Psychedelic Medicine
Volume 1, Number 1, 2023
© Mary Ann Liebert, Inc.
DOI: 10.1089/psymed.2022.0002

psychedelic medicine

RESEARCH ARTICLE

A Bayesian Reanalysis of a Trial of Psilocybin Versus Escitalopram for Depression

Sandeep M. Nayak,[1,*] Bilal A. Bari,[2] David B. Yaden,[1] Meg J. Spriggs,[3] Fernando E. Rosas,[3] Joseph M. Peill,[3] Bruna Giribaldi,[3] David Erritzoe,[3] David J. Nutt,[3] and Robin Carhart-Harris[4]

Abstract

Background: A trial of psilocybin (COMP360) versus escitalopram for major depressive disorder (MDD) was reported as negative, as there was no significant difference in the primary outcome, mean change in the 16-item Quick Inventory of Depressive Symptomatology–Self-Report (QIDS SR-16). However, analyses using three other depression scales (17-item Hamilton Depression Rating Scale [HAMD-17], Montgomery and Åsberg Depression Rating Scale [MADRS], and Beck Depression Inventory 1A [BDI-1A]) all significantly favored psilocybin, although without a prespecified plan for multiple comparisons correction.

Methods: Bayesian reanalysis of a trial of two doses of psilocybin (25 mg) versus 6 weeks of escitalopram (20 mg) was done in 59 patients with MDD. We used skeptical priors, which bias estimates toward zero, and Bayes factors, which quantify evidence for or against a hypothesis. We report posterior estimates for the difference between psilocybin and escitalopram for four different depression scales.

Results: Using Bayes factors and "skeptical priors" that bias estimates toward zero, for the hypothesis that psilocybin is superior by any margin, we found indeterminate evidence for QIDS SR-16, strong evidence for BDI-1A and MADRS, and extremely strong evidence for HAMD-17. For the stronger hypothesis that psilocybin is superior by a "clinically meaningful amount" (using literature-defined values of the minimally clinically important difference), we found moderate evidence against it for QIDS SR-16, indeterminate evidence for BDI-1A and MADRS, and moderate evidence supporting it for HAMD-17. For all scales, we found extremely strong evidence for psilocybin's noninferiority versus escitalopram. Findings were robust to prior sensitivity analysis.

Conclusions: The overall pattern of evidence provided by this Bayesian reanalysis supports the following inferences: (1) psilocybin did indeed outperform escitalopram in this trial, but not to an extent that was clinically meaningful and (2) psilocybin is almost certainly noninferior to escitalopram. These results provide a more precise and nuanced interpretation to previously reported results from this trial and support the need for further research into the relative efficacy of psilocybin therapy for depression with respect to current leading treatments.

Keywords: psilocybin, depression, Bayesian analysis, psychedelic

[1] Behavioral Pharmacology Research Unit, Center for Psychedelic and Consciousness Research, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
[2] Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA.
[3] Department of Medicine, Centre for Psychedelic Research, Imperial College, London, United Kingdom.
[4] Psychedelics Division, Neuroscape, Department of Neurology, University of California, San Francisco, California, USA.

*Address correspondence to: Sandeep M. Nayak, MD, Behavioral Pharmacology Research Unit, Center for Psychedelic and Consciousness Research, Johns Hopkins University School of Medicine, 5510 Nathan Shock Drive, Baltimore, MD 21224, USA, E-mail: smnayak1@gmail.com

Introduction

A recent trial investigating psilocybin's efficacy, relative to escitalopram, for major depressive disorder (MDD) reported no significant benefit relative to the standard of care.[1] Specifically, two high doses of psilocybin (25 mg) versus 20 mg escitalopram for 6 weeks did not show a significant difference with respect to the 16-item Quick Inventory of Depressive Symptomatology–Self-Report (QIDS SR-16) scores from 7 to 10 days preintervention to a 6-week endpoint, which was the primary outcome of this trial. However, a closer look at the results reveals that psilocybin significantly outperformed escitalopram on all secondary outcomes, including three clinically validated depression scales.

Because there was no prespecified plan for multiple comparisons corrections, the formally allowable frequentist interpretation was that the primary outcome was indeterminate and that the secondary outcomes were uninterpretable. A Bayesian approach has the potential to extract more interpretable information from the results of this trial, overcoming some key limitations of the previous frequentist analysis.

Frequentist and Bayesian approaches in clinical trials

The results of Carhart-Harris et al highlight several drawbacks of frequentist methods. First, frequentist methods suffer from several problems arising from multiple comparisons. Because *p*-values are uniformly distributed when the null hypothesis is true, 5% of tests will be positive by chance alone, when $\alpha = 0.05$. This necessitates special procedures to correct for multiple comparisons when multiple outcome measures are administered—a number of which can be arbitrary.[2] Second, frequentist methods do not convey the probability of any particular hypothesis, dealing instead with the probability of the data (or more extreme data) assuming the null hypothesis is true.

Because of this, *p*-values cannot be interpreted as measures of confidence on the findings. Third, these methods rigidly separate hypothesis testing from effect size estimation, and results are often reported that are statistically significant but clinically meaningless. Fourth, fixed sample sizes are chosen on the basis of *a priori* assumptions about the true effect size. If the actual effect size is smaller than anticipated, the trial is underpowered and may miss a real effect; hence, a null result provides no insight into whether this is due to a lack of power or due to a genuine absence of effect. However, if the actual effect size is much greater, then the trial collects superfluous participants.

An alternative approach is to employ methods of Bayesian inference. Although these methods are still less often used, they address many of the limitations of frequentist methods. First, with appropriately chosen priors, Bayesian inference can bypass the multiple comparisons problem.[3] Fewer false positive claims are made with confidence, which allows for more flexible use of multiple comparisons. Second, the Bayesian posterior distribution naturally allows for effect size estimation and hypothesis testing to be conducted simultaneously.

Third, and importantly for the specific case of clinical trials, Bayesian inference is flexible, modular, and allows for intuitive and meaningful clinical interpretations, rather than simple black/white dichotomization imposed by frequentist methods. In effect, the probability that a new intervention has any effect and the probability that it has a clinically meaningful effect (i.e., above an established criteria) can be determined naturally from the same posterior distribution. In addition, frequentist analyses can often be interpreted as special cases of Bayesian inference (i.e., when using uniform or ''flat'' priors), suggesting the two approaches are not entirely divorced from one another.[4]

Another important benefit of Bayesian analysis is that it allows us to quantify evidence for a hypothesis, rather than just evidence against a null, an advantage that we leverage here. Unlike *p*-values, which are simply positive or null, Bayes factors are tripartite, allowing us to distinguish positive, indeterminate, and null results.[5] Under a frequentist paradigm, null results may be truly null or may represent an underpowered study, and differentiating the two can be highly nontrivial. Because of this, no conclusions can be made in general from null results from a frequentist trial.

In contrast, Bayes factors naturally allow us to calculate the probability that a finding is truly negative versus indeterminate (requiring more data). This information can prove critical in determining whether to continue trials on a particular intervention (with a larger sample size) or to cease trials of said intervention all together. For these reasons, Bayesian analyses are becoming increasingly common in clinical medicine.

One useful example comes from the COVID STEROID 2 trial, which tested two different doses of dexamethasone in treating severe COVID-19 pneumonia. The study reported a null primary outcome, which was interpreted as null.[6] A Bayesian reanalysis concluded that the probability of any benefit of the higher dose was 95%, that of clinically important benefit was 62%, and that of clinically important harm was 0.2%.[7] Although not conflicting with the original frequentist study, this reanalysis offers a more complete clinically informative picture of the data.

Other examples include the ANDROMEDA-SHOCK trial[8] and a trial of Extra-Corporeal Membrane Oxygenation versus conventional ventilation,[9] each of which initially reported inconclusive primary outcomes with frequentist analyses, yet Bayesian reanalysis demonstrated high probability of benefit in each.[10,11] Each of these examples illustrates the usefulness of Bayesian reanalyses in better understanding clinical trial results that appeared ambiguous from the frequentist perspective.

Notably, it is not the case that Bayesian reanalyses simply convert null findings from frequentist trials into positive effects. On the contrary, a systematic review of Bayesian reanalyses of 82 studies in high-impact critical care journals found that discordance between frequentist and Bayesian results is uncommon.[12] In effect, in 78 of the 82 trials that were negative or indeterminate under frequentist criteria, Bayesian reanalysis found that clinically meaningful effects were probable in only seven trials (9%).

In 4 of the 82 trials with statistical significance for the intervention group, Bayesian reanalyses found positive results improbable in two trials (50%). As these findings demonstrate, Bayesian reanalyses are often more informative than the initial frequentist analysis—but Bayesian reanalyses do not represent a less conservative test of the purported benefit of a given intervention.

The present study

Given the success of Bayesian reanalyses, we suggest that the findings of Carhart-Harris et al can be better understood by subjecting them to a Bayesian reanalysis. In this study, we perform a Bayesian reanalysis of this trial to quantify the efficacy of psilocybin versus escitalopram in treating MDD. We test the hypothesis that psilocybin is superior to escitalopram using all four clinically validated depression inventories administered in the study, under both flat priors (largely equivalent to frequentist analyses) and skeptical priors (which bias effects toward zero and represent a more conservative approach).

Our results show that psilocybin indeed outperforms escitalopram, but not to an extent that is ''clinically meaningful''—defined using literature-defined scale-specific values of the minimally clinically important difference (MCID, see Methods section). Importantly, this reanalysis also provides additional insight into the seemingly incongruous ''null'' result on the QIDS, by distinguishing where evidence is truly indeterminate, and when it is in favor of the null. These results enrich and add context to the original trial, and support the need for further research into the relative efficacy of psilocybin therapy for depression, versus standard of care or any other viable active comparator with an evidence base.

Methods

Original study design

The original study compared 30 participants in the psilocybin group, and 29 participants in the escitalopram group. Participants in the psilocybin group received daily pill placebo and underwent two dosing sessions with high-dose psilocybin (25 mg) in psychologically supportive manner. The escitalopram group received escitalopram titrated to 20 mg daily for 6 weeks and underwent two dosing sessions with 1 mg psilocybin—effectively a placebo dose. The primary outcome was change in QIDS SR-16 at 6 weeks.

Bayesian linear regression

Bayesian linear models[13] were performed with all four depression scales used as outcome measures in the trial at 6 weeks: the QIDS SR-16, the 17-item Hamilton Depression Rating Scale (HAMD-17), the Montgomery and Åsberg Depression Rating Scale (MADRS), and the Beck Depression Inventory 1A (BDI-1A). All models took the following form, similar to the original analysis:

$$\mathrm{SCALE}_{\mathrm{FU}} = \beta_{\mathrm{C}} * \mathrm{Condition} + \beta_{\mathrm{BL}} * \mathrm{SCALE}_{\mathrm{BL}} + \nu,$$

where $\mathrm{SCALE}_{\mathrm{BL}}$ and $\mathrm{SCALE}_{\mathrm{FU}}$ are the values of a given scale at baseline and final follow-up, β_{C} and β_{BL} are the coefficients of a linear relationship between $\mathrm{SCALE}_{\mathrm{BL}}$ and condition (psilocybin or escitalopram group) as predictors of $\mathrm{SCALE}_{\mathrm{FU}}$, and ν is the residual of the regression. Put simply, the outcome variable was the follow-up score for each scale at 6 weeks, whereas condition and baseline depression scale score were used as independent variables.

Bayesian regression models need to specify prior distributions for their coefficients—in our case, for β_{C} and β_{BL}. For each outcome measure, two variants of the model were assessed that differed in the definition of their priors: a flat prior variant (which approximates frequentist methods) and a skeptical prior variant (which shrinks estimates closer to 0). Flat priors posit that any effect size is possible, and simply allow each parameter to take any value with uniform prior probability. Flat priors often produce results equivalent to frequentist approaches.

Skeptical priors instead posit that large effect sizes are unlikely. The skeptical priors were tuned such that the 95% highest density interval of the prior predictive distribution for group difference spans the magnitude of benchmark values for ''very much improved.'' In other words, this prior constrains effect sizes to be within a range that is considered clinically possible, and penalizes effects that are large. This skeptical prior signifies a belief that there is likely no group difference. Skeptical priors hence shrink estimates toward zero and are more conservative than flat priors and typical frequentist methods. Full details of these priors are available in the Supplementary Data.

For constructing the skeptical priors, the following benchmark values for ''very much improved'' were used. These criteria are based on values previously identified in the literature: QIDS 75% change from baseline[14];

HAMD-17 78% change from baseline, after averaging values from several citations[14–17]; and MADRS 82% change from baseline.[18] Finally, for BDI-1A, a 75% change from baseline was considered ''very much improved,'' following the benchmarks used for the other measures, since benchmark values of ''very much improved'' were not readily available in the literature for this scale.

Posterior distributions of depression scale scores were calculated for both psilocybin (COMPASS Pathways proprietary synthetic psilocybin, COMP360'') and escitalopram at the final follow-up (6-week timepoint), and the posterior distribution of their difference was calculated by subtracting one distribution from the other—yielding the ''posterior group difference.'' This posterior distribution can be summarized by its median value and by the upper and lower limits of the credible interval, which contains a given percentage (often 95%) of the posterior density.

Note that frequentist confidence intervals are often misinterpreted as denoting the probability that the interval contains the true value of a parameter of interest, or as capturing the number of times the true value would lie within the given interval if the study were run multiple times.[19] In contrast, the Bayesian credible interval can be interpreted more simply: given the data and the model, there is, for example, 95% probability that the true value lies within the interval.

Using the posterior group differences, the probabilities that psilocybin had (1) any superiority, (2) clinically meaningful superiority, and (3) noninferiority (NI) to escitalopram were calculated by taking the percentage of the posterior distribution (1) >0, (2) the MCID, and (3) the NI margin, respectively.

The MCID and NI margins were taken from the literature. The following values were used for MCID: QIDS 28.5% group difference,[14] HAMD-17 4 points,[20] MADRS 4.5 points,[20] and BDI-1A 29.64% group difference.[21] The following NI margins were used: QIDS −0.3 standardized difference from control,[22,23] MADRS −2.5 points,[24,25] and HAMD-17 – 2.5 points.[26,27] As NI margins were not readily available in the literature for BDI-1A, a conservative margin of −1 point was chosen.

All analyses were performed in R version 4.1.1[28] independently by two authors (S.M.N. and B.A.B.) to ensure similar results. Model parameters were estimated using Hamiltonian Markov Chain Monte Carlo simulations using both *brms*[29] and *rethinking*[13] packages, which are wrappers for the probabilistic programming language *Stan*. Analysis scripts are available at (https://osf.io/vfw7g/). Models were run with 4 chains and 2000 warm-up iterations, the default settings of the *brms* package.

Consent was not required for this reanalysis of previously collected data.

Bayes factors

We computed Bayes factors for two sets of hypotheses: that psilocybin outperforms escitalopram (1) by any amount and (2) by at least the MCID. Bayes factors comparing a specific H_1 (''experimental'' hypothesis) with H_0 (''null'' hypothesis) quantify the degree of evidence for H_1 versus H_0. For a given prior and posterior distribution, this Bayes factor (henceforth BF_{10}) can distinguish between null results and underpowered results—a useful property that is not possible with *p*-values.

For the hypothesis that psilocybin outperforms escitalopram by any amount, the experimental hypothesis is that the group difference is >0, whereas the null is that the group difference is 0. Mathematically:

$$\text{diff} = \text{SCALE}_{\text{FU}}^{\text{condition}=\text{escitalopram}} - \text{SCALE}_{\text{FU}}^{\text{condition}=\text{psilocybin}},$$

$$H_1 : \text{diff} \; 0,$$

$$H_0 : \text{diff} = 0.$$

To calculate BF_{10}, we take advantage of the following relationship:

$$\underbrace{\frac{P(H_1|D)}{P(H_0|D)}}_{\text{Posterior odds}} = \underbrace{\frac{P(D|H_1)}{P(D|H_0)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(H_1)}{P(H_0)}}_{\text{Prior odds}},$$

where the first term is the posterior odds, second term is the Bayes factor, and third term is the prior odds. We calculate the Bayes factor by dividing the posterior odds by the prior odds.

$$BF_{10} = \frac{P(D|H_1)}{P(D|H_0)} = \frac{P(H_1|D)}{P(H_0|D)} \Big/ \frac{P(H_1)}{P(H_0)}.$$

The prior odds can be interpreted as ''the odds of H_1 before seeing the data,'' and the posterior odds can be interpreted as ''the odds of H_1 after seeing the data.'' Greater values of the prior and posterior odds reflect greater plausibility of H_1 under those distributions.

BF_{10} is the ratio of these odds, where numbers >1 indicate more plausibility for H_1 after seeing the data, and numbers between 0 and 1 indicate more plausibility for H_0. For example, a BF_{10} of 5 means the data are five times more likely under H_1 than under H_0.

Using common convention, values of BF_{10} in the range 3–10 indicate moderate evidence, values in the range of 10–30 indicate strong evidence, values in the range of 30–100 indicate very strong, and values in the range >100 indicate extremely strong evidence for H_1.[30] These values can be inverted and interpreted similarly as evidence for H_0: a BF_{10} of 1/3–1/10 can be interpreted as strong evidence for H_0, with strength of evidence increasing as numbers approach 0. BF_{10} from 0.5 to 2 is usually considered to be indeterminate, requiring more evidence.

For the hypothesis that psilocybin is greater than escitalopram by a clinically meaningful amount (MCID), the following experimental and null hypotheses were used:

$$H_1 : \text{diff} > \text{MCID}\ H_0 : -\text{MCID} \leq \text{diff} \leq \text{MCID}.$$

Bayes factors were also computed for NI, using the following experimental and null hypotheses relative to the NI margin:

$$H_1 : \text{diff} > \text{NI}\ H_0 : \text{diff} < \text{NI}.$$

Prior sensitivity analysis

To ensure that results were not excessively impacted by the choice of priors, sensitivity analyses were performed using two additional sets of priors, in which the 95% highest density interval of the prior predictive distribution for group difference spanned 50% and 150% of the MCID. Further details about this procedure are available in the Supplementary Data.

Patient and public involvement

There was no patient nor public involvement in the design of this reanalysis.

Results

16-item Quick Inventory of Depressive Symptomatology–Self-Report

The median [95% CI] for the QIDS SR-16 group difference under a skeptical prior was 2.0 [−0.8 to 5.0] in favor of psilocybin, with a 92.0% probability for any positive effect and a 5.4% probability for a clinically meaningful difference. The Bayes factor for any positive effect was 1.2, indicating indeterminate evidence, which implies that the data are insufficient with respect to this question. The Bayes factor for a clinically meaningful difference was 0.14, indicating moderate evidence for the null of no clinically meaningful difference.

17-Item Hamilton Depression Rating Scale

The median [95% CI] for the HAMD-17 group difference under a skeptical prior was 5.3 [2.6–8.0] in favor of psilocybin, with a >99.99% probability for any positive effect and a 81.7% probability for a clinically meaningful difference. The Bayes factor for any positive effect was 363, indicating extremely strong evidence. The Bayes factor for a clinically meaningful difference was 6.1, indicating moderate evidence for a clinically meaningful difference.

Montgomery and Åsberg Depression Rating Scale

The median [95% CI] for the MADRS group difference under a skeptical prior was 7.0 [2.3–11.6] in favor of psilocybin, with a 99.7% probability for any positive effect and a 36.5% probability for a clinically meaningful difference. The Bayes factor for any positive effect was 25, indicating strong evidence. The Bayes factor for a clinically meaningful difference was 1.3, indicating indeterminate evidence.

Table 1. Adjusted Median Group Difference and Credible Interval [95%] in Depression Scale Scores at Final Follow-Up

Outcome	*Skeptical prior*	*Flat prior*
QIDS SR-16	2.0 [−0.8 to 5.0]	2.2 [−0.8 to 5.2]
HAMD-17	5.3 [2.6 to 8.0]	5.3 [2.3 to 8.2]
MADRS	7.0 [2.3 to 11.6]	7.2 [2.3 to 12.1]
BDI-1A	7.0 [1.6 to 12.2]	7.4 [1.8 to 12.9]

BDI-1A, Beck Depression Inventory 1A; HAMD-17, 17-item Hamilton Depression Rating Scale; MADRS, Montgomery and Åsberg Depression Rating Scale; QIDS SR-16, 16-item Quick Inventory of Depressive Symptomatology–Self-Report.

Beck Depression Inventory 1A

The median [95% CI] for the BDI-1A group difference under a skeptical prior was 7.0 [1.6–12.2] in favor of psilocybin, with a 99.4% probability for any positive effect and a 28.7% probability for a clinically meaningful difference. The Bayes factor for any positive effect was 12.6, indicating strong evidence, whereas the Bayes factor for a clinically meaningful difference was 1.0, indicating indeterminate evidence.

The probabilities (Bayes factor) for NI were QIDS: 99.67% (197), HAMD-17: >99.99% (infinite), MADRS: 99.98% (2831), and BDI-1A: 99.78% (398).

Sensitivity analyses using different priors did not substantially alter these results. Details of these analyses are available in the Supplementary Data.

Estimates for all four depression scales under skeptical and flat (not shown in text) priors are available in Table 1. Bayes factors for hypotheses of any superiority, clinically meaningful superiority, and NI are given in Table 2.

Table 2. Bayes Factors (BF_{10}) for Each of the Four Depression Scales on Three Hypotheses for Psilocybin Versus Escitalopram: Any Superiority, Clinically Meaningful Superiority, and Noninferiority

Outcome	*Any superiority*	*Clinically meaningful superiority*	*Noninferiority*
QIDS SR-16	1.2	0.14	197
HAMD-17	363	6.1	Infinite
MADRS	25	1.3	2831
BDI-1A	12.6	1.0	398

Values of BF_{10} in the range 3–10 indicate moderate evidence, values in the range of 10–30 indicate strong evidence, values in the range 30–100 indicate very strong, and values >100 indicate extremely strong evidence for the experimental hypothesis (H_1). Values of BF_{10} in the range of 0.33–0.1 can be interpreted as strong evidence for H_0, with strength of evidence increasing as numbers approach 0. Values of BF_{10} from 0.5 to 2 are usually considered to be indeterminate, requiring more evidence. Clinically meaningful superiority refers to a group difference greater than the minimally clinically important difference.

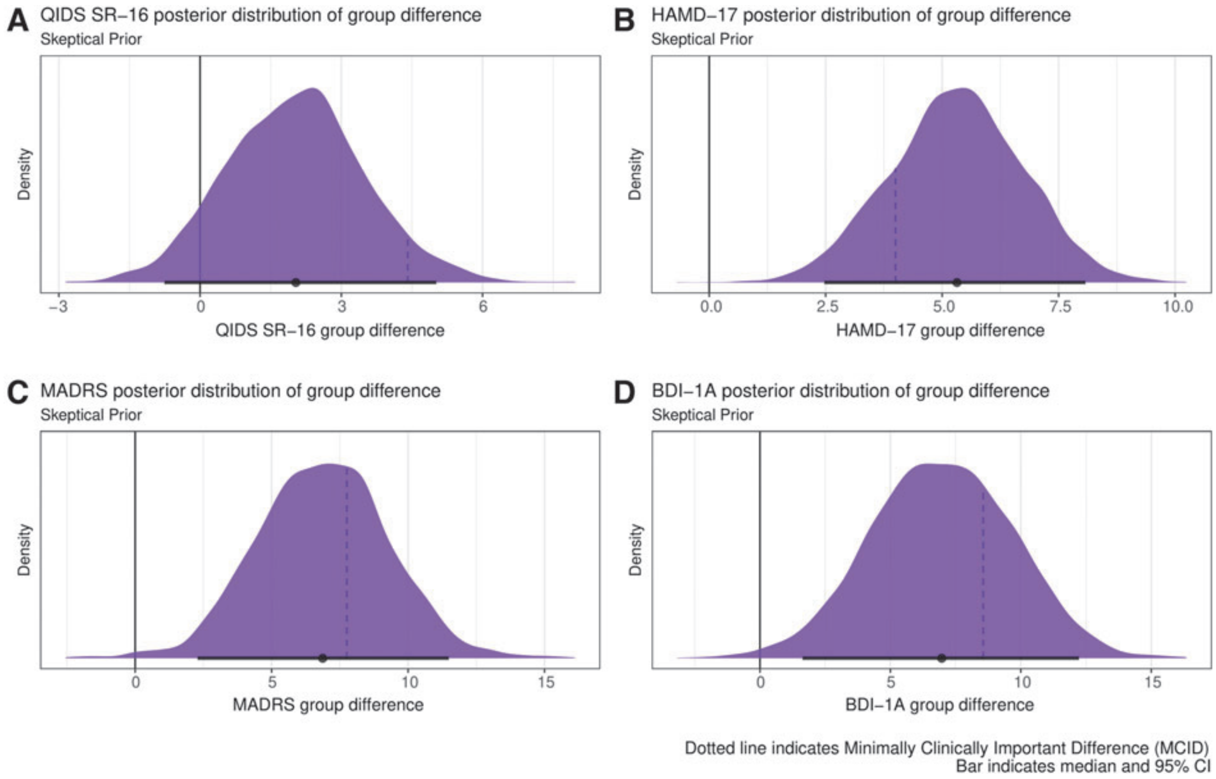

Fig. 1. Posterior distributions of group difference between psilocybin and escitalopram in the four depression scales used: QIDS SR-16 **(A)**, HAMD-17 **(B)**, MADRS **(C)**, and BDI-1A **(D)**.

Figure 1 shows the posterior distributions of group difference for all four scales under skeptical priors.

Discussion

This study presents a Bayesian reanalysis of a recently published study comparing psilocybin with escitalopram for the treatment of depression. Of the four depression scales included in this study, one failed to find a significant between-condition difference (QIDS SR-16) under the original frequentist analysis, whereas the remaining three found a significant difference in favor of psilocybin (BDI-1A, MADRS, HAMD-17). As the QIDS SR-16 was the predetermined primary outcome, the trial was considered indeterminate overall.

The Bayesian reanalysis presented here provides further insight into this trial's data, enabling clearer inferences to be made on them, and suggestions for future studies. Specifically, the results of the presented reanalysis suggest that psilocybin did indeed outperform escitalopram in this trial, but not to an extent that was clinically meaningful—whereas clarifying that more data are needed before these conclusions can be adopted with high confidence. In addition, results also support that psilocybin is almost certainly noninferior to escitalopram, as administered in this study. This Bayesian approach better allows full use of the available data in comparison with the frequentist approach.

Null hypothesis significance testing in the standard Neymann–Pearson methodology asks how probable the data are assuming H_0 is true, and is blind to the experimental hypothesis, H_1. Such a method can, therefore, not estimate the probability of H_0, or any other hypothesis. Alternatively, Bayesian methods can quantify the evidence for specific alternative and null hypotheses in intuitive probabilistic terms. This allows more direct answers to questions relevant to clinicians (e.g., ''what is psilocybin's effect on depression, how likely is that effect, and how certain can we be about it?'') rather than a mere dichotomous answer.

The current analysis investigated three hypotheses. For the hypothesis of any amount of superiority of psilocybin, there is indeterminate evidence (QIDS SR-16), strong evidence for H_1 (BDI-1A and MADRS), and extremely strong evidence for H_1 (HAMD-17). For the hypothesis that psilocybin is superior by a clinically meaningful amount, there is moderate evidence for H_0 (QIDS SR-16), indeterminate evidence (BDI-1A and MADRS), and moderate evidence for H_1 (HAMD-17). For all scales, there is extremely strong evidence for NI of psilocybin with respect to escitalopram.

Taken together, we can conclude that in this study population, psilocybin is probably superior to escitalopram, but not to a degree that is clinically meaningful, and that psilocybin is almost certainly noninferior to escitalopram. Although none of these conclusions conflicts with the results of the original article, they are much more informative and nuanced than the conclusions of frequentist analysis. Notably, psilocybin's adverse effects tended to be limited to the 24 h after the dosing sessions, in contrast to escitalopram.[1] Thus, even if psilocybin were noninferior but not superior to escitalopram, it may have a more favorable risk–benefit ratio.

In Carhart-Harris et al, the primary outcome measure (QIDS SR-16) yielded a nonsignificant result, whereas psilocybin was superior in every contrast using secondary efficacy outcome measures (including HAMD-17, MADRS, and BDI-1A). Nevertheless, frequentist conventions required this be reported as a null trial (i.e., that ''the primary outcome is indeterminate and the secondary outcomes uninterpretable''). As a thought experiment, imagine an alternative plausible outcome: the primary outcome significantly favored psilocybin and yet every secondary outcome was null. Although such results could be reported as proof of psilocybin's superiority over escitalopram, we suspect many readers would be skeptical of this interpretation—suspecting it to be a false positive.

Under a Bayesian analysis, the individual scales continue to offer contrasting evidence. For example, for the hypothesis of clinically meaningful superiority of psilocybin, there is moderate evidence against (i.e., H_0) according to the QIDS SR-16, whereas there is moderate evidence for (H_1) according to the HAMD. Future study could be done to address the relative strengths and weaknesses of the depressive symptom severity rating scales used in this trial, which may further aid our abilities to draw inferences on this trial's results and also may contribute to the design of future trials.

However, a Bayesian reanalysis with skeptical priors allows us to analyze the findings from each of the scales in their totality.[3] This provides a more informative picture of the results of the trial by considering all of the available data while remaining robust to problems resulting from multiple comparisons.

Bayesian methods have been critiqued as unnecessarily subjective, given the need for a prior distribution. We view this argument as a red herring, as frequentist clinical trials typically use substantial prior information in the design of the trial, particularly in estimating the number of subjects that must be enrolled to avoid an underpowered result. In addition, some frequentist methods are equivalent to Bayesian inference with uniform priors, demonstrating that priors are implicitly a feature of frequentism. The implicit flat prior distributions that characterize frequentist analyses are often inappropriate statistically (causing problems with model convergence) and logically (rendering extreme effect sizes as probable as small ones).[31]

Bayesian principles extend far beyond inference performed at the end of data collection, offering important advantages in the design of clinical trials. In powering a trial, frequentist methods typically establish a fixed sample size based on a prior assumption of effect size, which is often uncertain. If a null result is obtained, it can be unclear whether the result is truly null or underpowered, despite best attempts at collecting an appropriate number of subjects. Sequential designs are possible, and occasionally used, though this requires a rigid design with prespecified looks at the data.

A more flexible and intuitive approach is a Bayesian sequential trial.[32] A Bayesian sequential trial might, for example, target a specified strength of evidence (applicable to H_1 or H_0) using Bayes factors, and continue collecting participants until that strength of evidence is reached.[32–34] This method can not only allow continued data collection if results are indeterminate, but also permits ending trials earlier with lower sample sizes when effects are larger than expected.[35] Had the original study taken this approach, data collection could have continued until the evidence for QIDS SR-16 was no longer indeterminate.

Equally, a trial can be terminated early if there is sufficient evidence of no benefit (i.e., in support of H_0), which is often not possible with standard frequentist design. Bayesian sequential design also obviates problems related to findings that are statistically significant but not clinically significant, as the choice of H_1 can be a clinically meaningful difference.

Overall, this article illustrates several of the advantages of Bayesian methods for the design and analysis of clinical trials. First, specific alternative and null hypotheses can be clearly specified as the subject of the analysis. The evidence for these hypotheses can be presented in intuitive probabilistic terms, or through Bayes factors that provide a quantitative assessment about the strength of one hypothesis over another. When there is limited prior information to go on, as in the case of a psilocybin trial directed at a novel therapeutic indication, Bayesian sequential trials allow a more flexible trial design that may on average save resources[32] while remaining rigorous and principled. Given these advantages, we believe Bayesian methods deserve greater use in psychedelic clinical trials in particular and clinical trials in general.

Acknowledgments

We thank Allan Blemings for assistance in understanding the original analysis and Richard McElreath for his excellent Bayesian textbook, *Statistical Rethinking*.

Authors' Contributions

S.M.N. contributed to conceptualization, formal analysis, methodology, software, visualization, writing—original draft, and writing—review and editing. B.A.B. was involved in conceptualization, formal analysis, methodology, visualization, validation, and writing—review and editing. D.B.Y. carried out conceptualization and writing—review and editing. M.J.S. took charge of writing—review and editing and conceptualization. F.E.R. carried out writing—review and editing. J.M.P. was involved in data curation, writing—review and editing. B.G. carried out investigation, project administration, and writing—review and editing. D.E. was in charge of investigation and writing—review and editing. D.J.N. was involved in investigation and writing—review and editing. R.C.-H. was in charge of investigation, project administration, resources, supervision, and writing—review and editing.

Author Disclosure Statement

R.C.-H. reports receiving consulting fees or stock options from Journey Collab, Entheon Biomedical, Beckley Psytech, Mydecine, Tryp Therapeutics, and Maya Health; B.G. reports receiving consulting fees from Small Pharma Ltd; D.E. received consulting fees from Field Trip and Mydecine; and D.J.N. received consulting fees from Algernon, H. Lundbeck and Beckley Psytech, advisory board fees from COMPASS Pathways and lecture fees from Takeda and Otsuka and Janssen, plus owns stock in Alcarelle, Awakn, and Psyched Wellness. The other authors declare no competing interests.

Funding Information

Support for S.M.N. and D.B.Y. through the Johns Hopkins Center for Psychedelic and Consciousness Research was provided by Tim Ferriss, Matt Mullenweg, Blake Mycoskie, Craig Nerenberg, and the Steven and Alexandra Cohen Foundation. Support for B.A.B. comes from NIMH grant R25MH094612. R.C.-H. is the Ralph Metzner Chair of the Psychedelic Division, Neuroscape at University of California San Francisco.

Supplementary Material

Supplementary Data

References

1. Carhart-Harris R, Giribaldi B, Watts R, et al. Trial of psilocybin versus escitalopram for depression. N Engl J Med 2021;384(15):1402–1411.
2. Sjölander A, Vansteelandt S. Frequentist versus Bayesian approaches to multiple testing. Eur J Epidemiol 2019;34(9):809–821.
3. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff 2012;5(2):189–211.
4. Bayarri MJ, Berger JO. The interplay of Bayesian and frequentist analysis. Stat Sci 2004;19(1):58–80.
5. Keysers C, Gazzola V, Wagenmakers EJ. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. Nat Neurosci 2020;23(7):788–799.
6. Russell L, Uhre KR, Lindgaard ALS, et al. Effect of 12mg vs 6mg of dexamethasone on the number of days alive without life support in adults with COVID-19 and severe hypoxemia: The COVID STEROID 2 randomized trial. JAMA 2021;326(18):1807–1817.
7. Granholm A, Munch MW, Myatra SN, et al. Dexamethasone 12mg versus 6mg for patients with COVID-19 and severe hypoxaemia: A pre-planned, secondary Bayesian analysis of the COVID STEROID 2 trial. Intensive Care Med 2022;48(1):45–55.
8. Hernández G, Ospina-Tascón GA, Damiani LP, et al. Effect of a resuscitation strategy targeting peripheral perfusion status vs serum lactate levels on 28-day mortality among patients with septic shock: The ANDROMEDA-SHOCK randomized clinical trial. JAMA 2019;321(7):654–664.
9. Combes A, Hajage D, Capellier G, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. N Engl J Med 2018;378(21):1965–1975.
10. Zampieri FG, Damiani LP, Bakker J, et al. Effects of a resuscitation strategy targeting peripheral perfusion status versus serum lactate levels among patients with septic shock. A Bayesian reanalysis of the ANDROMEDA-SHOCK trial. Am J Respir Crit Care Med 2020;201(4):423–429.
11. Goligher EC, Tomlinson G, Hajage D, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial. JAMA 2018;320(21):2251.
12. Yarnell CJ, Abrams D, Baldwin MR, et al. Clinical trials in critical care: Can a Bayesian approach enhance clinical and scientific decision making? Lancet Respir Med 2021;9(2):207–216.
13. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd ed. Boca Raton, FL: CRC Press; 2020.
14. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54(5):573–583.
15. Furukawa TA, Akechi T, Azuma H, et al. Evidence-based guidelines for interpretation of the Hamilton rating scale for depression. J Clin Psychopharmacol 2007;27(5):531–534.
16. Leucht S, Fennema H, Engel R, et al. What does the HAMD mean? J Affect Disord 2013;148(2–3):243–248.
17. Bobo WV, Angleró GC, Jenkins G, et al. Validation of the 17-item Hamilton depression rating scale definition of response for adults with major depressive disorder using equipercentile linking to clinical global impression scale ratings: Analysis of pharmacogenomic research network antidepressa: Validation of HDRS definition of response. Hum Psychopharmacol 2016;31(3):185–192.
18. Leucht S, Fennema H, Engel RR, et al. What does the MADRS mean? Equipercentile linking with the CGI using a company database of mirtazapine studies. J Affect Disord 2017;210:287–293.
19. Hoekstra R, Morey RD, Rouder JN, et al. Robust misinterpretation of confidence intervals. Psychon Bull Rev 2014;21(5):1157–1164.
20. Hengartner MP, Plöderl M. Estimates of the minimal important difference to evaluate the clinical significance of antidepressants in the acute treatment of moderate-to-severe depression. BMJ Evid Based Med 2022;27(2):69–73.
21. Wilson HD. Minimum clinical important differences of health outcomes in a chronic pain population: Are they predictive of poor outcomes? ProQuest Dissertations and Theses. 2007; p. 230. Available from: https://www.proquest.com/dissertations-theses/minimum-clinical-important-differences-health/docview/304711319/se-2?accountid=11752 (accessed August 26, 2022).
22. Mechler J, Lindqvist K, Carlbring P, et al. Internet-based psychodynamic versus cognitive behaviour therapy for adolescents with depression: Study protocol for a non-inferiority randomized controlled trial (the ERiCA study). Trials 2020;21(1):587.
23. Mohr DC, Lattie EG, Tomasino KN, et al. A randomized noninferiority trial evaluating remotely-delivered stepped care for depression using internet cognitive behavioral therapy (CBT) and telephone CBT. Behav Res Ther 2019;123:103485.
24. Bauer M, Dell'Osso L, Kasper S, et al. Extended-release quetiapine fumarate (quetiapine XR) monotherapy and quetiapine XR or lithium as add-on to antidepressants in patients with treatment-resistant major depressive disorder. J Affect Disord 2013;151(1):209–219.
25. Andersson G, Hesser H, Veilord A, et al. Randomised controlled non-inferiority trial with 3-year follow-up of internet-delivered versus face-to-face group cognitive behavioural therapy for depression. J Affect Disord 2013;151(3):986–994.
26. Gibbons MBC, Gallop R, Thompson D, et al. Comparative effectiveness of cognitive and dynamic therapies for major depressive disorder in a community mental health setting: a randomized non-inferiority trial. JAMA Psychiatry 2016;73(9):904–911.

27. Szegedi A, Kohnen R, Dienel A, et al. Acute treatment of moderate to severe depression with hypericum extract WS 5570 (St John's wort): Randomised controlled double blind non-inferiority trial versus paroxetine. BMJ 2005;330(7490):503.
28. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2020. Available from: https://www.R-project.org/ (accessed August 26, 2022).
29. Bürkner PC. Advanced Bayesian multilevel modeling with the R package brms. R J 2018;10(1):395–411.
30. Quintana DS, Williams DR. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP. BMC Psychiatry 2018;18(1):1–8.
31. Van Dongen S. Prior specification in Bayesian statistics: Three cautionary tales. J Theor Biol 2006;242(1):90–100.
32. Schönbrodt FD, Wagenmakers EJ, Zehetleitner M, et al. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. Psychol Methods 2017;22(2):322–339.
33. Schönbrodt FD, Wagenmakers EJ. Bayes factor design analysis: Planning for compelling evidence. Psychon Bull Rev 2018;25(1): 128–142.
34. Wagenmakers EJ, Wetzels R, Borsboom D, et al. An Agenda for purely confirmatory research. Perspect Psychol Sci 2012;7(6):632–638.
35. Moerbeek M. Bayesian updating: Increasing sample size during the course of a study. BMC Med Res Methodol 2021;21(1):137.