PEER REVIEW HISTORY

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

ARTICLE DETAILS

TITLE (PROVISIONAL)	What are the Treatment Remission, Response, and Extent of Improvement Rates after up to Four Trials of Antidepressant Therapies in Real-World Depressed Patients? A Reanalysis of the STAR*D Study's Patient-Level Data with Fidelity to the Original Research Protocol
AUTHORS	Pigott, H.; Kim, Thomas; Xu, Colin; Kirsch, Irving; Amsterdam, Jay

VERSION 1 – REVIEW

REVIEWER	Moncrieff, Joanna University College London and North East London mental health trust, Mental Health Sciences
REVIEW RETURNED	05-May-2022

GENERAL COMMENTS	This is really important paper. It presents data on the primary outcome of the influential STAR*D study which has not been published before because the original authors presented different outcomes. The manuscript is clear and well-written, the authors have analysed the data appropriately, following the original protocol as closely as possible and describe the process in more than adequate detail. The STAR*D was a complex study, and I have a couple of suggestions for clarifying the design of the original study and adding some important information into the Results and Discussion: Abstract I suggest the Abstract should specify how many steps or levels of treatment there are in total (at present the Abstract does not make sense without prior knowledge of the STAR-D design) In Methods, Acute Treatment section and Appendix 2 I appreciate that the focus of this paper is not on the comparisons made between types of treatment in the STAR*D study, but on the overall results. However, the design of the study is difficult to understand from the text. I suggest the authors briefly describe Level 1 treatment in the text, then in Levels 2-4 specify that people could chose a number of treatment options involving switching or augmentation of the original antidepressant regime and were then randomised to a treatment option among those selected. The authors should also specify that CBT alone was one of the switch options (I assume).

	Results A lot of critical information is placed in Appendices which are sometimes difficult to access, and some readers will not bother to open them. I suggest the table in Appendix 6 is definitely placed in the main text since this neatly summarises the results of each level and overall, and helps convey the main results to readers plainly and accessibly. I would also consider placing the figures in Appendix 7 and 8 in the main text as these also provide the data for one of the main objectives of this analysis Discussion It would be good to know how the remission rate and HRSD change scores compare with that reported in placebo groups in randomised trials. I know the results of placebo-controlled trials are generally lower than those of comparator trials, nevertheless it would be good to know how the STAR*D result compares with results for people not taking antidepressants. I tried to find this figure from the Cipriani meta-analysis and it is not available in the published material, but just one or two examples, especially any that use the same HRSD criteria, would be helpful. The authors can qualify this by referencing the Rutherford et al studies that show that studies without an inactive control group produce superior outcomes, and also specify that placebo-controlled trials are much shorter, more selective and would therefore be expected to show lower remission rates than a study that lasted a year with a more morbid population I would also be interested to know how this data relates to the Pigott et al (2010) analysis that showed that only 108 participants had a sustained recovery and did not drop out by the end of the study.

REVIEWER	Plöderl, Martin Christian Doppler Clinic, Paracelsus Medical University, Dept. of Clinica Psychology and Dept. for Crisis Intervention and Suicide Prevention, University Clinic for Psychiatry, Psychotherapy, and Psychosomatics
REVIEW RETURNED	25-May-2022

GENERAL COMMENTS	In their RIAT analysis of all levels/steps of the famous STAR*D study, Pigott et al. again revealed discrepancies between study findings published by the original STAR*D study team and the results produced by following the pre-specified protocols. The re-analysis by Pigott et al. is very impressive and important work for at least two reasons. First, in this naturalistic study on antidepressants, rates of remission and response, if calculated per protocol, were substantially smaller than predicted by experts or found in randomized controlled trials (RCTs). Consequently, the frequently used argument that antidepressants, with their poor efficacy in RCTs, work better in actual clinical practice compared to RCTs is not supported by the STAR*D study. Second, the re-analysis by Pigott et al. is another example of biases in research, in this case resulting from protocol-deviations. It is troubling and depressing that this happens even in a large, non-industry funded, pre-registered study authored by many high-ranked academic psychiatrists. It is hard to believe that the original team of investigators simply did not report the main outcome as specified in the protocol, and that the results for response remained unpublished. Furthermore, there are still no results

	posted on clinicaltrial.gov where the STAR*D study was registered (https://clinicaltrials.gov/ct2/show/NCT00021528) and it seems no efforts were made by the NIMH to correct the results (https://www.nimh.nih.gov/funding/clinical-research/practical/stard/allmedicationlevels). Without doubt, the submitted paper by Pigott et al. is an important and necessary corrective. I have several comments which I consider minor. Generally, the paper might be hard to read for someone not familiar with the STAR*D study design, study results published by the original STAR*D team, and already existing critical publications. The structure of the paper could be improved by choosing other headers and placings of text. Throughout the paper, “STAR*D outcome” or “STAR*D results” is used for the results published by the original STAR*D study team (references 1-7, for example, Rush et al. 2006, Am J Psychiatry). However, this is ambiguous, because there are already several publications by Pigott et al. who provided the per-protocol results (“correct” STAR*D results). Both the original and the critical publications are “STAR*D results”. The STAR*D protocol was only available after a Freedom of Information request from the NIMH. Perhaps this deserves more explicit mentioning? I find it disturbing that such a large and expensive study funded by the public is lacking transparency. Is the protocol now available for the public? ABSTRACT The abstract seems to be a mixture of a description of the original STAR*D study and the RIAT-reanalysis and this is a bit confusing. The results paragraph starts with “We reanalyzed the STAR*D dataset with fidelity to the original research…” – IMO, this belongs into the methods part or in the objectives, not into the results part. Perhaps explain in the results-section that the 99 patients who remitted before the study started and the 125 patients who remitted when they initiated the next level should have been excluded from the analysis according to the study protocol. Perhaps consider to delete “actual” in the last sentence of the results paragraph in the abstract. The comparison of the main results with those from a meta-analysis seems to be a main goal of the study but nothing was mentioned in the abstract. INTRODUCTION First paragraph: “most consequential”, “oversized impact” of the STAR*D study – is there proof for these claims? The given

	references (1-7) are only references to the original publications, not about the consequences or impact of these publications. 3rd paragraph, first sentence: perhaps explain that this is a publication about the method of the STAR*D trial. Generally, the source of claims should be made explicit (protocol, publications about the study method or protocol, publications of the STAR*D results by original STAR*D team, publications by critical re-analysis following the protocol). p. 5 “As Pigott et al. document though, the investigators’ assumptions are not true in the real world, since patients who drop out are more likely to be treatment non-responders” Can you give references to this important claim? I have heard from PTSD research that a significant proportion of patients drop out of a study because of feeling better https://www.sciencedirect.com/science/article/pii/S0887618516301463#sec0070 I could not find studies about drop-outs in antidepressants because of feeling better, unfortunately. End of introduction: a brief summary of what has been published so far and what the current study adds would be informative. This is not easy to summarize for readers based on the information in the manuscript. Some of the findings were already published by the authors, but it is definitely informative to have the correct results for all levels/steps of STAR*D in one paper. METHOD “STAR*D investigators state in their level 1 article, ‘our primary analyses classified patients with missing exit HRSD scores as nonremitters a priori’ “. It is suprising that Trivedi et al. speak of an a-priori classification, but that this was not specified in the protocol. There were different a-priori’s, as it seems. Second paragraph (p. 6): “RIAT investigators published our response” in the BMJ. The response, however, was a rapid response to an existing BMJ-publication about RIAT, published by the BMJ not by the RIAT investigators. Or was the response reviewed by the RIAT initiative? Perhaps explain the difference between “steps” and “levels” in the STAR*D study somewhere in the text and by referring to Figure 1. Also consider to change the header “Acute Treatment” and use something like “Steps/Levels of acute treatment”. Otherwise, acute treatment may be conflated with level 1 or step 1 treatment. Furthermore, it is confusing when different expressions are used for “level” (trial, acute phase, initial trial). The STAR*D study has a complex methodology, and consistency in terms makes the paper much easier to read. Was the expression “aggressive medication” originally used by the developers of the STAR*D study?

	p. 7, 3rd line: “11 pharmacological distinct drug/drug combinations”. It was also possible to switch to a single other drug or to CBT. p. 7, 2nd paragraph, first sentence: perhaps mention that QIDS-C < 6 was a considered as clinician rated remission. Also, explain what “follow up” means. Header “research design” – refers to the original STAR*D study, not to the RIAT initiative. Generally, and in the abstract, the paper would be easier to read if there was a clearer distinction between the original STAR*D study plan and the RIAT re-analysis (e.g., changing headers “research design” to “research design of the STAR*D study”, “analytic plan” to “analytic plan of the RAIT reanalysis”. p. 7, last paragraph “First, the protocol is silent regarding patients who entered the study without a ROA administered HRSD score of ≥14. In their level 1 article, STAR*D investigators deemed such patients ineligible for analysis.[1]”. So these patients were not included for analysis of levels 1-4, not just for analysis of level 1? Could it be that some of these patients were getting worse (HRSD ≥ 14) and thus could enter another level of treatment? p. 8, 5th paragraph: “We then compared STAR*D’s outcomes to those found in a meta-analysis of 7,030 patients” – no reference is given. The Rutherford meta-analysis is from 2009 and thus quite outdated and likely more prone to publication bias. Is there an updated meta-analysis? RESULTS It would be helpful to know the number of patients randomized to the next level in Figure 1. Furthermore, it would be helpful to know what “Exit” and “Follow Up” means in a footnote to the Figure. Is Figure 2 necessary? There are only four percentages. “In step 1, these measures of improvement among STAR*D’s patients were approximately half that found in comparator trials,…” Numbers would be informative. DISCUSSION The structure of the discussion section can be improved. For example, the comparison of the STAR*D results with predictions by experts is found under “comparison with other studies” 1st paragraph: references to the “original publication” and “summary article” are missing. CONCLUSION

	The first paragraph seems out of place, because reporting the group differences was no topic in the methods and results section of the paper. Perhaps consider moving it into the discussion section or deleting it. The expression “try-try-try-and-try” again approach seems awkward.

REVIEWER	Ajilore, Olusola University of Illinois College of Medicine at Chicago, Psychiatry
REVIEW RETURNED	10-Sep-2022

GENERAL COMMENTS	The author provide a re-analysis of cumulative remission rates from the landmark STAR*D trial by utilizing the original pre-specified primary outcome - Hamilton Rating Scale for Depression. They found lower rates of response and remission using the HAM-D compared to Quick Inventory of Depression Symptoms - Self-Rated Scale. It's unclear why this study is needed as the first two STAR*D papers cited by the authors indicate that the Hamilton Rating Scale for Depression is the primary outcome and they report it as such in the cited papers. QIDS-SR is always mentioned as a secondary outcome. This is a quote from the results section of the 2006 AJP paper from Trivedi et al, "Remission rates were 28% (HAM-D) and 33% (QIDS-SR). ". This is a quote from the 2006 NEJM paper abstract from Rush et al, "Remission rates as assessed by the HRSD-17 and the QIDS-SR-16, respectively, were 21.3 percent and 25.5 percent for sustained-release bupropion, 17.6 percent and 26.6 percent for sertraline, and 24.8 percent and 25.0 percent for extended-release venlafaxine. QIDS-SR-16 response rates were 26.1 percent for sustained-release bupropion, 26.7 percent for sertraline, and 28.2 percent for extended-release venlafaxine." As written, the manuscript makes it sound like the original investigators behind STAR*D switched the primary outcomes for secondary outcomes which is not the case even in the papers cited by the authors. The authors need to address this as they incorrectly state, "However, despite its investigators numerous publications, neither change in HRSD depressive symptom severity nor HRSD response rates have been reported for STAR*D’s six trials and summary article". It would also interesting to look at whether the Hamilton scores were correlated with QIDS-SR scores to see whether the different in the rates of remission and response could have been due to psychometrics properties of the Hamilton which is not an adequate scale to capture all of the aspects of depression.

VERSION 1 – AUTHOR RESPONSE

Comments from Reviewers

Reviewer: 1
Dr. Joanna Moncrieff
University College London

1. Abstract

I suggest the Abstract should specify how many steps or levels of treatment there are in total (at present the Abstract does not make sense without prior knowledge of the STAR-D design)

Response: We agree and made the following addition to the Abstract's Research Design section: Patients who failed to gain adequate relief from their level 1 trial on the SSRI citalopram could receive up to three additional treatment trials in levels 2-4.

2. In Methods, Acute Treatment section and Appendix 2

I appreciate that the focus of this paper is not on the comparisons made between types of treatment in the STAR*D study, but on the overall results. However, the design of the study is difficult to understand from the text. I suggest the authors briefly describe Level 1 treatment in the text, then in Levels 2-4 specify that people could chose a number of treatment options involving switching or augmentation of the original antidepressant regime and were then randomised to a treatment option among those selected. The authors should also specify that CBT alone was one of the switch options (I assume).

Response: We agree and made several edits that will hopefully help readers more easily understand STAR*D's study's design. These include:

- The first two sentences of the Levels/Steps of Acute Treatment section's second paragraph now reads: "All patients were administered the SSRI citalopram for their level 1 treatment. Each treatment level consisted of 12 weeks of antidepressant therapy, with an additional 2 weeks for patients deemed close to remission." The renaming and first paragraph of this section were done to address Dr. Plöderl's recommendation to clarify for readers STAR*D's distinction between levels and steps of acute treatment.
- We added "Cognitive therapy was available as either a switch or citalopram augmentation option in level 2" as the last sentence of the third paragraph in the Levels/Steps of Acute Treatment section.
- Also, in the third paragraph of this section we state that for levels 2-4, "the treatment options available for randomization involved either switching to a new treatment or augmenting the patient's current treatment."

3. Results

A lot of critical information is placed in Appendices which are sometimes difficult to access, and some readers will not bother to open them. I suggest the table in Appendix 6 is definitely placed in the main text since this neatly summarises the results of each level and overall, and helps convey the main results to readers plainly and accessibly. I would also consider placing the figures in Appendix 7 and 8 in the main text as these also provide the data for one of the main objectives of this analysis.

Response: We agree and made Appendix 6 Table 2, and Appendix 7 and 8 figures 2 and 3. At Dr. Plöderl's recommendation, we made our original Figure 2 Appendix 6 freeing up space for these changes.

4. Discussion

It would be good to know how the remission rate and HRSD change scores compare with that reported in placebo groups in randomised trials. I know the results of placebo-controlled trials are generally lower than those of comparator trials, nevertheless it would be good to know how the STAR*D result compares with results for people not taking antidepressants. I tried to find this figure from the Cipriani meta-analysis and it is not available in the published material, but just one or two examples, especially any that use the same HRSD criteria, would be helpful. The authors can qualify this by referencing the Rutherford et al studies that show that studies without an inactive control group produce superior outcomes, and also specify that placebo-controlled trials are much shorter, more selective and would therefore be expected to show lower remission rates than a study that lasted a year with a more morbid population.

Response: We agree in theory but decided against it for the following reasons:

- Dr. Kirsch only obtained the HRSD change scores from Dr. Rutherford for the comparator trials, not the placebo-controlled trials.
- While the average remission rate in the placebo arms of Rutherford et al's placebo-controlled trials was 39.7% (range 22 to 62%), the bulk of these studies had strict exclusion criteria such that approximately 75% (or more) of STAR*D's "real-world" patients would likely have been excluded due to their extensive medical and/or psychiatric comorbidities (see Wisniewski et al, 2009).
- We know of no placebo-controlled trial that used STAR*D's enrollment methodology of including only patients seeking care (vs recruited) and/or had similarly limited exclusion criteria.
- As we state in the Results and Discussion sections, when comparing Rutherford et al's comparator trials to STAR*D's, we found that STAR*D's level 1 outcomes "were approximately half that found in comparator trials, and improvement grew progressively worse in each subsequent treatment episode." Would the remission rate in the placebo arm with patients like STAR*D's have had a similar 50% reduction in effectiveness…more/less? No one knows.

Given our concerns, we assessed that there would need to be so many qualifiers for an estimate of what the placebo rate might be in patient's similar to STAR*D's, that it would distract from our paper and be ripe for misinterpretation. Instead, we added the sentence, "As there was neither a placebo nor waitlist control group during any phase of the STAR*D study, it is impossible to know if even the meager results that were observed were due to the pharmacologic effects of the prescribed medications, placebo effects, or the mere passage of time."

We also added a sentence in the second paragraph of the Comparison with other studies section highlighting antidepressant's suspect efficacy with reference to two recent meta-analyses published in BMJ Open and BMJ.

5. Discussion

I would also be interested to know how this data relates to the Pigott et al (2010) analysis that showed that only 108 participants had a sustained recovery and did not drop out by the end of the study.

Response: Great question that at this time we do not have the answer. Instead, we plan to analyze and report the follow-up data in subsequent efforts tied to each compared treatment. We made this decision for the following reasons:

- As both Drs. Moncrieff and Plöderl noted, STAR*D was a very complex and difficult to follow study. It has been a challenge to present our reanalysis of the up to four acute care treatments in an easily understood format for readers. This, combined with word count limitations, made us restrict the scope of this paper to reanalyzing the levels/steps 1-4 data.

- During follow-up, STAR*D clinicians stopped using their measurement-based system of care in which during each monthly clinic visit the QIDS-SR, IDS-C and measures of side-effects and medication adherence were administered and recorded in the patient-level database. Consequently, it will be near impossible to determine definitively when/if a patient dropped out during follow-up.
- The Pigott et al report of only 108 patients having a sustained recovery and not dropping out by the end of follow up, was based on the table above the survival analysis figure in figure 3 of STAR*D's summary article. Pigott confirmed this understanding in an email exchange with Dr. Wisniewski, STAR*D's chief biostatistician, on July 1, 2008.
- During follow-up, as per protocol, the telephonic IVR-version of the QIDS was administered monthly and designed to be used as a secondary outcome measure with the blinded HRSD administered every three months as the primary measure. STAR*D used the QIDS-IVR to report their follow-up survival analysis outcomes. The data for each cell in the table above the survival analysis figure is the number of patients who completed at least 1 (out of the scheduled 3) QIDS-IVR assessments during that quarter and did not score as having relapsed. It turns outs that the QIDS-IVR had a low completion rate such that there were likely some patients who did not complete even 1 QIDS-IVR during a quarter yet remained in follow-up and may have taken the quarterly HRSD.

For these reasons, we plan to report the follow-up HRSD data in our upcoming efforts. We agree though that it would be helpful for readers to be aware of the durability of treatment effects issue. For this reason, we added a second paragraph to the Principal findings and comparison with original STAR*D publication section. It states, "Our reanalysis did not assess the durability of treatment effects during the 12-month follow up phase. In their summary article though, STAR*D authors reported an overall relapse rate of 46.1% for the 1,729 patients for whom they had at least one assessment (of up to 12 possible) during follow up using a telephonic-administered version of the QIDS [7, table 5] whereas Pigott et al found a far lower sustained recovery rate when incorporating patient dropout in the analysis.[12]"

Reviewer 2
Dr. Martin Plöderl
Christian Doppler Clinic
Paracelsus Medical University

1. Introduction

The STAR*D protocol was only available after a Freedom of Information request from the NIMH. Perhaps this deserves more explicit mentioning? I find it disturbing that such a large and expensive study funded by the public is lacking transparency. Is the protocol now available for the public?

Response: We agree and changed the last sentence of the Introduction's third paragraph to now read, "The protocol, obtained via a Freedom of Information Act request by Pigott, [10] states:" We just completed a Google search and were not able find STAR*D's research protocol online.

2. Abstract

The abstract seems to be a mixture of a description of the original STAR*D study and the RIAT-reanalysis and this is a bit confusing. The results paragraph starts with "We reanalyzed the STAR*D dataset with fidelity to the original research…" – IMO, this belongs into the methods part or in the objectives, not into the results part.

Perhaps explain in the results-section that the 99 patients who remitted before the study started and the 125 patients who remitted when they initiated the next level should have been excluded from the analysis according to the study protocol.

Perhaps consider to delete “actual” in the last sentence of the results paragraph in the abstract.

Response: We agree. The Objective now reads, “Reanalyze the patient-level dataset of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study with fidelity to the original research protocol and related publications.” As for the 99 and 125 patients scoring as remitted at study outset or prior to starting their next-level treatment, Results section states, “These patients should have been excluded from data analysis.” Actual was deleted from the last sentence in the Results section.

3. Abstract

The comparison of the main results with those from a meta-analysis seems to be a main goal of the study but nothing was mentioned in the abstract.

Response: We agree but we’d already exceeded BMJ Open’s word count for the Abstract. If we added this, we’d also want to add the comparison with the protocol’s expert predictions of outcome. We currently reference both in Strengths & Limitations section.

4. Introduction

First paragraph: “most consequential”, “oversized impact” of the STAR*D study – is there proof for these claims? The given references (1-7) are only references to the original publications, not about the consequences or impact of these publications.

Response: We partially agree and deleted “most consequential” but left “oversized impact.” Given its two NEJM publications, 100+ additional publications by STAR*D authors in high impact psychiatric journals, innumerable citations by other researchers, and prominent role STAR*D’s claims play in many depression treatment guidelines, it has had an oversized impact compared to any other antidepressant study that we are aware of. The sentence now reads, “The 35-million US dollar Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study is the largest prospective antidepressant trial ever conducted, with over 100 journal articles published by study investigators and innumerable citations of STAR*D’s findings by other researchers giving it an oversized impact on the treatment of depression world-wide.[1-7]”

5. Introduction

3rd paragraph, first sentence: perhaps explain that this is a publication about the method of the STAR*D trial. Generally, the source of claims should be made explicit (protocol, publications about the study method or protocol, publications of the STAR*D results by original STAR*D team, publications by critical re-analysis following the protocol).

Response: We agree. To this end, we’ve modified our original sentence to read: “Our reanalysis of STAR*D examines key methodological deviations from its research protocol and related publications, and these deviations’ impact on its investigators report of outcomes.” Then followed by, “In STAR*D’s Rationale and Research Design article, and repeated in the level 1-4 published study outcomes, STAR*D investigators stated, “the primary outcome is depressive symptom severity, measured by the 17-item Hamilton Rating Scale for Depression (HRSD).”[8, p. 120]. Elsewhere in the paper we’ve sought to be explicit regarding the source of claims.

6. Introduction

p. 5 “As Pigott et al. document though, the investigators’ assumptions are not true in the real world, since patients who drop out are more likely to be treatment non-responders”
Can you give references to this important claim? I have heard from PTSD research that a significant proportion of patients drop out of a study because of feeling better
https://www.sciencedirect.com/science/article/pii/S0887618516301463#sec0070.
I could not find studies about drop-outs in antidepressants because of feeling better, unfortunately.

Response: We agreed, until we reread the Barbui et al’s paroxetine meta-analysis. It found that paroxetine subjects were significantly more likely to drop out of RCTs due to adverse treatment side effects and emergent suicidality than placebo subjects. We’ve changed this sentence to read, “As Pigott et al. document though, the investigators’ assumptions are not true in the real world since more patients dropped out than remitted in each STAR*D treatment level,[11] and patients who drop out are more likely to have had adverse treatment side effects and/or emergent suicidality. [12]”

7. Introduction

End of introduction: a brief summary of what has been published so far and what the current study adds would be informative. This is not easy to summarize for readers based on the information in the manuscript. Some of the findings were already published by the authors, but it is definitely informative to have the correct results for all levels/steps of STAR*D in one paper.

Response: We partially agree and focused on highlighting what the current study adds to the Pigott et al 2010 article versus summarizing Pigott’s six prior STAR*D publications. We added the sentence, “This effort builds on Pigott et al’s 2010 article that focused on deconstructing STAR*D investigators’ levels 1-4 and summary articles [1-7] by analyzing STAR*D’s patient-level dataset obtained from NIMH in 2019 with fidelity to the original research protocol and inclusion in data analysis criteria.[11]”

8. Method

Second paragraph (p. 6): “RIAT investigators published our response” in the BMJ.
The response, however, was a rapid response to an existing BMJ-publication about RIAT, published by the BMJ not by the RIAT investigators. Or was the response reviewed by the RIAT initiative?

Response: Our Call to Action was first reviewed by RIAT investigators and then essentially greenlighted with BMJ.

9. Method

Perhaps explain the difference between “steps” and “levels” in the STAR*D study somewhere in the text and by referring to Figure 1. Also consider to change the header “Acute Treatment” and use something like “Steps/Levels of acute treatment”. Otherwise, acute treatment may be conflated with level 1 or step 1 treatment. Furthermore, it is confusing when different expressions are used for “level” (trial, acute phase, initial trial). The STAR*D study has a complex methodology, and consistency in terms makes the paper much easier to read.

Response: We agree. We renamed this section, “Levels/Steps of Acute Treatment”. Starting with the second sentence of the first paragraph it now reads, “Appendix 2 describes the antidepressant therapies available in levels 1-4 while steps refer to the numeric order of treatments. As seen in figure 1, treatment steps 1 and 2 correspond to level 1 and 2 treatments. Similarly, for most patients their level 3 and 4 treatments correspond to treatment steps 3 and 4. For level/step 2 patients though who failed to respond adequately to cognitive therapy alone or combined with citalopram and chose to

continue in the study, their third treatment step was designated level 2A and they were randomized to one of two level 2 switch medications. For these patients, their level 2A treatment was their third treatment step. For level 2A patients who did not adequately benefit from this medication trial and chose to continue in the study, they entered a fourth treatment step consisting of level 3 treatments." Used MS Word's search function to better ensure consistency of terms.

10. Method

Was the expression "aggressive medication" originally used by the developers of the STAR*D study?

Response: No. STAR*D investigators wrote their measurement-based system of care was used to ensure every patient received "a fully adequate dose for a sufficient time." The sentence now reads, "This system was used to guide medication management of a fully adequate dose for a sufficient time to "ensure that the likelihood of achieving remission was maximized and that those who did not reach remission were truly resistant to the medication."

11. Method

p. 7, 3rd line: "11 pharmacological distinct drug/drug combinations". It was also possible to switch to a single other drug or to CBT.

Response: We agree and added the sentence, "Cognitive therapy was also available as either a switch or citalopram augmentation option in level 2."

12. Method

p. 7, 2nd paragraph, first sentence: perhaps mention that QIDS-C < 6 was a considered as clinician rated remission. Also, explain what "follow up" means.

Response: We agree and created a heading for this paragraph titled, "STAR*D Follow Up Phase". The first two sentences of this section now reads, "In each treatment trial for levels 1-4, patients who scored <6 on their last QIDS-Clinician version (QIDS-C) were considered clinician-rated remissions and encouraged to enter the 12-month follow-up phase. During follow-up, patients continued their "previously effective acute treatment medication(s) at the doses used in acute treatment but that any psychotherapy, medication, or medication dose change could be used."[7, p.1908]

13. Method

Header "research design" – refers to the original STAR*D study, not to the RIAT initiative. Generally, and in the abstract, the paper would be easier to read if there was a clearer distinction between the original STAR*D study plan and the RIAT re-analysis (e.g., changing headers "research design" to "research design of the STAR*D study", "analytic plan" to "analytic plan of the RAIT reanalysis".

Response: We agree and changed the headers to "Research Design of the STAR*D Study" and "Analytic Plan of the RIAT Reanalysis."

14. Method

p. 7, last paragraph "First, the protocol is silent regarding patients who entered the study without a ROA administered HRSD score of ≥14. In their level 1 article, STAR*D investigators deemed such patients ineligible for analysis.[1]". So these patients were not included for analysis of levels 1-4, not

just for analysis of level 1? Could it be that some of these patients were getting worse (HRSD ≥ 14) and thus could enter another level of treatment?

Response: We agree and Dr. Plöderl is correct. We edited the sentences to now read, “In their level 1 article, STAR*D investigators deemed the 931 such patients who lacked this marker of depression severity ineligible for inclusion in data analysis.[1] We do the same and extend this exclusion for such patients who continued on to levels 2-4 because their extent of depression severity at study outset is not known.” We did not assess if some of the 931 patients who lacked a HRSD score of ≥14 at study outset, yet continued to level 2, scored ≥14 on the HRSD at exit from level 1.

15. Method

p. 8, 5th paragraph: “We then compared STAR*D’s outcomes to those found in a meta-analysis of 7,030 patients” – no reference is given. The Rutherford meta-analysis is from 2009 and thus quite outdated and likely more prone to publication bias. Is there an updated meta-analysis?

Response: We added the reference and are not aware of an updated meta-analysis of comparator trials.

16. Results

It would be helpful to know the number of patients randomized to the next level in Figure 1. Furthermore, it would be helpful to know what “Exit” and “Follow Up” means in a footnote to the Figure.

Response: We agree and added the number of patients randomized to the next-level treatment and defined study exit and follow up as footnotes to the figure.

17. Results

Is Figure 2 necessary? There are only four percentages.

Response: We agree, it is not necessary. We moved this figure to the Appendix.

18. Results

“In step 1, these measures of improvement among STAR*D’s patients were approximately half that found in comparator trials,…” Numbers would be informative.

Response: We partially agree. In response to Dr. Moncrief, we made Appendix 7 and 8 which visually presents this data, figures 2 and 3. This should make it easy for readers to visually confirm this statement (since the figures would now be immediately accessible in the PDF file) as well as having the numbers for comparing comparator trials to STAR*D in terms of remission, response, and HRSD mean change for steps 1-4.

19. Discussion

1st paragraph: references to the “original publication” and “summary article” are missing.

Response: We agree and references have been added.

20. Discussion

The structure of the discussion section can be improved. For example, the comparison of the STAR*D results with predictions by experts is found under “comparison with other studies.”

Response: We partially agree. We moved the last paragraph of the “Comparison to other studies” to the second paragraph of this section. We kept the expert predictions in this section since as highlighted, “The underlying assumptions of these estimates come largely by inferences from results of published RCTs.” In other words, STAR*D’s results are being compared to these experts understanding of the published RCT literature up to 1999.

We conclude this section with “This discrepancy further highlights the relative ineffectiveness of antidepressants in real-world depressed patients, compared to those reported in conventional studies. Consequently, the claim that antidepressants, with their suspect efficacy [35,36], work better in real-world clinical practice is not supported by the STAR*D study when its patient-level data are analyzed as per protocol.” The last sentence echoes Dr. Plöderl’s summation of the implications of our reanalysis in the first paragraph of his review and we believe it is an excellent summation and segue to the Conclusion.

21. Conclusion

The first paragraph seems out of place, because reporting the group differences was no topic in the methods and results section of the paper. Perhaps consider moving it into the discussion section or deleting it.

Response: We agree and deleted this paragraph and replaced it with, “Bias in the clinical literature is commonly associated with industry-funded RCTs, not publicly funded ones.[36] Our RIAT reanalysis though documents numerous scientific errors in this NIMH-funded study. These errors inflated STAR*D investigators’ report of positive outcomes, taking a failed study and portraying it as positive.”

We share Dr. Plöderl’s concerns stated in the first paragraph of his review, regarding undisclosed protocol-deviations in publicly-funded research and that NIMH has done nothing to correct the results. We hope this change captures that sentiment. Also, in the Introduction’s fifth paragraph, we added a quote from NIMH’s Director Dr. Thomas Insel, claiming a “roughly 70%” STAR*D remission rate. When the NIMH Director makes such a published claim back in 2009, it easy to see how this became accepted clinical wisdom.

22. Conclusion

The expression “try-try-try-and-try” again approach seems awkward.

Response: We agree and deleted it. The last paragraph now reads, “The STAR*D summary article’s claim of a 67% cumulative remission rate was published in 2006. If STAR*D’s outcomes had been reported as prespecified, its measurement-based treat-to-remission model of care would likely have faced much stronger criticism 16 years ago and fueled a more vigorous search for evidence-based treatment alternatives.”

Reviewer: 3
Dr. Olusola Ajilore
University of Illinois College of Medicine at Chicago

1. It's unclear why this study is needed as the first two STAR*D papers cited by the authors indicate that the Hamilton Rating Scale for Depression is the primary outcome and they report it as

such in the cited papers. QIDS-SR is always mentioned as a secondary outcome. This is a quote from the results section of the 2006 AJP paper from Trivedi et al, "Remission rates were 28% (HAM-D) and 33% (QIDS-SR)."

This is a quote from the 2006 NEJM paper abstract from Rush et al, "Remission rates as assessed by the HRSD-17 and the QIDS-SR-16, respectively, were 21.3 percent and 25.5 percent for sustained-release bupropion, 17.6 percent and 26.6 percent for sertraline, and 24.8 percent and 25.0 percent for extended-release venlafaxine. QIDS-SR-16 response rates were 26.1 percent for sustained-release bupropion, 26.7 percent for sertraline, and 28.2 percent for extended-release venlafaxine."

Response: We respectfully disagree. Several points:

- First, we never claim that STAR*D's protocol-specified primary measure, the HRSD/HAM-D, was not used to report remission rates in STAR*D's levels 1-4 articles published in AJP and NEJM. We apologize to Dr. Ajilore if this was not made sufficiently clear to him and therefore likely other readers. We therefore edited the first point in the Introduction section of our published STAR*D criticisms to read, "While STAR*D investigators used the used HRSD to report remission rates in their levels 1-4 articles,[1-6] the QIDS-SR was used as the sole measure to report remission, response, and extent of improvement rates in their summary article.[7]"
- Second, Dr. Ajilore never comments on the fact that as we stated in the Introduction's paragraph 3, "STAR*D's research protocol specifically excluded all clinic-administered assessments, such as the QIDS-SR, from use as research outcome measures since they were not blinded and instead, used to guide patient care" and then we quoted the protocol. We'd welcome to hear Dr. Ajilore's explanation why it was scientifically acceptable for STAR*D investigators, without disclosure, to use a measure banned by the protocol to report remission rates as a secondary outcome, and sole measure to report response rates, in the AJP and NEJM articles he cites.
- Third, Dr. Ajilore also never comments on the fact that as we stated in the Abstract, Introduction and Methods, STAR*D (without disclosure) failed to adhere to their inclusion in data analysis criteria by including 99 patients who scored as remitted on the HRSD at study outset as well as 125 who scored as remitted when initiating their next-level treatment. Again, we'd welcome to hear Dr. Ajilore's explanation why this was scientifically acceptable on STAR*D investigators' part.

2. As written, the manuscript makes it sound like the original investigators behind STAR*D switched the primary outcomes for secondary outcomes which is not the case even in the papers cited by the authors. The authors need to address this as they incorrectly state, "However, despite its investigators numerous publications, neither change in HRSD depressive symptom severity nor HRSD response rates have been reported for STAR*D's six trials and summary article".

Response: We respectfully disagree. Neither in the papers cited by Dr. Ajilore, nor in any other STAR*D publication, do its investigators report change in HRSD depressive symptom severity nor HRSD response rates. If Dr. Ajilore is aware of such a STAR*D publication, we'd welcome him to cite it for us to then incorporate it into our resubmission.

3. It would also interesting to look at whether the Hamilton scores were correlated with QIDS-SR scores to see whether the different in the rates of remission and response could have been due to psychometrics properties of the Hamilton which is not an adequate scale to capture all of the aspects of depression.

Response: We respectfully disagree. Examining the correlation between the protocol-specified and blindly-administered HRSD, and the non-blinded QIDS-SR that the protocol disallowed from being used as a research measure, is beyond the scope of our reanalysis. Similarly, assessing the

adequacy of the HRSD and QIDS-SR as measures of depression is beyond the scope of our BMJ Open submission.

VERSION 2 – REVIEW

REVIEWER	Moncrieff, Joanna University College London and North East London mental health trust, Mental Health Sciences
REVIEW RETURNED	23-Nov-2022

GENERAL COMMENTS	The authors have done a good job of clarifying the structure of the STAR-D study. I have a few further suggestions to improve the presentation of the paper. This is such an important paper, it is better to keep the language neutral. The results stand on their own, and readers can make their own judgements about whether the study fulfilled the hopes of its authors. Abstract: In the conclusion, suggest that 'half than' is changed to 'half of' Introduction: In the 3rd paragraph on P 4, it is a bit confusing to focus on the secondary outcome of response and change in symptoms. It does not make clear how the primary outcome was also not clearly reported in the initial reports of the STAR-D study. Therefore I suggest that a sentence is added before the sentence that starts 'However, despite is investigators..' to the effect of 'Remission as defined by the HRSD was not presented in the summary paper, and only presented as a subsidiary analysis in other papers' (if the latter point is not correct then just the first part of the sentence applies). For the same reasons (clarifying how the primary outcome was presented first), in the first bullet point made on P 5 I suggest that the sentence 'While the STAR-D investigators used the HRSD to report remission rates..' is moved to the beginning of this point. This clarifies that the outcome was reported, but not in the summary paper. Results I suggests that a column for cumulative remission rate is added to Figures 2 and 3. Discussion I suggest the phrase 'suspect efficacy' is either deleted or changed to 'whose efficacy has been questioned'. This is to ensure the paper is not dismissed by those who are convinced that antidepressants have been shown to be efficacious. I would also not use the Stone meta-analysis to support this point, as many people view it as evidence of efficacy. Conclusion I would not describe the STAR-D as a 'failed study'. The study produced poor outcomes and many more people dropped out than estimated, but there were no clear criteria of success or failure by which it can be judged. Also, it sounds a bit biased and might lead readers to dismiss the paper.

REVIEWER	Plöderl, Martin Christian Doppler Clinic, Paracelsus Medical University, Dept. of Clinica Psychology and Dept. for Crisis Intervention and Suicide

	Prevention, University Clinic for Psychiatry, Psychotherapy, and Psychosomatics
REVIEW RETURNED	26-Nov-2022

GENERAL COMMENTS	The authors adequately addressed the concerns of us reviewers and I have only minor suggestions without the need for another round of review. The paper and the complex design of STAR*D is now easier to follow. I want to express my respect for the authors for the massive effort to do this important re-analysis. There is the widespread believe that two-thirds improve in “real world” antidepressant treatment, a belief resulting from the misleading an incorrect analysis from the original STAR*D investigators, deviating from the original research protocol. The study by Pigott et al. will hopefully help to correct this belief and also point out that even non-industry trials can be affected by substantial biases. One important novel characteristic of this paper is that the re-analysis was based on patient level data. I know that the title is already very long, but if the patient level data is what makes the study special, then it should be added in the title. Abstract, Participants: “...seeking care (versus recruited)…” I guess Pigott et al. refer to the difference of the STAR*D study from typical clinical trials, but this may not be obvious for all readers. Perhaps make explicit. Furthermore, if this is mentioned in the abstract, than it should also be mentioned in the introduction. Abstract, Main Outcome Measures: If word-count allows, perhaps make explicit that this was remission and response defined as in the original protocol (for example, “According to STAR*D’s protocol,…”). If I understood correctly, the QIDS-SR is a self-report questionnaire handed out at meetings in the clinic (“clinic-administered” may be a bit confusing here). Thus, I wonder if blinding is really an issue, as patients fill out these questionnaires on their own. However, it seems that the original STAR*D investigators saw a concern here, and independent if (un)blinding is an issue or not for the QUID-SR, using it instead of the HRSD is a deviation from the protocol. Introduction, p. 5: “In their summary article, STAR*D investigators used the QIDS-SR as the sole measure to report remission, response, and extent of symptom improvement.” This contrasts with what reviewer Dr. Ajilore pointed out when he quoted the Rush et al. 2006 paper: “Remission rates as assessed by the HRSD-17 and the QIDS-SR-16, respectively, were 21.3 percent and 25.5 percent for sustained-release bupropion, 17.6 percent and 26.6 percent for sertraline, and 24.8 percent and 25.0 percent for extended-release venlafaxine.” So, if I understood correctly, the deviation from the protocol was for only for response and quantitative symptom improvement, where the QIDS-SR instead of the HRSD was resported. Introduction, p. 5:

	“patients who drop out are more likely to have had adverse treatment side effects and/or emergent suicidality. [12]” Perhaps make explicit that this is from randomized placebo-controlled trials. Perhaps reconsider the expression “Oversized Impact” in the introduction. Typo: Introduction, p. 5, line 45: “ used the used HRSD “ Discussion: “As there was neither a placebo nor waitlist control group during any phase of the STAR*D study, it is impossible to know if even the meager results that were observed were due to the pharmacologic effects of the prescribed medications, placebo effects, or the mere passage of time.” I guess all three causes contribute to symptom reductions and they do not exclude each other. Discussion, p. 14, first line: The claim that antidepressants work better in real practice is very common in discussions, but it would be good to reference to such claims in the literature. Furthermore, the recent PANDA study (Lewis et al., 2019) also found very small and non-significant findings for treatment with an antidepressant (sertraline) in its “real-world” design and may be worth mentioning in the discussion as confirming the finding from the STAR*D study. Lewis et al. (2019). The clinical effectiveness of sertraline in primary care and the role of depression severity and duration (PANDA): A pragmatic, double-blind, placebo-controlled randomised trial. The Lancet Psychiatry, 6(11), 903–914. https://doi.org/10.1016/S2215-0366(19)30366-9 I apologize for criticizing the term “aggressive treatment” – it really seems to be common. Pigott et al. correctly point out the important issue that the myth of ~70% remitters have became common clinical knowledge. And it seems there is no felt need for correction. A Google search with “STAR*D” immediately leads to the NIMH information about the STAR*D study, which still has the wrong and misleading finding: “In conclusion, about half of participants in the STAR*D study became symptom-free after two treatment levels. Over the course of all four treatment levels, almost 70 percent of those who did not withdraw from the study became symptom-free” https://www.nimh.nih.gov/funding/clinical-research/practical/stard/allmedicationlevels The myth is also repeated in a recent NYT article: ““If you look at the STAR*D, better than 60 percent of those patients actually had a very good response after going through those various levels of treatment,” said Dr. Gerard Sanacora, a professor of psychiatry at the Yale School of Medicine.“

	https://www.nytimes.com/2022/11/08/well/mind/antidepressants-effects-alternatives.html?smid=nytcore-ios-share&referringSource=articleShare I guess there are myriads of similar repetitions of this myth, highlighting the importance of publishing the re-analysis by Pigott et al.

REVIEWER	Olivier, Jake University of New South Wales, School of Mathematics and Statistics
REVIEW RETURNED	10-Mar-2023

GENERAL COMMENTS	I have conducted a statistical review of the submission. There are some content issues that I am unsure of, but I have not commented on them as content-specialist reviewers have provided reviews. No analysis plan has been provided in the submission. How was the STAR*D data analysed by the original researchers? Were those analyses reasonable or adequate? Can the current authors provide evidence their analytic plan is better? These are important questions and their answers are not clear. Last observation carried forward (LOCF) is widely considered a poor approach to dealing with missing data. Here is one reference but there are many criticisms of this method. It is also unclear to me what the current authors did with participants who exited without an HRSD score. But, certainly using LOCF would not be a valid approach after treatment was given. Lastly, was loss to follow up related to response? From what I understand, the STAR*D research would be roughly correct if they are not related, but would be problematic otherwise. https://www.tandfonline.com/doi/full/10.1080/10543400903105406 Abstract: Please provide remission rate values instead of unquantified statements. Strengths/Limitations: Please provide brief results of comparator trials. It is not clear whether your results are similar or not to these trials. I am not sure I am comfortable with emotive language like "oversized impact" or "unrealistic provisos" or "taking a failed study" without there being evidence to support such statements. The research by the STAR*D authors may have had an appropriate impact on the discipline or maybe not. That needs to be demonstrated and not assumed to be true or false.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Dr. Joanna Moncrieff
University College London

The authors have done a good job of clarifying the structure of the STAR-D study. I have a few further suggestions to improve the presentation of the paper. This is such an important paper, it is better to

keep the language neutral. The results stand on their own, and readers can make their own judgements about whether the study fulfilled the hopes of its authors.

Response: We agree and made edits to keep language neutral.

Abstract:
In the conclusion, suggest that 'half than' is changed to 'half of'

Response: We agree and made said change.

Introduction:
In the 3rd paragraph on P 4, it is a bit confusing to focus on the secondary outcome of response and change in symptoms. It does not make clear how the primary outcome was also not clearly reported in the initial reports of the STAR-D study. Therefore I suggest that a sentence is added before the sentence that starts 'However, despite is investigators..' to the effect of 'Remission as defined by the HRSD was not presented in the summary paper, and only presented as a subsidiary analysis in other papers' (if the latter point is not correct then just the first part of the sentence applies).

Response: We agree and added the sentence, "Remission as defined by the HRSD was not presented in the summary paper."

For the same reasons (clarifying how the primary outcome was presented first), in the first bullet point made on P 5 I suggest that the sentence 'While the STAR-D investigators used the HRSD to report remission rates..' is moved to the beginning of this point. This clarifies that the outcome was reported, but not in the summary paper.

Response: We agree and made said change.

Results
I suggest that a column for cumulative remission rate is added to Figures 2 and 3.

Response: We respectfully disagree. We tried to do it with Figure 2 but it looked cluttered and potentially confusing for readers.

Discussion
I suggest the phrase 'suspect efficacy' is either deleted or changed to 'whose efficacy has been questioned'. This is to ensure the paper is not dismissed by those who are convinced that antidepressants have been shown to be efficacious. I would also not use the Stone meta-analysis to support this point, as many people view it as evidence of efficacy.

Response: We agree. After taking into consideration Dr. Plöderl's request for references to claims in the literature that antidepressants work better in real-world clinical practice, we decided to delete the entire sentence and go straight to the Conclusion. On reflection, we don't think the sentence adds much and could be a point of contention.

Conclusion
I would not describe the STAR-D as a 'failed study'. The study produced poor outcomes and many more people dropped out than estimated, but there were no clear criteria of success or failure by which it can be judged. Also, it sounds a bit biased and might lead readers to dismiss the paper.

Response: We agree and made said change.

Reviewer 2
Dr. Martin Plöderl
Christian Doppler Clinic
Paracelsus Medical University

Comments to the Author:
The authors adequately addressed the concerns of us reviewers and I have only minor suggestions without the need for another round of review. The paper and the complex design of STAR*D is now easier to follow. I want to express my respect for the authors for the massive effort to do this important re-analysis. There is the widespread believe that two-thirds improve in “real world” antidepressant treatment, a belief resulting from the misleading an incorrect analysis from the original STAR*D investigators, deviating from the original research protocol. The study by Pigott et al. will hopefully help to correct this belief and also point out that even non-industry trials can be affected by substantial biases.

One important novel characteristic of this paper is that the re-analysis was based on patient level data. I know that the title is already very long, but if the patient level data is what makes the study special, then it should be added in the title.

Response: We agree and made said change.

Abstract, Participants: “...seeking care (versus recruited)…”
I guess Pigott et al. refer to the difference of the STAR*D study from typical clinical trials, but this may not be obvious for all readers. Perhaps make explicit. Furthermore, if this is mentioned in the abstract, than it should also be mentioned in the introduction.

Response: We agree and made said change in both the abstract and introduction..

Abstract, Main Outcome Measures: If word-count allows, perhaps make explicit that this was remission and response defined as in the original protocol (for example, “According to STAR*D’s protocol,…”).

Response: We agree and made said change.

If I understood correctly, the QIDS-SR is a self-report questionnaire handed out at meetings in the clinic (“clinic-administered” may be a bit confusing here). Thus, I wonder if blinding is really an issue, as patients fill out these questionnaires on their own. However, it seems that the original STAR*D investigators saw a concern here, and independent if (un)blinding is an issue or not for the QUID-SR, using it instead of the HRSD is a deviation from the protocol.

Response: It is important to know the procedures for each clinic visit to understand why STAR*D investigators emphasized in the protocol that “Research outcomes assessments are not collected at the clinic visits. They are not collected by either clinicians or Clinical Research Coordinators.” At each clinic visit, prior to being seen by the CRC the patient completed (Clinical Procedures Manual, page 75):

- The QIDS-SR,
- The Frequency & Intensity of Side Effect Rating,
- The Global Rating for Side Effect Burden,
- The Medication Adherence Questionnaire, and
- The Patient Rated Inventory of Side Effects

The CRC reviewed these measures and then meet with the patient to administer the clinician-version of the QIDS, discuss any symptoms and side effects that the patient may be experiencing and administer that session's Patient Education material. Patient education involved teaching the "mechanism of action" for patients' current antidepressant and educating patients that "depression is a disease, like diabetes or high blood pressure" and "can be treated as effectively as other illnesses," etc. We think the STAR*D investigators recognized the potential for these bi-weekly procedures to overtime bias how some patients answered the QIDS-SR and therefore excluded all clinic-administered measures from use as research outcome assessments.

Introduction, p. 5:
"In their summary article, STAR*D investigators used the QIDS-SR as the sole measure to report remission, response, and extent of symptom improvement." This contrasts with what reviewer Dr. Ajilore pointed out when he quoted the Rush et al. 2006 paper: "Remission rates as assessed by the HRSD-17 and the QIDS-SR-16, respectively, were 21.3 percent and 25.5 percent for sustained-release bupropion, 17.6 percent and 26.6 percent for sertraline, and 24.8 percent and 25.0 percent for extended-release venlafaxine." So, if I understood correctly, the deviation from the protocol was for only for response and quantitative symptom improvement, where the QIDS-SR instead of the HRSD was reported.

Response: In regards to the QIDS-SR, the protocol deviations were using it in the level 1-4 articles to report remission rates as a secondary outcome and sole measure of response rates and then using it as the sole measure to report remission, response and extent of symptom improvement in the Summary article.

Introduction, p. 5:
"patients who drop out are more likely to have had adverse treatment side effects and/or emergent suicidality. [12]" Perhaps make explicit that this is from randomized placebo-controlled trials.

Response: We agree and made said change.

Perhaps reconsider the expression "Oversized Impact" in the introduction.

Response: We agree and made said change.

Typo: Introduction, p. 5, line 45: " used the used HRSD "

Response: We agree and made said change.

Discussion:
"As there was neither a placebo nor waitlist control group during any phase of the STAR*D study, it is impossible to know if even the meager results that were observed were due to the pharmacologic effects of the prescribed medications, placebo effects, or the mere passage of time."
I guess all three causes contribute to symptom reductions and they do not exclude each other.

Response: We agree and made said change.

Discussion, p. 14, first line: The claim that antidepressants work better in real practice is very common in discussions, but it would be good to reference to such claims in the literature.
Furthermore, the recent PANDA study (Lewis et al., 2019) also found very small and non-significant findings for treatment with an antidepressant (sertraline) in its "real-world" design and may be worth mentioning in the discussion as confirming the finding from the STAR*D study.

Lewis et al. (2019). The clinical effectiveness of sertraline in primary care and the role of depression severity and duration (PANDA): A pragmatic, double-blind, placebo-controlled randomised trial. The Lancet Psychiatry, 6(11), 903–914. https://doi.org/10.1016/S2215-0366(19)30366-9

Response: We found conflicting claims in the literature regarding antidepressants relative effectiveness in clinical practice vs clinical trials. We therefore decided to delete the entire sentence and go straight to the Conclusion. On reflection, we do not think the sentence adds much and could be a point of contention.

Pigott et al. correctly point out the important issue that the myth of ~70% remitters have became common clinical knowledge. And it seems there is no felt need for correction.

A Google search with "STAR*D" immediately leads to the NIMH information about the STAR*D study, which still has the wrong and misleading finding: "In conclusion, about half of participants in the STAR*D study became symptom-free after two treatment levels. Over the course of all four treatment levels, almost 70 percent of those who did not withdraw from the study became symptom-free" https://www.nimh.nih.gov/funding/clinical-research/practical/stard/allmedicationlevels

The myth is also repeated in a recent NYT article: ""If you look at the STAR*D, better than 60 percent of those patients actually had a very good response after going through those various levels of treatment," said Dr. Gerard Sanacora, a professor of psychiatry at the Yale School of Medicine." https://www.nytimes.com/2022/11/08/well/mind/antidepressants-effects-alternatives.html?smid=nytcore-ios-share&referringSource=articleShare

I guess there are myriads of similar repetitions of this myth, highlighting the importance of publishing the re-analysis by Pigott et al.

Response: Thank you for the link to the recent NYT's article. It is striking how even today, STAR*D's Summary article is the go-to article purporting to show antidepressants' effectiveness in clinical practice. We therefore added the sentence…"More recently (2022), a New York Times' article claimed that half of STAR*D's participants "had significantly improved after using either the first or second medication, and nearly 70 percent of people had become symptom-free by the fourth antidepressant."

The "symptom-free" line was lifted from the NIMH press release. Although an HRSD score of <8 is the common criterion for classifying remission, such a score is by no means synonymous with patients becoming "symptom-free" because patients could have up to seven HRSD symptoms mildly expressed and still met this criterion (or several depressive symptoms moderate to severely expressed).

Reviewer: 4
Dr. Jake Olivier
University of New South Wales
Comments to the Author:

I have conducted a statistical review of the submission. There are some content issues that I am unsure of, but I have not commented on them as content-specialist reviewers have provided reviews.

No analysis plan has been provided in the submission. How was the STAR*D data analysed by the original researchers? Were those analyses reasonable or adequate? Can the current authors provide evidence their analytic plan is better? These are important questions and their answers are not clear.

Response: STAR*D investigators used the non-blinded/clinic-administered QIDS-SR as the sole measure to report treatment remission, response, and extent of improvement rates in their widely-cited summary article. This is contrary to the research protocol which barred all clinic-administered measures from being used to report outcomes.

The investigators also deviated from their protocol and related publications by including in their summary article 931 patients who lacked a baseline HRSD score of ≥14 including 99 patients who scored <8 on their baseline HRSD—indicating these patients met STAR*D's remission criterion at study outset and should not have been included in their report of outcomes. STAR*D investigators also Included in their analyses 125 patients who scored as remitted at entry into their next-level treatment. This occurred despite STAR*D investigators prespecifying that, "patients who begin a level with HRSD <8 will be excluded from analyses." For these reasons, we believe that STAR*D investigators' analyses were neither reasonable nor adequate since they deviated from the protocol.

We sought to reanalyze STAR*D's patient-level dataset with fidelity to the original research protocol and related publications. As we state in the "Analytic Plan of the RIAT Reanalysis" section, "where the protocol was silent, we used other STAR*D publications to guide our analysis" and then describe the four times we relied on related publications to guide our reanalysis. We believe our analytic plan is superior because it aligns with STAR*D's protocol and related publications.

Last observation carried forward (LOCF) is widely considered a poor approach to dealing with missing data. Here is one reference but there are many criticisms of this method. It is also unclear to me what the current authors did with participants who exited without an HRSD score. But, certainly using LOCF would not be a valid approach after treatment was given.

Response: We agree and felt uncomfortable using LOCF which is why we highlighted twice in the manuscript that there were 1,330 patients with missing exit HRSD scores across all treatments and presented STAR*D's cumulative remission rate two ways. Unfortunately, STAR*D's protocol is silent regarding how they planned to handle missing data. It is only in related publications that they repeatedly state, "our primary analyses classified patients with missing exit HRSD scores as nonremitters a priori." In their level 2-4 articles, STAR*D investigators used a correspondence table to map the final QIDS-SR score to the HRSD for patients missing their exit HRSD score to assess the impact of their approach to counting such patients as "nonremitters a priori" (see: IDS/QIDS (ids-qids.org). Each time they reported "consistent findings indicated that the results were not affected by this approach to missing data" but this was used only for statistical comparison of remission rates between treatments; not descriptive statistics of each treatment's remission rate.

We have therefore dropped LOCF and instead used the correspondence table to map the final QIDS-SR score to the HRSD for patients missing their exit HRSD score and used it to calculate the mean HRSD exit, mean change, and combined HRSD & QIDS-SR response rates for all treatments. This approach resulted in no missing exit data and increased the mean change and combined HRSD/QIDS-SR response rates from what we previously reported. For example in step 1, we had reported mean change on the HRSD of 6.41 points and a response rate of 32.3% ; now it is 8.4 points and 40.5% respectively.
Given the potential biasing of the final QIDS-SR score for patients missing their exit HRSD (as discussed in our response to Dr. Plöderl), we believe this approach provides an upper limit "best case scenario" for mean change and response rates but is more accurate than our use of LOCF.

Lastly, was loss to follow up related to response? From what I understand, the STAR*D research would be roughly correct if they are not related, but would be problematic otherwise.

Response: In a preliminary analysis in which we used the QIDS-SR to estimate missing posttreatment HRSD scores, we found that the posttreatment severity scores for patients who dropped out were different from patients who did not at level 1 ($p<0.00$). We found that the mean posttreatment HRSD score was 14.2 (SD=8.2) for patients who dropped out (n=1074) vs. 13.1 (SD=8.5) for patients who did not drop out (n=2036). We had used LOCF originally because we were concerned a multiple imputation approach would incorrectly estimate lower/better posttreatment HRSD scores for missing outcomes. However, our new methodology (i.e., using the correspondence table to map available QIDS-SR to missing posttreatment HRSD scores), we believe, provides a more accurate estimate of posttreatment severity scores, and therefore also, response and remission rates.

Abstract: Please provide remission rate values instead of unquantified statements.

Response: The Abstract currently states that the cumulative remission rate was 35%.

Strengths/Limitations: Please provide brief results of comparator trials. It is not clear whether your results are similar or not to these trials.

Response: We agree and made said changes.

I am not sure I am comfortable with emotive language like "oversized impact" or "unrealistic provisos" or "taking a failed study" without there being evidence to support such statements. The research by the STAR*D authors may have had an appropriate impact on the discipline or maybe not. That needs to be demonstrated and not assumed to be true or false.

Response: We agree and made said changes.

VERSION 3 – REVIEW

REVIEWER	Moncrieff, Joanna University College London and North East London mental health trust, Mental Health Sciences
REVIEW RETURNED	27-Apr-2023

GENERAL COMMENTS	I am happy that the reviewers have addressed all my comments. I recommend timely publication of this important paper.

REVIEWER	Plöderl, Martin Christian Doppler Clinic, Paracelsus Medical University, Dept. of Clinica Psychology and Dept. for Crisis Intervention and Suicide Prevention, University Clinic for Psychiatry, Psychotherapy, and Psychosomatics
REVIEW RETURNED	01-May-2023

GENERAL COMMENTS	Pigott et al. addressed the concerns of the reviewers and I have only very minor suggestions with no need for another round of review. I agree that imputing patients' missing HDRS exit scores with the self-report QUID-SR measure is a better way than to use the LOCF appraoch. I would like to comment on the concerns of reviewer #4, when he wants to know if the approach by Pigott et al. is better than the

	approach by the original investigators of the STAR*D trial. I agree with Pigott et al. when they say „We believe our analytic plan is superior because it aligns with STAR*D's protocol...". The important goal of the paper by Pigott et al. is to compare the results of the original publications of the STAR*D trial, which were based on unjustified/unexplained deviations from the protocol. Now Pigott et al. contrast this with the results from the data-analysis according to the protocol. There is no doubt that unjustified/unexplained deviations from the protocol are scientific no-go's. It still strikes me that the biggest depression trial of all times, despite being publicly funded, and published in prestigious journals, was not analyzed as planned in the protocol, and none of the deviations from the protocol were made explicit or justified, despite leading to very different results (which I assume the original investigators were aware of). Thus, the current paper by Pigott et al. is very important to correct not only the scientific record but perhaps also clinical practice which still may be guided by the widespread belief that antidepressant work better in real-world clinical practice than in clinical trials (where antidepressants are hardly superior to placebo). A belief which was much the result of the the misleading original publications of the STAR*D trial. Minor issues: Introduction: "Remission as defined by the HRSD was not presented in STAR*D's summary article." Perhaps change to "Remission as defined by the HRSD (according to the protocol) was not presented in STAR*D's summary article." Results: "In step 1, these measures of improvement among STAR*D's patients were one-third or more less than that found in comparator trials" Consider using changing the expression: "were one-third or more less"

REVIEWER	Olivier, Jake University of New South Wales, School of Mathematics and Statistics
REVIEW RETURNED	03-May-2023

GENERAL COMMENTS	Thank you for responding to my earlier comments. When I asked for analysis plans, both for STAR*D and your reanalysis, I wanted a plan for how the data would be analysed. This could have been a linear or generalised linear model, or some approach with random effects to account for within-subject dependence. Can you provide that and compare what was planned and performed for the STAR*D trial and your reanalysis? Deviating from an analysis plan can be problematic in its own right, but I am also concerned as to whether the meta-analysis performed is comparable to the STAR*D and your analyses.

	It could be that you performed the same "type" of analysis as the STAR*D trial and the only difference is in the processing of the data. But this is not clear to me.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 1

Dr. Joanna Moncrieff

University College London

I am happy that the reviewers have addressed all my comments. I recommend timely publication of this important paper.

Response: We agree and thank you for your many helpful comments.

Reviewer: 2

Dr. Martin Plöderl

Christian Doppler Clinic

Paracelsus Medical University

Pigott et al. addressed the concerns of the reviewers and I have only very minor suggestions with no need for another round of review. I agree that imputing patients' missing HDRS exit scores with the self-report QUID-SR measure is a better way than to use the LOCF approach.

Response: We agree and thank you for your many helpful comments.

I would like to comment on the concerns of reviewer #4, when he wants to know if the approach by Pigott et al. is better than the approach by the original investigators of the STAR*D trial. I agree with Pigott et al. when they say „We believe our analytic plan is superior because it aligns with STAR*D's protocol...".

The important goal of the paper by Pigott et al. is to compare the results of the original publications of the STAR*D trial, which were based on unjustified/unexplained deviations from the protocol. Now Pigott et al. contrast this with the results from the data-analysis according to the protocol. There is no doubt that unjustified/unexplained deviations from the protocol are scientific no-go's. It still strikes me that the biggest depression trial of all times, despite being publicly funded, and published in prestigious journals, was not analyzed as planned in the protocol, and none of the deviations from the protocol were made explicit or justified, despite leading to very different results (which I assume the

original investigators were aware of). Thus, the current paper by Pigott et al. is very important to correct not only the scientific record but perhaps also clinical practice which still may be guided by the widespread belief that antidepressant work better in real-world clinical practice than in clinical trials (where antidepressants are hardly superior to placebo). A belief which was much the result of the misleading original publications of the STAR*D trial.

Minor issues:

Introduction:

“Remission as defined by the HRSD was not presented in STAR*D’s summary article.”

Perhaps change to “Remission as defined by the HRSD (according to the protocol) was not presented in STAR*D’s summary article.”

Response: We agree and made said change.

Results:

“In step 1, these measures of improvement among STAR*D’s patients were one-third or more less than that found in comparator trials;” Consider using changing the expression: “were one-third or more less.”

Response: We agree and changed it to read: “In step 1, these measures of improvement among STAR*D’s patients were at least one-third less than that found in comparator trials.”

Reviewer: 4

Dr. Jake Olivier, University of New South Wales Comments to the Author:

Thank you for responding to my earlier comments.

Response: Thank you for encouraging us to drop LOCF. We think the change we made clearly improved the paper.

When I asked for analysis plans, both for STAR*D and your reanalysis, I wanted a plan for how the data would be analysed. This could have been a linear or generalised linear model, or some approach with random effects to account for within-subject dependence. Can you provide that and compare

what was planned and performed for the STAR*D trial and your reanalysis? Deviating from an analysis plan can be problematic in its own right, but I am also concerned as to whether the meta-analysis performed is comparable to the STAR*D and your analyses. It could be that you performed the same "type" of analysis as the STAR*D trial and the only difference is in the processing of the data. But this is not clear to me.

Response: In this paper, we focused on replicating the Rush et al. summary article which used descriptive statistics to present the remission, response, and extent of symptomatic improvement for all 14 antidepressant therapies as well as the overall remission rate based on the QIDS-SR.[1] We did the same type of analyses with the key differences compared to those presented in STAR*D's summary article being: 1) ours is based on the protocol-specified HRSD and only used the QIDS-SR for those patients missing their exit HRSD and 2) we only included patients who met the inclusion for data analysis criteria stipulated in the research protocol and related publications.

The inferential statistical analyses for the STAR*D trial are presented in their treatment levels 2-4 articles where they ran logistic regression models to identify whether there were differences in remission based on medication received, as well as controlling for predictors such as demographic/clinical covariates.[2-6] This is consistent with the protocol's analytic plan. Our plan is to conduct these inferential statistical analyses in replication of the treatment levels 2-4 in future papers yet include only those patients who met the inclusion for data analysis criteria.

References:

1. Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, Niederehe G, Thase ME, Lavori PW, Lebowitz BD, McGrath PJ, Rosenbaum JF, Sackheim HA, Kupfer DJ, Luther J, Fava M. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry. 2006;163:1905–1917.

2. Rush AJ, Trivedi MH, Wisniewski SR, Stewart JW, Nierenberg AA, Thase ME, Ritz L, Biggs MM, Warden D, Luther JF, Shores-Wilson K, Niederehe G, Fava M. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. N Engl J Med. 2006;354:1231–1242.

3. Trivedi MH, Fava M, Wisniewski SR, Thase ME, Quitkin F, Warden D, Ritz L, Nierenberg AA, Lebowitz BD, Biggs MM, Luther JF, Shores-Wilson K, Rush AJ. Medication augmentation after the failure of SSRIs for depression. N Engl J Med. 2006;354:1243–1252.

4. Fava M, Rush AJ, Wisniewski SR, Nierenberg AA, Alpert JE, McGrath PJ, Thase ME, Warden D, Biggs MM, Luther JF, Niederehe G, Ritz L, Trivedi MH. A comparison of mirtazapine and nortriptyline following two consecutive failed medication treatments for depressed outpatients: a STAR*D report. Am J Psychiatry. 2006;163:1161–1172.

5. Nierenberg AA, Fava M, Trivedi MH, Wisniewski SR, Thase ME, McGrath PJ, Alpert JE, Warden D, Luther JF, Niederehe G, Lebowitz BD, Shores-Wilson K, Rush

AJ. A comparison of lithium and T3 augmentation following two failed medication treatments for depression: a STAR*D report. Am J Psychiatry. 2006;163:1519–1530.

6. McGrath PJ, Stewart JW, Fava M, Trivedi MH, Wisniewski SR, Nierenberg AA, Thase ME, Davis L, Biggs MM, Shores-Wilson K, Luther JF, Niederehe G, Warden D, Rush AJ. Tranylcypromine versus venlafaxine plus mirtazapine following three failed antidepressant medications trials for depression: a STAR*D report. Am J Psychiatry. 2006;163:1531–1541.

VERSION 4 – REVIEW

REVIEWER	Olivier, Jake University of New South Wales, School of Mathematics and Statistics
REVIEW RETURNED	22-May-2023

GENERAL COMMENTS	Thanks for addressing my previous concerns about the STAR*D and your analytic plan for this paper. However, I believe the manuscript should reflect that. That is, explicity state you performed a descriptive analysis for this paper as did the STAR*D paper by Rush et al (something like the first paragraph in your response should suffice).

VERSION 4 – AUTHOR RESPONSE

Response to the Reviewer #4

Dr. Jake Olivier

University of New South Wales

Thanks for addressing my previous concerns about the STAR*D and your analytic plan for this paper. However, I believe the manuscript should reflect that. That is, explicitly state you performed a descriptive analysis for this paper as did the STAR*D paper by Rush et al (something like the first paragraph in your response should suffice).

Response: We agree. Thank you for your helpful comments that have improved the paper. We've now changed the Introduction section's final paragraph to read:

"This reanalysis article uses the patient-level dataset obtained from NIMH to replicate the STAR*D summary article which used descriptive statistics to present the remission, response, and extent of symptomatic improvement for 14 antidepressant therapies based on the QIDS-SR.[7] We perform the same descriptive analyses with the key differences compared to those presented in STAR*D's summary article being: 1) ours is based on the protocol-specified HRSD and only uses the QIDS-SR for those patients missing their exit HRSD and 2) we only included patients who met the inclusion for data analysis criteria stipulated in the research protocol and related publications. Future efforts will use inferential statistics to reanalyze STAR*D's levels 2-4 semi-randomized comparator trials, including the extent of emergent suicidal ideation and 12-month follow-up outcomes tied to each compared treatment."