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], Robin Carhart-Harris[4]

[1]Center for Psychedelic and Consciousness Research, Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[2]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA

[3]Centre for Psychedelic Research, Department of Medicine, Imperial College London, UK

[4]Psychedelics Division, Neuroscape, Department of Neurology, University of California, San Francisco, CA, USA

Corresponding Author:

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

smn@jhmi.edu

Word count: 3558

Abstract

Objectives: To perform a Bayesian reanalysis of a recent trial of psilocybin (COMP360) versus escitalopram for Major Depressive Disorder (MDD) in order to provide a more informative interpretation of the indeterminate outcome of a previous frequentist analysis.

Design: Reanalysis of a two-arm double-blind placebo controlled trial.

Participants: Fifty-nine patients with MDD.

Interventions: Two doses of psilocybin 25mg and daily oral placebo versus daily escitalopram and 2 doses of psilocybin 1mg, with psychological support for both groups.

Outcome measures: Quick Inventory of Depressive Symptomatology–Self-Report (QIDS SR-16), and three other depression scales as secondary outcomes: HAMD-17, MADRS, and BDI-1A.

Results: Using Bayes factors and 'skeptical priors' which bias estimates towards 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. Furthermore, across the board we found extremely strong evidence for psilocybin's non-inferiority versus escitalopram. These findings were robust to prior sensitivity analysis.

Conclusions: This Bayesian reanalysis supports the following inferences: 1) that psilocybin did indeed outperform escitalopram in this trial, but not to an extent that was clinically meaningful—-and 2) that psilocybin is almost certainly non-inferior to escitalopram. The present 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.

Trial registration number: NCT03429075

Strengths and limitations:

Relative to the original frequentist analysis, this Bayesian reanalysis allows several advantages:

- Provides a more intuitive, probabilistic interpretation of trial results that is robust to multiple comparisons.
- Clarifies where indeterminate frequentist results are null versus underpowered.
- Quantifies the evidence for several meaningful hypotheses (any effect, clinically meaningful effect, and non-inferiority).

This study nonetheless shares the same limitations of the original trial's study design, namely unbliding and expectancy effects that may inflate group differences. This is mitigated to some extent by the use of skeptical priors which bias effect estimates towards zero.

Introduction

A recent trial investigating psilocybin's efficacy, relative to escitalopram, for major depressive disorder reported no significant benefit relative to the standard of care (Carhart-Harris et al. 2021). Specifically, psilocybin did not show a significantly difference with respect to the Quick Inventory of Depressive Symptomatology–Self-Report (QIDS SR-16) scores from 7-10 days pre-intervention 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 pre-specified 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. (2021) 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 $\alpha = .05$. This necessitates special procedures to correct for multiple comparisons when multiple outcome measures are administered—a number of which can be arbitrary (see Sjölander and Vansteelandt 2019). 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. On the other hand, 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. Firstly, with appropriately chosen priors, Bayesian inference can bypass the multiple comparisons problem (Gelman, Hill, and Yajima 2012). 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. Additionally, 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 (Bayarri and Berger 2004).

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 which 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 (Keysers, Gazzola, and Wagenmakers 2020). Under a frequentist paradigm, null results may be truly null or may represent an underpowered study, and differentiating the two can be highly non-trivial. Because of this, no conclusions can be made in general from a null results from a frequentist trial. In contrast, Bayes factors naturally allow us to calculate the probability that a finding is truly negative vs 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 (Russell et al. 2021). A Bayesian reanalysis concluded that the probability of any benefit of the higher dose was 95%, of clinically important benefit was 62%, and of clinically important harm was 0.2% (Granholm et al. 2021). While 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 (Hernández et al. 2019) and a trial of Extra-Corporeal Membrane Oxygenation vs conventional ventilation (Combes et al. 2018), each of which initially reported inconclusive primary outcomes with frequentist analyses, yet Bayesian reanalysis demonstrated high probability of benefit in each (Zampieri et al. 2020; Goligher et al. 2018). Each of these examples illustrate 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 (Yarnell et al. 2021). 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 7 (9%). In 4 of the 82 trials with statistical significance for the intervention group, Bayesian reanalyses found positive results improbable in 2 (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 the Carhart-Harris et al. (2021) trial can be better understood by subjecting them to a Bayesian reanalysis. Here, we

perform a Bayesian reanalysis of Carhart-Harris et al. (2021) to quantify the efficacy of psilocybin versus escitalopram in treating major depressive disorder. 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 towards 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). 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 Carhart-Harris et al. (2021), 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

Bayesian linear regression

Bayesian linear models (McElreath 2020) were performed with each of the depression scales that were used as outcome measures in the trial: the 16-item Quick Inventory of Depressive Symptomatology–Self-Report (QIDS SR-16), the 17-item Hamilton Depression Rating Scale (HAMD-17), the Montgomery and Asberg Depression Rating Scale (MADRS), and the Beck Depression Inventory 1A (BDI-1A). All models took the following form, similar to the original analysis:

$$\text{SCALE}_{\text{FU}} = \beta_C * \text{Condition} + \beta_{BL} * \text{SCALE}_{\text{BL}} + v,$$

where SCALE_{BL} and SCALE_{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 SCALE_{BL} and condition (psilocybin or escitalopram group) as predictors of SCALE_{FU} and v is the residual of the regression. Put simply, the outcome variable was the follow-up score for each scale at 6 weeks, while 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 materials.

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 (Rush et al. 2003); HAMD-17 78% change from baseline, after averaging values from several citations (Rush et al. 2003; Furukawa et al. 2007; Leucht et al. 2013; Bobo et al. 2016); MADRS 82% change from baseline (Leucht et al. 2017). 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 (Hoekstra et al. 2014). In contrast, the Bayesian credible interval can be interpreted more simply: given the data and the model, there is a e.g. 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) non-inferiority to escitalopram were calculated by taking the percent of the posterior distribution 1) greater than 0, 2) the minimally clinically important difference (MCID), and 3) the non-inferiority margin, respectively.

The MCID and non-inferiority margins were taken from the literature. The following values were used for MCID: QIDS 28.5% group difference (Rush et al. 2003); HAMD-17 4 points (Hengartner and Plöderl 2021); MADRS 4.5 points (Hengartner and Plöderl 2021); BDI-1A 29.64% group difference (Wilson 2007). The following non-inferiority margins were used: QIDS -0.3 standardized difference from control (Mechler et al. 2020; Mohr et al. 2019); MADRS -2.5 points (Bauer et al. 2013; Andersson et al. 2013); HAMD-17 -2.5 points (Gibbons et al. 2016; Szegedi et al. 2005). As non-inferiority 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 (R Core Team 2020) independently by two authors (SMN and BAB) to ensure similar results. Model parameters were estimated using Hamiltonian Markov Chain Monte Carlo simulations using both *brms* (Bürkner 2018) and *rethinking* (McElreath 2020) packages, which are wrappers for the probabilistic programming language *Stan*. Visual inspection of posterior predictive checks (demonstrating that simulated data adequately approximate real data) and trace plots (showing adequate chain mixing) suggested reasonable model specification. Analysis scripts are available at https://osf.io/vfw7g/.

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) to 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 greater than zero, while the null is that the group difference is zero. Mathematically:

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

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

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

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 prior to 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 greater than 1 indicate more plausibility for H_1 after seeing the data, and numbers between 0 and 1 indicating more plausibility for H_0. For example, a BF_{10} of 5 means the data are 5 times more likely under H_1 than 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, 30–100 very strong, and greater than 100 extremely strong evidence for H_1 (Quintana and Williams 2018). 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 0.5–2 are 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 non-inferiority, using the following experimental and null hypotheses relative to the non-inferiority margin (NI):

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

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 can be found in the supplemental material.

Results

QIDS SR-16

The median [95% CI] for QIDS SR-16 group difference under a skeptical prior was 2.0 [-0.8, 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.

HAMD-17

The median [95% CI] for HAMD-17 group difference under a skeptical prior was 5.3 [2.6, 8.0] in favor of psilocybin, with a 100% 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.

MADRS

The median [95% CI] for 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.

BDI-1A

The median [95% CI] for 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, while the Bayes factor for a clinically meaningful difference was 1.0 indicating indeterminate evidence.

Estimates for all four depression scales under skeptical and flat (not shown in text) priors is available in Table 1.

The probabilities (Bayes factor) for non-inferiority were QIDS: 99.67% (197), HAMD-17: 100% (infinite), MADRS: 99.98% (2831), BDI-1A: 99.78% (398).

Sensitivity analyses using different priors did not substantially alter these results. Details of these analyses can be found in the supplementary material.

Discussion

This study presents a Bayesian reanalysis of data from a recently published study comparing psilocybin to 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, while the remaining three found a significant difference in favor of psilocybin (BDI-1A, MADRS, HAMD-17). As the QIDS SR-16 was the pre-determined 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 suggests that psilocybin did indeed outperform escitalopram in this trial, but not to an extent that was clinically meaningful—while clarifying that more data is needed before these conclusions can be adopted with high confidence. In addition, results also support that psilocybin is almost certainly non-inferior to escitalopram, as administered in this study.

Null hypothesis significance testing in the standard Neymann-Pearson methodology asks how probable the data are under the assumption that H_0 is true, and is blind to the experimental hypothesis, H_1. Such a method can therefore not directly 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 offering a mere dichotomous answer.

Harnessing this capacity, 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). Across the board there is extremely strong evidence for non-inferiority of psilocybin with respect to escitalopram.

Taken together, we can conclude that in this study population psilocybin is probably superior to escitalopram, but not clearly to a degree that is clinically meaningful, and that psilocybin is

almost certainly non-inferior to escitalopram. While none of these conclusions conflicts with the results of the original manuscript, they are much more informative and nuanced than the conclusions of frequentist analysis.

In Carhart-Harris (2021), the primary outcome measure (QIDS SR-16) yielded a non-significant result, while 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, while there is moderate evidence for (H_1) according to the HAMD. Future work 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 re-analysis with skeptical priors allows us to analyze the findings from each of the scales in their totality (Gelman, Hill, and Yajima 2012). 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) (Van Dongen 2006).

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 (Schönbrodt et al. 2017). 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 (Schönbrodt et al. 2017; Schönbrodt and Wagenmakers 2018; Wagenmakers et al. 2012). 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 (Moerbeek 2021). Had Carhart-Harris et al. (2021) taken this approach, data collection would have been allowed to continue until the evidence for QIDS SR-16 was no longer intederminate. 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. Firstly, 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 via 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 allows a more flexible trial design that may on average save resources (Schönbrodt et al. 2017) 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.

Table 1. Adjusted median group difference and credible interval [95%] in depression scale scores at final follow-up.

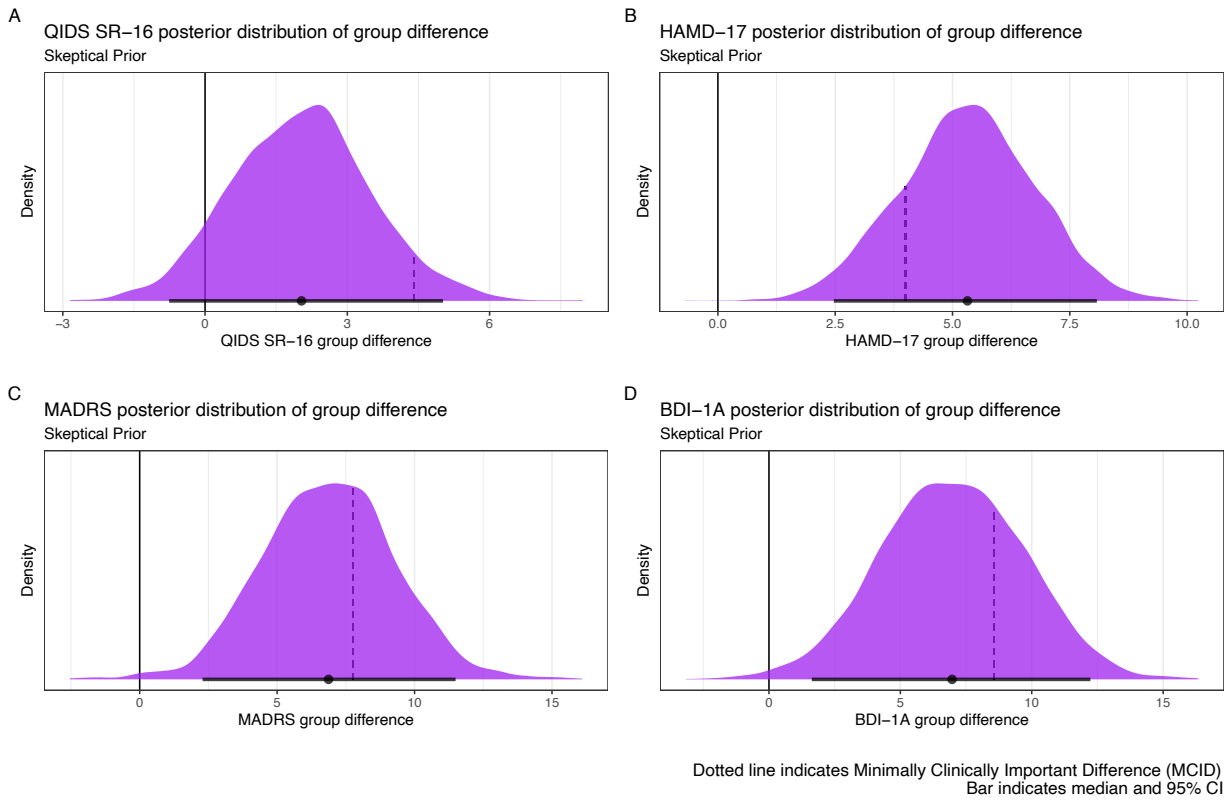
Outcome	Skeptical prior	Flat prior
QIDS SR-16	2.0 [-0.8, 5.0]	2.2 [-0.8, 5.2]
HAMD-17	5.3 [2.6, 8.0],	5.3 [2.3, 8.2]
MADRS	7.0 [2.3, 11.6]	7.2 [2.3, 12.1]
BDI-1A	7.0 [1.6, 12.2]	7.4 [1.8, 12.9]

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 non-inferiority.

Outcome	Any superiority	Clinically meaningful superiority	Non-inferiority
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

Note. Values of BF_{10} in the range 3–10 indicate moderate evidence, values in the range of 10–30 indicate strong evidence, 30–100 very strong, and greater than 100 extremely strong evidence for the experimental hypothesis (H_1) (Quintana and Williams 2018). 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. BF_{10} from 0.5–2 are usually considered to be indeterminate, requiring more evidence. Clinically meaningful superiority refers to a group difference greater than the Minimally Clinically Important Difference (MCID).

Figure 1. Posterior distributions of group difference between psilocybin and escitalopram in the four depression scales used



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

Author statement: SMN: Conceptualization, Formal Analysis, Methodology, Software, Visualization, Writing - Original Draft, and Writing - Review & Editing. **BAB:** Conceptualization, Formal Analysis, Methodology, Visualization, Validation, and Writing - Review & Editing. **DBY:** Conceptualization, and Writing - Review & Editing. **MJS:** Writing - Review & Editing and Conceptualization. **FER:** Writing - Review & Editing. **JMP:** Data curation, Writing - Review & Editing. **BG:** Investigation, Project administration, Writing - Review & Editing. **DE:** Investigation, Writing - Review & Editing. **DJN:** Investigation and Writing - Review & Editing. **RCH:** Investigation, Project Administration, Resources, Supervision, and Writing - Review & Editing.

Funding statement: The original trial was funded by a private donation from the Alexander Mosley Charitable Trust and by the founding partners of Imperial College London's Centre for Psychedelic Research.

Support for SMN and DBY 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 BAB comes from NIMH grant R25MH094612. RCH is the Ralph Metzner Chair of the Psychedelic Division, Neuroscape at University of California San Francisco.

Competing interests: RCH reports receiving consulting fees or stock options from Journey Collab, Entheon Biomedical, Beckley Psytech, Mydecine, Tryp Therapeutics and Maya Health; BG reports receiving consulting fees from Small Pharma Ltd; DE received consulting fees from Field Trip and Mydecine. DJN 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.

Data availability: Data and scripts are available at https://osf.io/vfw7g/

References

Andersson, Gerhard, Hugo Hesser, Andrea Veilord, Linn Svedling, Fredrik Andersson, Owe Sleman, Lena Mauritzson, et al. 2013. "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* 151 (3): 986–94. *https://doi.org/10.1016/j.jad.2013.08.022*.

Bauer, Michael, Liliana Dell'Osso, Siegfried Kasper, William Pitchot, Eva Dencker Vansvik, Jürgen Köhler, Leif Jørgensen, and Stuart A. Montgomery. 2013. "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* 151 (1): 209–19. *https://doi.org/10.1016/j.jad.2013.05.079*.

Bayarri, M Jésus, and James O Berger. 2004. "The Interplay of Bayesian and Frequentist Analysis." *Stat Sci* 19 (1): 58–80.

Bobo, William V., Gabriela C. Angleró, Gregory Jenkins, Daniel K. Hall-Flavin, Richard Weinshilboum, and Joanna M. Biernacka. 2016. "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 Psychopharm Clin* 31 (3): 185–92. *https://doi.org/10.1002/hup.2526*.

Bürkner, Paul-Christian. 2018. "Advanced Bayesian Multilevel Modeling with the R Package Brms." *The R Journal* 10 (1): 395–411. *https://doi.org/10.32614/RJ-2018-017*.

Carhart-Harris, Robin, Bruna Giribaldi, Rosalind Watts, Michelle Baker-Jones, Ashleigh Murphy-Beiner, Roberta Murphy, Jonny Martell, Allan Blemings, David Erritzoe, and David J. Nutt. 2021. "Trial of Psilocybin Versus Escitalopram for Depression." *N Engl J Med* 384 (15): 1402–11. *https://doi.org/10.1056/NEJMoa2032994*.

Combes, Alain, David Hajage, Gilles Capellier, Alexandre Demoule, Sylvain Lavoué, Christophe Guervilly, Daniel Da Silva, et al. 2018. "Extracorporeal Membrane Oxygenation for Severe Acute Respiratory Distress Syndrome." *N Engl J Med* 378 (21): 1965–75.

Furukawa, Toshi A., Tatsuo Akechi, Hideki Azuma, Toru Okuyama, and Teruhiko Higuchi. 2007. "Evidence-Based Guidelines for Interpretation of the Hamilton Rating Scale for Depression." *J Clin Psychopharm* 27 (5): 531–34. *https://doi.org/10.1097/JCP.0b013e31814f30b1*.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. "Why We (Usually) Don't Have to Worry about Multiple Comparisons." *J Res Educ Eff* 5 (2): 189–211.

Gibbons, Mary Beth Connolly, Robert Gallop, Donald Thompson, Debra Luther, Kathryn Crits-Christoph, Julie Jacobs, Seohyun Yin, and Paul Crits-Christoph. 2016. "Comparative Effectiveness of Cognitive and Dynamic Therapies for Major Depressive Disorder in a Community Mental Health Setting: A Randomized Non-Inferiority Trial." *JAMA Psychiatry* 73 (9): 904–11. *https://doi.org/10.1001/jamapsychiatry.2016.1720*.

Goligher, Ewan C., George Tomlinson, David Hajage, Duminda N. Wijeysundera, Eddy Fan, Peter Jüni, Daniel Brodie, Arthur S. Slutsky, and Alain Combes. 2018. "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* 320 (21): 2251. *https://doi.org/10.1001/jama.2018.14276*.

Granholm, Anders, Marie Warrer Munch, Sheila Nainan Myatra, Bharath Kumar Tirupakuzhi Vijayaraghavan, Maria Cronhjort, Rebecka Rubenson Wahlin, Stephan M. Jakob, et al. 2021. "Dexamethasone 12 Mg Versus 6 Mg for Patients with COVID-19 and Severe Hypoxaemia: A Pre-Planned, Secondary Bayesian Analysis of the COVID STEROID 2 Trial." *Intens Care Med*, November. *https://doi.org/10.1007/s00134-021-06573-1*.

Hengartner, Michael P., and Martin Plöderl. 2021. "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*, February. *https://doi.org/10.1136/bmjebm-2020-111600*.

Hernández, Glenn, Gustavo A. Ospina-Tascón, Lucas Petri Damiani, Elisa Estenssoro, Arnaldo Dubin, Javier Hurtado, Gilberto Friedman, et al. 2019. "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* 321 (7): 654–64. *https://doi.org/10.1001/jama.2019.0071*.

Hoekstra, Rink, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. “Robust Misinterpretation of Confidence Intervals.” *Psychon Bull Rev* 21 (5): 1157–64. *https://doi.org/10.3758/s13423-013-0572-3*.

Keysers, Christian, Valeria Gazzola, and Eric-Jan Wagenmakers. 2020. “Using Bayes Factor Hypothesis Testing in Neuroscience to Establish Evidence of Absence.” *Nat Neurosci* 23 (7): 788–99.

Leucht, Stefan, Hein Fennema, Rolf R. Engel, Marion Kaspers-Janssen, Peter Lepping, and Armin Szegedi. 2017. “What Does the MADRS Mean? Equipercentile Linking with the CGI Using a Company Database of Mirtazapine Studies.” *J Affect Disord* 210 (March): 287–93. *https://doi.org/10.1016/j.jad.2016.12.041*.

Leucht, Stefan, Hein Fennema, Rolf Engel, Marion Kaspers–Janssen, Peter Lepping, and Armin Szegedi. 2013. “What Does the HAMD Mean?” *J Affect Disord* 148 (2-3): 243–48. *https://doi.org/10.1016/j.jad.2012.12.001*.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. CRC press.

Mechler, Jakob, Karin Lindqvist, Per Carlbring, Peter Lilliengren, Fredrik Falkenström, Gerhard Andersson, Naira Topooco, et al. 2020. “Internet-Based Psychodynamic Versus Cognitive Behaviour Therapy for Adolescents with Depression: Study Protocol for a Non-Inferiority Randomized Controlled Trial (the ERiCA Study).” *Trials* 21 (1). *https://doi.org/10.1186/s13063-020-04491-z*.

Moerbeek, Mirjam. 2021. “Bayesian Updating: Increasing Sample Size During the Course of a Study.” *Bmc Med Res Methodol* 21 (1): 137. *https://doi.org/10.1186/s12874-021-01334-6*.

Mohr, David C., Emily G. Lattie, Kathryn Noth Tomasino, Mary J. Kwasny, Susan M. Kaiser, Elizabeth L. Gray, Nameyeh Alam, Neil Jordan, and Stephen M. Schueller. 2019. “A Randomized Noninferiority Trial Evaluating Remotely-Delivered Stepped Care for Depression Using Internet Cognitive Behavioral Therapy (CBT) and Telephone CBT.” *Behav Res Ther* 123 (December): 103485. *https://doi.org/10.1016/j.brat.2019.103485*.

Quintana, Daniel S, and Donald R Williams. 2018. “Bayesian Alternatives for Common Null-Hypothesis Significance Tests in Psychiatry: A Non-Technical Guide Using JASP.” *BMC Psychiatry* 18 (1): 1–8.

R Core Team. 2020. “R: A Language and Environment for Statistical Computing.” Manual. Vienna, Austria. *https://www.R-project.org/*.

Rush, A. John, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, et al. 2003. “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 Psychiat* 54 (5): 573–83. *https://doi.org/10.1016/S0006-3223(02)01866-8*.

Russell, Lene, Kis Rønn Uhre, Ann Louise Syraach Lindgaard, Jette Fredlund Degn, Mik Wetterslev, Praleene Sivapalan, Carl Thomas Anthon, et al. 2021. “Effect of 12 Mg Vs 6 Mg 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* 326 (18): 1807–17.

Schönbrodt, Felix D., and Eric-Jan Wagenmakers. 2018. “Bayes Factor Design Analysis: Planning for Compelling Evidence.” *Psychonomic Bulletin & Review* 25 (1): 128–42. *https://doi.org/10.3758/s13423-017-1230-y*.

Schönbrodt, Felix D., Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. “Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences.” *Psychol Methods* 22 (2): 322–39. *https://doi.org/10.1037/met0000061*.

Sjölander, Arvid, and Stijn Vansteelandt. 2019. “Frequentist Versus Bayesian Approaches to Multiple Testing.” *Eur J Epidemiol* 34 (9): 809–21. *https://doi.org/10.1007/s10654-019-00517-2*.

Szegedi, A, R Kohnen, A Dienel, and M Kieser. 2005. “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* 330 (7490): 503. *https://doi.org/10.1136/bmj.38356.655266.82*.

Van Dongen, Stefan. 2006. “Prior Specification in Bayesian Statistics: Three Cautionary Tales.” *J Theor Biol* 242 (1): 90–100.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit. 2012. “An Agenda for Purely Confirmatory Research.” *Perspect Psychol Sci* 7 (6): 632–38. *https://doi.org/10.1177/1745691612463078*.

Wilson, Hilary D. 2007. “Minimum Clinical Important Differences of Health Outcomes in a Chronic Pain Population: Are They Predictive of Poor Outcomes?” *ProQuest Dissertations and Theses*. PhD thesis. *https://www.proquest.com/dissertations-theses/minimum-clinical-important-differences-health/docview/304711319/se-2?accountid=11752*.

Yarnell, Christopher J, Darryl Abrams, Matthew R Baldwin, Daniel Brodie, Eddy Fan, Niall D Ferguson, May Hua, et al. 2021. “Clinical Trials in Critical Care: Can a Bayesian Approach Enhance Clinical and Scientific Decision Making?” *Lancet Respir Med* 9 (2): 207–16. *https://doi.org/10.1016/S2213-2600(20)30471-9*.

Zampieri, Fernando G., Lucas P. Damiani, Jan Bakker, Gustavo A. Ospina-Tascón, Ricardo Castro, Alexandre B. Cavalcanti, and Glenn Hernandez. 2020. “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* 201 (4): 423–29. *https://doi.org/10.1164/rccm.201905-0968OC*.