\textsubscript{1} (no efficacy).
Suppose two drugs (Drug A and Drug B) were independently tested against placebo and their $BF_{10}$ were 20 and 0.2, respectively. Then, three conclusions can be drawn from these results: (1) the existence of the effect of Drug A was supported 20 times more than the absence of the effect of Drug A (i.e., $20 \div 1 = 20$), (2) the absence of the effect of Drug B was supported 5 times more than the existence of the effect of Drug B (i.e., $1 \div 5 = 0.2$). Note that the evidence points towards the alternative hypothesis for Drug A, but towards the null hypothesis for Drug B, and (3) the existence of the effect of Drug A was supported 100 ($=20/0.2$) times more than that of Drug B. A $BF_{10}$ can be calculated based on a result from a single trial, but when multiple trials are conducted to test the efficacy of a single drug, the overall strength of evidence can be quantified as a meta-analytic Bayes factor (meta-BF). Moreover, the overall Bayesian estimation of the effect size, i.e. the posterior distribution of the effect size, can be also obtained when calculating a meta-BF. In sum, the Bayes factor can be used to quantify evidence in favor of either the null hypothesis or the alternative hypothesis, given the data. This is crucially different from the classical $P$-value, which only looks at the probability of data at least as extreme as those observed if the null hypothesis were true. As such, evidence in favor of the absence of an effect cannot be obtained with $P$-values and all antidepressants would eventually be concluded to be effective if one conducted enough trials.

Obviously, there is a correlation between the $BF$ and the number of conducted trials when there is an underlying effect (i.e. more evidence means more evidential strength). However, the difference between $P$-values and the $BF$ is that with larger data (more trials), a $BF$ either gets closer to infinity (stronger support for the existence of the effect) or closer to 0 (stronger support for the absence of the effect), while a $P$-value always becomes smaller and more likely to reject the null hypothesis.

To obtain an overview and to differentiate between antidepressants with respect to the strength of evidence for the existence of the effect and the effect size, the meta-BF was
calculated for each drug by using the R-package *BayesFactor* (version 0.9.10-2). The prior distribution (rscale) of the effect size was set to $\sqrt{2}/2$ (default) and one-sided analyses were performed since we assumed that the effect of the drug was either 0 or positive, and not negative. All t-statistics and the overall sample sizes of the drug and placebo groups for both fixed- and flexible-dose designs were combined for each antidepressant to calculate the meta-analytic BF and the posterior effect size. The prior distribution was set to $\sqrt{2}/2$ (default) to calculate meta-analytic BFs and the posterior effect size. Trials with fixed- and flexible-dose designs were combined based on the results of Khan et al., in which the authors did not find a dose-response relationship between antidepressants and symptom reduction.

**Sensitivity analysis**

To examine how the meta-BFs may differ based on the scale selection of the prior distribution, sensitivity analyses were conducted, by varying the scale from small ($1/3 \times \sqrt{2}$) to large ($3 \times \sqrt{2}$). A small scale for the prior distribution indicates a prior distribution of effect size sharply peaked around 0, which is a skeptical prior, whereas the large scale sets a prior distribution with a broader positive effect size, which is an optimistic prior. All statistical analyses, except for the calculation of the pooled effect sizes and CIs, were performed by using R version 3.2.3. All the R code used in the present study is presented as a supplement.

**Results**

**Meta-analytic Bayes factors**

*Figure 1* shows the results of meta-BFs in parentheses, together with the pooled effect sizes and CIs calculated in a classical way. To facilitate visualization, bupropion was set as a reference and the tested antidepressants were divided into four groups according to the logarithms of the meta-BF values. *Figure 1* shows that the strength of evidence for the efficacy
(i.e., the evidence load for the existence of the effect) among FDA approved antidepressants varies strongly. For instance, the evidence load for the efficacy of duloxetine was 5,120,300,699 (=20,481,202,796/4) times higher than for bupropion. The highest evidence load for efficacy was found for duloxetine, followed by venlafaxine and paroxetine. The lowest evidence load for efficacy was found for bupropion. A BF-range of 20-25 is often suggested to indicate “strong evidence”. Using this range, all antidepressants, except bupropion, were found to have at least “strong evidence” for their efficacy. Figure 1 also shows that the estimated pooled effect sizes and CIs were relatively similar for the different drugs, making them of limited use to differentiate between drugs.

**Posterior distribution of the effect sizes**

The posterior effect sizes for each of the antidepressants are presented in Figure 2, which displays the posterior density distribution of each antidepressant’s effect size estimation. The figure illustrates the variation across antidepressants in the certainty of effect size estimation. For instance, duloxetine shows a distribution with a relatively peaked density, indicating a higher certainty of the effect size estimation, whereas venlafaxine XR showed a distribution with a broader density, indicating a lower level of certainty. The peak of the posterior distribution indicates the most probable estimation of the effect size for each antidepressant. For instance, paroxetine shows a peak around 0.4, indicating the most probable effect size estimation lies around 0.4.

Similarly to the classical effect size estimations in Figure 1, the peaks of distributions in Figure 2 lay mostly between 0.2 and 0.4. However, Figure 2 illustrates additional differentiation between the drugs. Venlafaxine has the highest estimated effect size, followed by paroxetine and venlafaxine XR, while duloxetine has the highest estimated certainty for the effect size, followed by escitalopram. Bupropion has the lowest estimated effect size with moderate density
compared to the rest of the tested antidepressants, indicating that the effect size of bupropion is estimated to be the lowest among tested drugs (around 0.15) with some degree of certainty of the estimation. Importantly, Figure 2 shows that the total sample sizes do not necessarily correspond with the meta-BFs, i.e., a large total sample size does not necessarily mean that the meta-BF is high and vice versa.

**Sensitivity analysis**

The meta-BFs derived from small and large prior scales are presented in Table 1. With the skeptical prior (small scale) for the effect size, all the meta-BFs were higher than those with the medium scale, and meta-BFs with the medium scale were higher than those with the optimistic prior (large scale). This result is explained by the fact that the skeptical prior puts a higher expectation on effect sizes around 0, which is the case in this study, where effect sizes typically ranged from 0.2 to 0.4. Since the skeptical prior fit better to the data than the more optimistic priors, the resulting meta-BFs were highest. However, regardless of the prior scale selection, the efficacy of all FDA-approved second-generation antidepressants, except for bupropion, are supported by meta-BFs.

**Discussion**

To quantify the comparative evidence load for the efficacy of antidepressants, the efficacy of FDA-approved second-generation antidepressants to treat depression was re-evaluated in a Bayesian framework. It was demonstrated that the meta-BFs can be used to differentiate between second-generation antidepressant based on the available evidence-load. Although all studied antidepressants are approved by the FDA as “efficacious” drugs based on P-values, the meta-BFs varied strongly between them. Moreover, the presented posterior distributions of the effect sizes highlighted the differences between drugs in terms of the certainty of the effect size estimations. The results showed that whereas effect size estimations
overlapped across drugs, the actual evidence-load for each antidepressant’s efficacy varied strongly across drugs. The estimated effect size was shown to be the highest for venlafaxine, followed by paroxetine. Duloxetine had a lower effect size but the highest certainty of the effect size estimation. Bupropion had the lowest effect size as well as the lowest evidence for its efficacy. Interestingly, the results showed that meta-BFs are not necessarily related to the total sample size: large total sample sizes were not always associated with large meta-BFs. This is interesting because in the classical approach, when the sample size is large, the P-value is more prone to indicate a significant effect than when the sample size is small29. The results of this study have important clinical and theoretical implications.

Clinically, the evidence load and effect size certainty offers valuable information to help clinicians to select a drug. The APA guideline12 and several meta-analyses30-32 concluded that all second-generation drugs are equally efficacious to treat depression, although some meta-analyses found different results33-36. The current results provide a new perspective and demonstrate that the strength of evidence for the drugs’ efficacy varies considerably which can be used to guide drug-prescription in clinical practice. For example, given that all the tested drugs showed strong evidence for the existence of the treatment effect except for bupropion, clinicians could choose venlafaxine since this drug showed the highest effect size of all drugs with high evidence. Of course, additional clinically relevant aspects need to be considered when prescribing antidepressants (e.g. side effects, acceptability and comorbidity), but the evidence load for an antidepressant could be a good basic criterion for prescribers to keep in mind at all times. To offer the up-to-date knowledge of antidepressants efficacy, it is important to keep updating the current results with the unbiased results/evidence as more trials are conducted. Fortunately, meta-BFs and posterior distribution of the effect sizes can easily be updated by adding the new trial’s t-value and sample size (see the supplement code). Under the condition
that new trials are unbiased, like FDA documented studies, we can keep updating the meta-BFs and posterior effect size distributions.

From a theoretical perspective, the current results have several implications. First, an important feature of BFs is that they can distinguish between “evidence for absence” (e.g., BF\text{10} = 0.1) and “absence of evidence” (e.g., BF\text{10} = 1). Using the BF, the important distinction between these scenarios is possible, whereas it is not when using P-values. Second, comparisons between a classical and Bayesian way of evaluating drugs highlighted fundamental differences between the two approaches. The classical approach evaluates a trial in a dichotomous manner, typically with P<0.05, whereas the Bayesian approach quantifies the evidence in a continuous manner. This is important since all the tested drugs were ultimately concluded to be efficacious by the FDA, but strong variations were observed in the actual evidence load for efficacy. Problems relating to the use of P-values as a measure of evidence have often been discussed\textsuperscript{37-40} and the use of CIs and effect sizes has been encouraged\textsuperscript{41,42} to avoid dichotomous decisions based on P-values. However, a recent study showed that even CIs are typically misunderstood\textsuperscript{43} and they do not “index the plausibility that the true parameter is included in the interval”\textsuperscript{44}. That is, CIs cannot be used as a measure of estimation precision, while the Bayesian approach does show the estimate precision as reflected by the density distribution of the effect size certainty. Third, the Bayesian approach has been recommended over the classical approach\textsuperscript{18}, but has been difficult to perform in practice due to limited computational speed and lack of useable software. However, thanks to technical developments that have radically increased the speed of computers over the past decade, and the development of user-friendly programs\textsuperscript{26,45}, performing Bayesian analysis has become more feasible and simpler. As a result, it has become easier and more efficient to perform Bayesian analysis, including sensitivity analyses with various prior distributions. This is an important development that allows one to address the typical criticism of the Bayesian approach that the
subjective selection of the prior distribution determines the outcome to a large extent. Finally, by explicitly modeling the missing data reflecting the fact that the trials had P-values larger than .05, we provide estimates for the effect sizes that are likely to be much closer to the true value than, for instance, simply deleting those trials.

The current study has some limitations. First, we only analyzed data from trials which were registered with the FDA and the result could have been different if trials were included that were not used to obtain marketing approval from the FDA. Second, the current study did not address all the relevant questions to subscribe an antidepressant. The role of known relevant factors, such as initial severity\textsuperscript{46,47}, side effects\textsuperscript{48,49}, costs\textsuperscript{42} or comorbidity\textsuperscript{34} were not incorporated in the analyses. Utility analyses could be performed in the future to weight the evidence load and efficacy for each drug according to these factors. Third, the differences between the estimated BF$s and the effect sizes may due to the different study designs (such as the length of the studies, initial severity of patients) or data handling method (such as how the missing values were dealt to calculate t-statistics or P-values). These factors play an important role on comparing drug efficacies, but they are not reflected on the current study. A key strength of the present study was that the included trials are much less liable to the effects of publication bias.

In conclusion, the current study successfully differentiated the evidence load of the effect size estimates between the FDA-approved second-generation antidepressants in the treatment of depression. This can be a useful reference for clinicians when prescribing an antidepressant.
Reference


33. Lepola UM, Loft H, Reines EH. Escitalopram (10–20 mg/day) is effective and well tolerated in a placebo-controlled study in depression in primary care. *International clinical psychopharmacology*, 2003;18(4);211-217.
38. Ioannidis JP. Why most published research findings are false. *Chance*. 2005;18(4);40-47.


Table 1 Meta-BFs with three different prior scales

<table>
<thead>
<tr>
<th>Drug</th>
<th>Scale for the prior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small ($\frac{1}{3} \times \frac{\sqrt{2}}{2}$)</td>
</tr>
<tr>
<td>duloxetine</td>
<td>27,535,219,111</td>
</tr>
<tr>
<td>venlafaxine</td>
<td>13,966,627,690</td>
</tr>
<tr>
<td>paroxetine*</td>
<td>1,493,126,853</td>
</tr>
<tr>
<td>escitalopram</td>
<td>21,402,452</td>
</tr>
<tr>
<td>mirtazapine</td>
<td>19,357</td>
</tr>
<tr>
<td>venlafaxine XR</td>
<td>10,859</td>
</tr>
<tr>
<td>sertraline*</td>
<td>5,493</td>
</tr>
<tr>
<td>fluoxetine</td>
<td>2,944</td>
</tr>
<tr>
<td>citalopram</td>
<td>2,280</td>
</tr>
<tr>
<td>paroxetine CR</td>
<td>456</td>
</tr>
<tr>
<td>nefazodone</td>
<td>208</td>
</tr>
<tr>
<td>bupropion</td>
<td>8</td>
</tr>
</tbody>
</table>

Note. Drug names with * indicates that some test statistics were missing in the FDA reviews and therefore modeled and estimated by using JAGS.
Figure 1 The relationships between Hedges’ g and meta BFs

Meta-analytic Bayes factors are shown in brackets. The dots indicate effect size (Hedges’ g) presented in Turner et al\textsuperscript{15}. The intervals show 95% CIs of the effect sizes.
Figure 2. Posterior effect size differences between drugs

Density reflects the certainty of the estimated effect size and the peak of the distribution indicates the most probable estimated effect size.