	Reviewer's comment	Our response	Change made
1	Ed: Along with a number of reviewers and our statistical advisor I continue to think the paper would be stronger if you performed imputation. Performing these analyses would also demonstrate that you are doing your best to be fair and make the best and highest use of the data. There are arguments on both sides, of course, so we will not insist on this, particularly since readers of the prepublication history for the paper will see the back and forth about this matter and will be able to judge the matter for themselves. FN: Concerning efficacy analyses, I agree that the analysis pre-specified in the protocol is the analysis that must be done. AW: With reference to utilising more appropriate means of analysis, albeit these were not in the study protocol, as requested by 2 reviewers (Hilde PA van der Aa, comments 1 and 4, Sarah Hetrick comment 9) and the committee (Loder comment 9), I do not think the argument for continuing to use LOCF alone ("It continues to be widely used") is valid. Although the technique is still seen it is well known to potentially give biased results. … However the authors do argue (letter to Loder) that the point of a RIAT is "not to repeat all that was done in a published paper but rather what should have been done according to study protocol". I do not know RIAT well but it does appear that box 2 point 6 of the original RIAT paper does not preclude the necessity in some cases for analyses additional to the protocol. Additionally the authors argue that "over time and with much back and forth, we ended up deciding that the choice of analytic approach was a potential source of bias (our own bias)", but I do not think that this a strong argument either since an attempt at multiple imputation would seem standard today (as was requested by 2 reviewers plus the committee). In defence of the authors, the argument that "if more 'modern' methods of data imputation could have in any way redeemed this study, one imagines GSK would have done so" does seem reasonable to me. Furthermore, the OC and LOCF results are similar which also suggests that conclusions would not be	On careful reflection, we have decided against using imputation, largely for reasons that we have already set out. We note that we have the support of two of four of this round's reviewers, that AW is in favour of imputation, but acknowledges contrary arguments, and that DH is silent on the issue.	None

	changed by more appropriate imputation. PD: I read through the other reviewers' concern about modern statistical methods and agree with the RIAT authors' response that the primary purpose of RIAT is to stick to the protocol as best possible. While it is perfectly valid to subject the data to additional analyses, such re-analyses would need to be clearly labeled post-hoc, and can always happen in subsequent papers by different authors given the public availability of the trial data. But unless there is a strong argument that the statistical methods in the original study protocol are not just outdated but simply WRONG, I would agree with the RIAT authors position that the analyses should be conducted according to the original protocol.		
2	AW: In box 1, 'Missing values' paragraph: "are frequently preferred" (referring to MI and MM) should be replaced by "are shown to be superior".	Agreed	Now reads 'are superior'.
3	AW: The authors should add confidence intervals to the estimates given in the abstract (addition to response to Loder 9).	Done	10.73 [9.134, 12.328], 8.95 [7.356, 10.541] and 9.08 [7.450, 10.708] points, LS MEAN [95%CI]. We have also added CIs to table 3 and the accompanying text in the results section.
4	DH: Double-blind should be mentioned under 'Design'	Done	Design: Double- blind randomised placebo-controlled trial.
5	DH: The authors state: "Clinically significant increases in harms, including suicidal ideation and behaviour and other serious adverse events, were observed in the paroxetine group." They have chosen to make this prominent in the abstract. They have taken the privilege of featuring a difference in harms between active and control as 'clinically significant' without assessing statistical significance. … The authors make a virtue of not applying statistical analyses to harms and yet want to highlight the differences. For the record, a chi square test with 2 degrees of freedom is significant applied to the psychiatric AE data across	Statistical significance does not confer clinical significance, and we contest the view that we cannot claim clinical significance – just look at the data (for example in table 11)! We stand by the opinion that the AE profile of paroxetine is the most striking finding from this dataset and should therefore be reported prominently.	None

	the two active and one control group. Assuming the summary estimates of psychiatric AEs are not affected by un-blinded assessment (see below) I think they should present statistical analyses of the main harms.		
6	DH: They say that they have discussed the reasons for not applying statistical significance testing to harms but Box 3 is minimal in this regard.	We disagree. We believe that our current text addresses the issue: 'In our opinion, statistically significant or not, all relevant primary and secondary outcomes, and harms outcomes, should be explicitly reported. Testing for statistical significance is most appropriately undertaken for the primary outcome measures. We have not undertaken statistical tests for harms, since we know of no valid way of interpreting them. To get away from a dichotomous (statistically significant/non significant) presentation of evidence, we opted to present all original and recoded evidence to allow readers their own interpretation. The data presented in Appendix 2 and related worksheets lodged at www.xxx will, however, readily permit other approaches to data analysis for those interested, and we welcome other analyses.'	No change
7	Ed: I agree with Professor Henry that the abstract needs attention and encourage you to adopt the more neutral and balanced wording he suggests in several places, particularly when discussing the balance between efficacy and harms. DH: The authors write: "Paroxetine was neither well tolerated nor effective for major depression in adolescents. Imipramine, given in high doses, was also poorly tolerated and was not shown to be effective" I don't know what "well" or "poorly tolerated" means in this context. The term often applies to common non-serious symptomatic AEs (eg nausea, dry mouth visual blurring) that interfere with daily activities. Based on what they have presented an alternative would be to say "Neither paroxetine nor imipramine demonstrated efficacy in	We deliberately paraphrased the wording of the Keller paper, but are happy to accept this modification. We have significantly reworked the abstract (see also #8 & #9 below), and it is the better for it.	Neither paroxetine nor high-dose imipramine demonstrated efficacy for major depression in adolescents, and there was an increase in harms with both drugs.

	adolescents, and there was an apparent increase in harms with both drugs." The key AEs could be quantified here.		
8	DH: "This study has demonstrated that when there is access to primary data, trial conclusions will ordinarily be provisional rather than authoritative." I think that's a big call and too difficult to introduce in the abstract. Something like "the re-analysis of trial 329 illustrates the value of making primary trial data available" would be OK	This is not a paper about Paxil so much as a paper about authorship and the validity of conclusions in scientific papers. We have rephrased for clarity.	Access to primary data from trials has important implications for both clinical practice and research, including that published conclusions about efficacy and safety should not be read as authoritative. The re-analysis of Study 329 illustrates the necessity of making primary trial data available to increase the rigour of the evidence base.
9	DH: Introduction. They have to introduce two concepts we could argue which gets mentioned first. My preference would be to mention Trial 329 and the clinical context before introducing RIAT. I accept arguments can be made for both sequences.	We see the point, but think it is best left as is, for reasons noted in #8 above and now made clear in our abstract.	The primary objective of the trial was to compare the efficacy and safety of paroxetine and imipramine to placebo in the treatment of adolescents with unipolar major depression. The objective of the restoration was to see whether access and reanalysis of a full dataset from a randomised controlled trial (GSK's Study 329, published by Keller et al. in 2001) under the Restoring Invisible and Abandoned Trials (RIAT) initiative would have clinically relevant implications for evidence based medicine.
10	DH: Methods. Reordering of the sections of the paper would be helpful. I think the details of randomization and assignment should follow the description of the interventions. In the current version they appear very late in the Methods	We were following the RIAT proposal, but have no objection to changing.	We have moved the sample size, randomization and blinding such that they now follow the description of the interventions.
11	DH: The position of Box 1 (challenges to carrying out RIAT) is awkward and breaks up the flow of the manuscript. It should be repositioned in the production phase.	We are happy for the editor to position Box 1 as judged appropriate.	
12	DH: The authors state "Only for six events from the eleven serious adverse event narratives was it not possible to be blind. This was 0.005% of events." I think we need to know whether the un-blinded assessment of these 6 serious AEs has a possible	DH has misunderstood, and we can see that it is confusing. These were not the serious AEs that were coded unblind, but "extra" non-serious events described within the SAE	We have deleted the sentence 'Only for six events from the eleven serious adverse event narratives was it not possible to be blind. This was 0.005% of

	effect on the results – what do the results look like if they are removed? For instance what does Table 12 (Discussion) look like? … how do Tables 5-7 change if the un-blinded adjudications of harms are removed? … Table 12 has been referred to above in regard to un-blinded assessments.	narratives that had been left uncoded or were coded and never transcribed. It was not possible to be blind to these, because allocation status was written into the narratives. We had two MedDRA trained coders review redacted SAEs and both coded the events in the same way, so to avoid other readers being confused we now think it appropriate to delete the qualification.	events.'
13	DH: In view of their importance and since the unblended SAEs are small in number could those not have been recoded with allocation status masked in some way?	Done as requested	See #12
14	DH: The Harms section has too many tables. Table 4 could be in text.	We believe it is clearer in a table.	No change
15	DH: Severity ratings in Table 7 could be added to Table 5	Agreed	New table 5 with suitably changed legend
16	DH: The sections on discontinuations and withdrawals are long and perhaps the authors could decide what could be placed in an Appendix.	This is important material – and not usually covered in a paper – its important to let people know material like this exists	No change
17	DH: The text in the Discussion section is brief. In part this is because some elements appear in other parts of the paper. A tight edit could identify these and move them – that is a style/editorial decision.		No changes made
18	DH: Box 3 is useful but the word 'confounder' in the title has a technical meaning in epidemiology and its use here is not accurate.	Agreed	Box 3. Potential barriers to accurate reporting of harms; have also removed another reference to confounders
19	PD: Abstract - Setting. Give exact dates (not just years).	Done	20 April 1994 to 15 February 1998.
20	PD: Tables 5 & 6. Three columns contain the text "additional AEs found in 93 CRFs". These cannot all be n=93. Please give the respective number of CRFs reviewed for paroxetine, imipramine, and placebo.	Agreed	Changed in both tables to 31/40/22
21	PD: The authors states that they followed the April 17, 1994 protocol. However the protocol I see uploaded to Scholar One is	We thank the reviewer for catching this discrepancy. We have, in fact, used the 1996	Have made clear that we are following the 1994 version apart from the

	dated June 12, 1993. Was this intentional? The protocol that is available on GSK's website appears to be the one dated from 1996. So unless I have missed it, I don't see the 1994 protocol. I would suggest that the authors provide the 1994 protocol (which I agree as the final protocol prior to patient enrollment is the appropriate one to use) and ensure that it is available online with the published paper.	protocol that contains two approved amendments. The first amendment, replacing the K-SADS-P with the K-SADS-L, was approved on 17 April 1994 (prior to enrolment of the first patient). The second amendment, approved on 28 October 1996, involved re-estimating the required sample size for the acute phase, and proposals on how to proceed with the continuation phase (both in response to limited medication supply), as well as a change in the contact information of the medical monitors. These changes do not have any impact on our reanalysis. We have edited the manuscript to reflect the correct version of the protocol used.	reduction in numbers in 1996, on pp 5, 9 & 14.
22	PD: The authors mention further correspondence with GSK asking them for documentation to support GSK's claim that the outcomes introduced in the Keller paper which did not appear in the 1993, 1994, or 1996 trial protocols was nonetheless defined prior to breaking the blind. The authors state GSK was not forthcoming with this documentation. I would suggest that BMJ ask the authors whether there are any updates on this correspondence.	GSK have still not provided any evidence to refute our statement that 'we have no contemporaneous documentation of that claim, despite having repeatedly requested it from GSK'	No change required
23	PD: In their response #21 to the editors, the authors write that the "periscope" model of data access (which GSK required in order for the authors to read CRFs) prevented them from printing off materials and submitting them to a panel of coders in an effort to reduce bias, etc. I think this is a very valuable observation and should be stated in the Discussion of the paper.	We have reformulated the discussion as requested.	The single screen remote desktop interface (we called the "periscope") proved to be an enormous challenge. The efficacy analysis required multiple spreadsheet tables be opened simultaneously, with much copying, pasting, cross-checking, and the space was highly restrictive. Gaining access to the CRFs required extensive correspondence with GSK.[11] Although GSK ultimately provided CRFs, they were even harder to manage, given that could we see only one page at a time. It required of the order of one thousand

			hours to examine only a third of the CRFs. Being unable to print was a significant handicap. There were no means to prepare packets for multiple independent coders to decrease bias; to make annotations or use marginalia; or to sort and collate the Adverse Event reports. Our experience highlights the crucial value of hard copies is for an enterprise like this.
24	PD: In response to my query #7 regarding the missing pages of CRFs, the authors write "See Loder, query 25." But I do not see any answer to my question here. Please clarify as I still think this is an important point.	We have clarified, see change.	At least 1000 pages were missing from the CRFs reviewed with no discernible pattern to missing information; for example, one CRF came with a page inserted stating that pages 114 to 223 are missing without indicating why.
25	PD: The authors response to my query #11 is helpful, regarding the reasons for how they chose the MedDRA SOC classes. I did not see the authors include this rationale in the paper itself, however, and think it should be included.	Have added explanatory phrase	(consistent with the published paper)
26	PD: RIATAR. Section 24 states that the protocol used was the one in CSR Appendix A. But isn't this the 1996 protocol, not the 1994 one the authors used? Please clarify	See above #21.	
27	PD: Also, in RIATAR section 17a ("Outcomes and estimation"), the authors include numerous sections of the CSR including data tables. This confuses me because based on the Conclusions section of Box 1 ("Challenges in carrying out RIAT"), my impression was that the authors used the electronic IPD they had access to via GSK, and not efficacy data from the CSR. Please clarify.	We understand PD to be saying that the RIATAR should only point to the sources we used for the RIAT. Given the nature of our work, which often involved cross-checking multiple documents and sections to verify things, the RIATAR was developed and used to refer to all places in the documents we had that have information relevant to the item in question, so that others could check our work.	No change
28	PD: under this same section of the RIATAR (i.e. 17a), the authors list "Data Source Tables: Safety, pages 113-260". This confuses	We considered safety data, in addition to efficacy data, as outcomes, which is why we	No change

	me because safety data usually is not included in 17a.	included it in 17a as well.	
29	PD: Methods. I would revise the sentence in the first paragraph, inserting the BOLDED words, "...the INDIVIDUAL PARTICIPANT LEVEL data access system SAS Solutions OnDemand,[10] on which GSK SUBSEQUENTLY ALSO posted some Study 329 documents (available only to users approved by GSK..."	Done	the individual participant level data access Solutions OnDemand,[10] on which GSK subsequently also posted some Study 329 documents (available only to users approved by GSK).
30	Additional change related to newly identified publication: Ebrahim S, Sohani ZN, Montoya L, Agarwal A, Thorlund K, Mills EJ, Ioannidis JP. Reanalyses of randomized clinical trial data. JAMA. 2014 Sep 10;312(10):1024-32	This shows our comment about being possibly the first RCT republished by independent authors was incorrect.	Now reads: Very few previously published RCTs have been reported in published papers by different teams of authors.
31	Table 11 gives a misleading impression of Keller et al's reporting of psychiatric AEs.	This was unintentional on our part, now corrected.	See changes to table 11 and preceding text.