Re: Manuscript ID BMJ.2014.022376 entitled "A randomized, controlled trial of the efficacy and harms of paroxetine and imipramine in the treatment of adolescent major depression: Restoring Study 329"
Response to the reviewers and committee providing, point by point, replies to the comments made by the reviewers and the editors, and explaining how we have dealt with them in the paper.

| Reviewer comment | comment to editor | change made |
|---|---|---|
| Loder and committee | | |
| 1. The full text online version of your article, if accepted after revision, will be the indexed citable version (full details are athttp://resources.bmj.com/bmj/about-bmj/the-bmjs-publishing-model), while the print and iPad BMJ will carry an abridged version of your article, usually a few weeks afterwards. This abridged version of the article is essentially an evidence abstract called BMJ pico, which we would like you to write using a template and then email it to papersadmin@bmj.com (there are more details below on how to write this using a template). Publication of research on bmj.com is definitive and is not simply interim "epublication ahead of print", so if you do not wish to abridge your article using BMJ pico, you will be able to opt for online only publication. Please let us know if you would prefer this option. | We would like to abridge | - |
| 2. As well as submitting your revised manuscript, we also require a copy of the manuscript with changes highlighted. Please upload this as a supplemental file with file designation 'Revised Manuscript Marked copy'. | | done |
| 3. did you register the study in an approved trial registry? | we have not done a trial; we have provided the trial registration details in our paper | None required |

| | | |
|---|---|---|
| 4. how many versions of the protocol are there, and if there was more than one, how did you choose which one to follow? | To our knowledge, there were 2 full versions: 1993 (signed by Principal Investigator 2 June), and April 1994 (amended 24 March; approved 17 April). The 1994 version that we followed included Amendment 1, which specified substitution of K-SADS-L instead of K-SADS-P; optional external review of diagnosis; additional safety measures; and a replacement SB Medical Monitor. There was also Amendment 2 in 1996 (28 October; detailed in the CSR, pp. 000027-000028) reducing the sample size to approximately 275 patients, but otherwise unchanged. | None required |
| 5. * We agree with several of the reviewers that the problem of potential bias and conflict of interest needs more attention. We would like to hear your thoughts about these matters and we think some comment in the paper itself might be necessary. | One point behind a RIAT article is by making the data available we hope to allow others to make judgements about the possible influence conflicts of interest that authors might not themselves see.<br>With standard COI declarations, readers then have to guess whether itemised conflicts have had an influence or not.<br>We aim to make it possible for readers to do their own analysis and if this analysis differs from ours there may be a pattern to the difference that is indicative of some kind of bias.<br>We expect GSK to re-analyze the data and produce an alternate reading but we welcome this. One of our main points behind the article is to show that there is no privileged interpretation that someone who has no conflicts of interest can arrive at. | COIs now incorporated into document title page, + see new box 1 |
| 6. We did not agree, for example, that it makes sense to move symptoms such as dizziness and headache out of the nervous system cluster. | The published paper put psychiatric events into a cluster called Nervous System. MedDRA separates Psychiatric from nervous system.<br>We argue that it is inappropriate to code dizziness and headache as neurological issues. The most likely cause of dizziness given the drugs involved is cardiovascular. Headaches most commonly stem from muscles and blood vessels to the scalp – not part of the CNS. | The relevant section (now shifted to Discussion) now reads:<br>'The CSR and CRF figures also differ substantially from other figures quoted in Keller et al, because we did not code 'dizziness' and 'headache' under Nervous System, since the former is more likely to |

| | | |
|---|---|---|
| | The important point behind our coding is not where we put these items but rather drawing attention to the fact that wherever items like this are put can significantly affect the interpretation. | be attributable to 'cardiovascular' while headaches most commonly stem from muscles and blood vessels to the scalp.' |
| 7. We agree with reviewers that coding of adverse events needs to be redone by people who are independent of your group. | 1. 100% of the initial coding was done blind – all of the extra coding from the CRFs was done blind – as the drug name was not on the list<br>It was only for the eleven SAEs where it was not possible to be blind and not all of these gave extra codes – so 99.995% of the coding was blind<br>2. JLN was recruited to the group specifically because of her expertise to carry out these analyses. Neither she nor JN, who did the analysis independently and agreed on all ratings, has COI.<br>3. JJ, who has provided expert opinion in a class action, did not analyse the efficacy or the adverse event data.<br>4. We know of no precedent for analyses to be carried out outside a research group. | Added:<br>'All of the initial coding from the the clinical descriptions in the CSR was done blind, as was coding from the CRFs. Only for six events from the eleven serious adverse event narratives was it not possible to be blind. This was 0.005% of events.' |
| 8. We also agree with several of the reviewers that extrapolation of AEs from the non-random sample of CRFs is unwise. This analysis should be removed from the paper. (Table 6) | Agreed, will remove | Deleted from various tables |
| 9. * Please present a true ITT analysis (in other words, analyze all subjects in the groups to which they were randomised, regardless of whether they received the study drug or not). Our statistician suggests that you consider having several columns in your results table. The first would present an ITT analysis using LOCF, the second using imputation and correcting for strata (12 centres). The third column could show the per protocol or complete case analysis using LOCF and the fourth the per protocol or | The Protocol called for evaluation of the OC [Observed Case] data and the LOCF [Last Observation Carried Forward] dataset with the latter being definitive. The LOCF method for correcting missing values was the standard at the time the study was conducted. It continues to be widely used, though newer models such as Multiple Imputation or Mixed Models are now frequently preferred. We chose to stick to the Protocol and use the LOCF method rather than introduce a post hoc analytic tool. | See new figure 2, and commentary in new box 1 |

| | | |
|---|---|---|
| complete case analysis using imputation. This would allow readers to judge for themselves the effects, if any, of using more modern methods of analysis, while still showing the originally intended efficacy analysis. | | |
| 10. * We would also like to see the results of pairwise comparisons. | We conducted the protocol-specified omnibus analyses, which are negative as shown. Nevertheless, the pairwise results were confirmed as non-significant as reported by Keller et al. These are tabled in the appendix 2. | Figure 1&2<br>Appendix 2 – Table i |
| 11. * Can you please also include a table that contrasts all of your findings with those of the original paper? You do this for AEs but not for the efficacy outcomes. Many editors commented that it was difficult to understand how and where the reanalyses differ from the original ones. | The contrast of relevance is with the CSR rather than the published paper, but there is no significant discrepancy between our results and GSK's | 'There were no discrepancies between any of our analyses and those contained in the CSR.' added to efficacy results |
| 12. * we were disappointed that you did not examine the CRFs for all subjects. This seems a serious problem. It is, we understand, a major undertaking to review all of these documents, but seems necessary to set the record straight. After all, the trial itself was a major effort on the part of the original investigators. | See also letter to Dr Loder.<br>We are no longer making even tentative extrapolations from the audit (and we no longer use that term), so the primary justification for completing it no longer exists.<br>Furthermore:<br>1. Completing the audit would take about 2000 hours because GSK's method of permitting access to the data is so burdensome and this would essentially all have to be done by one person. (The difficulties faced by the RIAT team are several orders of magnitude greater than the GSK team who did the original write up. GSK could have made it a lot easier for us to do the audit expeditiously and safely but chose not to do so.)<br>2. It would give only an illusion of completeness as we have already found 1000 missing pages so that there are likely 3000+ missing pages.<br>3. Because of the enormous burden in gaining access to | all extrapolations from the 'audit' removed.<br>Also have changed the way we report CRF findings (see, e.g., table 5) – no longer as part of a total, but the number of additional cases identified, which we think is more informative – no one actually knows what the total is. |

| | | |
|---|---|---|
| | and auditing CRFs, no other team (apart from GSK) is likely to have resources to check our audit. We would therefore prefer a model that sees publication of this version of the paper and then has BMJ in an editorial calling on GSK and other companies to make the data available to researchers in a user friendly format so that the audit can be readily audited by others. | |
| 13. * We believe the original investigators in the trial should be acknowledged in the paper. | the roles of the various investigators, authorship and related issues are thoroughly discussed in reference 3. | Added to Background:<br>'We acknowledge the work of the original investigators.' |
| 14. You mention that in some cases it was not clear what happened in the original study, for example why some secondary outcomes were changed. Did you make any attempt to ask the original investigators? If not, why not? | We have had considerable correspondence with GSK has published in a series of rapid responses in the BMJ. In particular GSK has not been able to produce a copy of the putative 'analytic plan' | None required |
| 15. Did you have any funding for this reanalysis? | No | Added to Abstract:<br>'No funding was obtained to support this restoration' |
| 16. * The abstract contains no numerical findings. Please present the figures for the principal study outcomes in the abstract. | Done | The following sentence has been added to the Abstract's Results section:<br>'HAM-D scores decreased by 10.73, 8.95 and 9.08 points, respectively, for the paroxetine, imipramine and placebo groups (p = 0.204).' |
| 17. * We thought that information about the alleged problems with the original study could be dealt with in a single paragraph in the introduction. | We already write: Keller et al., which was largely ghostwritten,[3] claimed efficacy and safety for paroxetine at odds with the data,[4] This is problematic because the article has been influential in the literature supporting the use of antidepressants in adolescents.[5] | None required |
| 18. Please be careful not to include ad hominem remarks. | If you can identify any such remarks, we would be happy to remove them | None found by us |

| | | |
|---|---|---|
| 19. Has the previous paper been retracted? If not, how will readers of that paper know about this one? | The previous paper has not been retracted; perhaps the publication of this paper will provide more incentive for JAACAP to do so. | None required |
| 20. * We thought you should comment on the matter of dropouts. These seemed higher in the placebo group. | They are not higher in the placebo group. We already discuss dropouts. | None required |
| 21. We also wondered whether 8 weeks is too soon to see any possible benefit of an antidepressant. Several editors who are practicing physicians and use these drugs thought that 8 weeks might be too soon to expect the drugs to diverge from placebo. Could you comment on this? | Tedeschini et al.'s (2011) pooled analysis of 104 clinical trials revealed that 'Four weeks is the minimum adequate length of a trial in order to reliably detect drug versus placebo differences'.<br><br>Tedeschini E, Fava M, Papakostas GI. Placebo-controlled, antidepressant clinical trials cannot be shortened to less than 4 weeks' duration: a pooled analysis of randomized clinical trials employing a diagnostic odds ratio-based approach. J Clin Psychiatry. 2011 Jan;72(1):98-113.<br><br>Keller et al. commented that 8 weeks might not be sufficient to achieve a full clinical response (p. 770). Similarly it might not be sufficient for ADRs to emerge. | |
| 22. * The methods section should give more information about how subjects were recruited, number of centers involved in the study and how they were chosen. Who did the interviews? How were they trained? You say that children signed an informed consent form, but should this not be "assent?" | As specified in the manuscript, there were 12 study centers (10 in the United States and 2 in Canada). This is now stated in the abstract as well as the Methods section. | Now reads:<br>'An undisclosed number of patients identified by telephone screening as potential participants were subsequently evaluated at the study site **by a senior clinician (psychiatrist or psychologist)**.'<br>The following sentence has been added:<br>'Multiple meetings and teleconferences were held by the sponsoring company with site study investigators to ensure standardization across sites.' |

| | | |
|---|---|---|
| | | We have added: ''The centers were affiliated with either a university or a hospital psychiatry department and had experience with adolescent patients. The investigators were selected for their interest in the study and their ability to recruit study patients.' There was no assent form. We have added:<br>'the study informed consent form was signed by both patient and parent; there is no mention of a separate assent form in the protocol or in the clinical study report.' |
| 23. Please explain how the decision was made to reduce the number of subjects from 300 to 275. | We already explain this:<br>The protocol called for 300 subjects, but this was reduced to 275. Recruitment was slower than expected and, reportedly because of limited medication supplies (mainly placebo) due to expiry, a midcourse evaluation of 189 patients was carried out, without breaking the blind, revealing less variability in HAM-D scores (SD 8) than anticipated. Therefore the recruitment target was reduced on the grounds that it would have no negative impact on the estimated 80% power required to detect a four-point difference between placebo and active drug groups. | See also Naudet, query 8, and Doshi, query 6. Now under sample size as:<br>'Recruitment was slower than expected, and reportedly medication supplies (mainly placebo) were limited due to expiry. Therefore a midcourse evaluation of 189 patients was carried out, without breaking the blind, revealing less variability in HAM-D scores (SD 8) than anticipated. Therefore the recruitment target was reduced to 275 on the grounds that it would have no negative impact on the estimated 80% power required to |

| | | |
|---|---|---|
| | | detect a four-point difference between placebo and active drug groups.' |
| 24. In describing the intervention, please clarify the definition of "non responder." | There is no explicit definition for non-response, just implicit one, considering the definition of response.<br>According to the CSR, section 5.2.4 Sustained Response, page 000078, "Survival analysis was performed for time until sustained response, defined as response lasting until endpoint of the acute phase. Response was defined as a HAM-D total score less than or equal to 8 or a decrease from baseline in HAM-D total score of 50% or greater. Patients were classified as being a responder or non-responder." | Revised to read<br>'Non-responders (those failing to reach responder criteria)...' |
| 25. Although subjects could be titrated up to 60 mg paroxetine or 300 mg imipramine, how many actually did achieve these doses? Can you provide information about the mean final dose in each group and the range? | We have already reported mean final dose and range. We have added number reaching highest dose for imipramine and paroxetine. | Now reads:<br>'The paroxetine group was titrated to a dose of 20mg/day by week 4, with 55% moving to a higher dose (mean 28.0 mg/day, SD 8.4 mg) by week 8. The imipramine group was titrated to 200 mg/day by week 4, with 40% going higher (mean 205.8 mg/day, SD 63.9 mg) by week 8. 28 patients reached the highest permissible dose of 40 mg of paroxetine, and 20 patients were titrated to the maximum 300 mg of imipramine.' |
| 26. * How many subjects were screened for the study? Please show this in Figure 1. | We have not been able to find this information. | Added:<br>'An undisclosed number of patients…' |
| 27. Figure 1 also needs to show the number analyzed for the complete case outcome at 8 weeks. | Displayed in Data Table | Figure 1 |

| | | |
|---|---|---|
| Reviewer: 1 (Florian Naudet) | | |
| 1. comments in the method section and in the results section which are generally not the place to discuss choices and results. Please see for example:- in the introduction : “Consequently, we have reanalysed Study 329 according to the RIAT statement.. To this end, we have used the Clinical Study Report (CSR; GSK's 'Final Clinical Report') available on the GSK website,[7] other publically available documents,[8] and the data access system SAS Solutions OnDemand,[9] on which GSK has posted some Study 329 documents (available only to users approved by GSK). Following negotiation,[10] GSK posted de-identified individual case report forms (CRFs) on that site. A table of sources of data consulted in preparing each part of this paper is available as Appendix 1.” This should appear in the method section; | Agreed | Moved to Methods section |
| 2. - in the method section, authors state “These imipramine doses are high for adolescents. In the six comparator studies submitted by SKB as part of their 1991 Approval NDA for paroxetine in adults, the mean imipramine dose overall was 140mg, with a mean endpoint dose of 170mg” | Agreed | Moved to Results below table 4 |
| 3. - in the method section we can read “(we acknowledge differing opinions about this issue in the statistical literature).” This comment has no reference. | Reference added | Kline RB. Beyond Significance Testing. Statistics Reform in the Behavioral Sciences, 2013, p81. |
| 4. - in the result section “(with a difference of 4 points being pre-specified as clinically significant)” : it is in already in the method | agreed | Deleted |

| | | |
|---|---|---|
| section and should not appear in the results which are descriptive ; | | |
| 5. - in the result section '(Scores on the HAM-D can vary from zero to a maximum of 52)' that should appear in the method section. | agreed | Moved to Methods |
| 6. - in the result section "the protocol also listed the relapse rate in the continuation phase for responders as a secondary outcome variable. Our calculation differed from the CSR calculation because we included those whose HAM-D scores rose above the 'response' range and those who intentionally overdosed." | We think this needs to stay where it is in order to make what follows intelligible | Not changed |
| 7. - in the results section authors states that "alternative treatments of the data could give different results." It must be in the discussion section and not in the results. | No longer relevant as estimates from audit now excluded | deleted |
| 8. - I also think that, for clarity purpose, the information about changes in sample size can be presented after the sample size calculation for clarity purposes. | | Moved and edited for clarity, see Loder, query 23 |
| 9. there were two pre-specified outcome variables, with three groups. Was there a correction for multiple comparisons mentioned in the protocol? These points must be detailed. | No correction | See new box 1 |
| 10. If I understand, there was also a change of primary outcome criteria which was done a posteriori and after breaking the blind. | Yes, but this is discussed in detail in a previous publication, and we think it need not be rehearsed here | No change |
| 11. Can authors give the date of:- Breaking the blind; Changes made in the outcomes criteria | this is discussed in detail in a previous publication, and we think it need not be rehearsed here | No change |
| 12. It would be also helpful to list and compare all the outcomes reported in the published paper by Keller et al. | we think it better to follow Doshi's recommended approach and restrict discussion of Keller to the introduction and discussion | See Doshi section for changes |
| 13. In the sentence: "Global impression scale?" | Agree confusing, we were being obsessional about | change to 'Clinical Global |

| | | |
|---|---|---|
| please suppress the "?" and explain that it is the CGI (as reported in the table). | accuracy, but have changed for clarity | Impression (CGI)' |
| 14. The primary efficacy variable reported in the statistical methods and in the primary outcome variables are not the same. Please explain or correct. | We have rewritten this section, which we agree was confusing | Now reads:<br>'One of the two primary efficacy variables, proportion of responders (response), and one secondary efficacy variable, proportion of patients relapsing, were treated as categorical variables. The second primary efficacy variable, change in total HAM-D score over the acute phase, and the remaining secondary efficacy variables were treated as continuous variables.' |
| 15. In Table 1: please legend (mean [SD]). | Assume this refers to table 3 | Done |
| 16. Figures are represented for OC analysis, please provide the data for W8 (endpoint) ITT analysis with LOCF which was defined as the principal population of analysis. | Under "Patients Valid For The Efficacy Analysis", the Protocol states, "All patients randomized to study treatment and for whom at least one valid post-treatment efficacy evaluation is available will be valid for inclusion in an 'intent-to-treat' analysis." | |
| 16a. Please also indicate the number of patient in each group under the figure for each time point. | This data is too cumbersome for main paper, so have added as an appendix | See Table xiv in Appendix 2. |
| 17. I understand that it is time consuming and difficult, but I think that the analysis of CRF should be complete to avoid any misinterpretation. It is indeed important since this audit process gave rise to additional AEs. Indeed, since this analysis is not complete, and since it was not at random, it is a major limitations and one can be very critic on this point. | See above | |
| 18. In tables where the CRF estimates are | We agree | deleted |

| | | |
|---|---|---|
| presented, I think that this estimates are highly speculative and that the data cannot be analysed in this way. I suggest to delete this column and to analyse all the CRF. | | |
| 19. SAE have a specific definition in MEDRA. I'm not sure that it is strictly overlapping with the notion of severity. Thus the comparison with Keller's et al. paper is very difficult as stated by the authors. For MEDRA, a SAE is serious when it results in death, life-threatening, hospitalization (initial or prolonged), a disability or Permanent Damage, in a congenital Anomaly/Birth Defect, it required Intervention to Prevent Permanent Impairment, and for other Serious (Important Medical Events).<br>This last category is a crucial point and it is probably not stricly overlapping with the notion of severe AE (used by the authors) : it is when the event may jeopardize the patient and may require medical or surgical intervention (treatment) to prevent one of the other outcomes. Examples include allergic brochospasm (a serious problem with breathing) requiring treatment in an emergency room, serious blood dyscrasias (blood disorders) or seizures/convulsions that do not result in hospitalization. The development of drug dependence or drug abuse would also be examples of important medical events. | The problem with SAE as used by Keller et al is that a component of these stems from the judgement of the doctor – we cannot replicate this.<br>Lodging the data with BMJ means that anyone will be able to go in to our spreadsheets and see exactly what was coded and how and will be able to come up with alternate codings.<br>No matter who does the coding, it will be possible for other groups to make a case that in between 1 – 5% of cases that they would have done things slightly differently.<br>This is simply the nature of the beast. Coding is not something you can get right – it is inherently collaborative. | See response to Loder, query 7 |
| 20. When authors state that "The majority of patients stopped at this point were designated by GSK as lack of efficacy (see Table 11). Investigators in four centres reported lack of | We provide the data for others to make their own interpretation. Others are quite welcome to code these as GSK have done. GSK simply don't provide us with a basis for going along with what they have done. Our approach | No change |

| | | |
|---|---|---|
| efficacy as a reason for stopping six placebo patients even though the HAM-D score was in the responder range and as low as 2 or 3 points in some instances." I would like to see more details. Additionaly, I think that the change of coding between Lack of Efficacy and Adverse Event is difficult and could be misleading. Many times, discontinuation occurs for both lack of efficacy and adverse events, since one can easily consider that adverse events like dry mouth can be more acceptable in the case of treatment efficacy. This point could be addressed in the discussion and I'm not sure that a a posteriori interpretation of the CRF can give a perfect information about the individual patient experience (even if it is very better than aggregated data of course…). Moroever, I also think that a lack of efficacy can be considered for patients even if they are responder upon the HDRS. Patients are not just a score on a scale. The authors' a posteriori proposal for recoding this can be thus erroneous. | has more face validity – but could as he says be wrong. We're not afraid to be wrong. | |
| 21. Please explain, in the discussion, for readers that the interpretation of qualitative information in CRF is very subjective and prone to an interpretation bias (including for the first manuscript and for this one). Please explain why it is not possible to collect AE in an otherway (or explain how they should be collected) and the interest of MEDRA. | First interpreting the data on the CRF was essentially blind. There was no indication on the document that indicated which drug was involved.<br>In so far as coding is an act of interpretation, then yes there was interpretation and the risk of bias. This was something that could not be overcome owing to the limitations GSK imposed on us. We could not print off the material and submit it to panels of coders in an effort to reduce bias and we could not convene panels of coders around the periscope.<br>The collection of AE was done 16-20 years ago - not by | |

| | | |
|---|---|---|
| | us. It was done in the usual ways it is done in drug trials then and now. This is a very poor system. It would be possible to design much better systems if you were interested to discover adverse events but this is a different topic and we are stuck with what GSK in fact did | |
| 22. Table 5 can be deleted since it presents results that are also presented in table 6. | | Table 5 moved to Appendix 2. |
| 23. Legend of table 6 is missing (SOC*). | | fixed |
| 24. In table 11, please legend what is "RIAT proposed" ? | | fixed |
| 25. It is stated that "Roughly 1000 pages were missing from the CRFs audited". Can authors precise why? | In some cases GSK state these are missing, in some cases they are simply missing without note; we could detect no pattern to this | Added:<br>'with no discernible pattern to missing information' |
| 26. In the box Patient 00039, please detail wether it was AE or SAE. | This was severe AE – but not serious SAE | Patient 00039, who had a severe (but not serious) AE |
| 27. In the discussion section, when authors state that "The RIAT approach […] outcome variables." It must be recalled that the message is very different since the Keller's report state in the abstract that "Paroxetine demonstrated significantly greater improvement compared with placebo in HAM-D total score < or = 8, HAM-D depressed mood item". | We can't see that anything is being requested here | |
| 28. When they state "In our opinion, statistically significant or not, all relevant primary and secondary outcomes, and harms outcomes, should be explicitly reported". I'm not sure that it is only the opinion of this paper'authors. RCTs are often underpowered for detecting these changes. | We can't see that anything is being requested here | |
| 29. The URL www.xxx is not exactly the good URL… Please do not test… and correct… | This URL is a placeholder until we find out where the documents will be housed | Pending confirmation docs will be housed on BMJ website, and on a |

| | | |
|---|---|---|
| | | dedicated study329 site. |
| 30. Where they state : “They reveal evidence consistent with dependence on and withdrawal from paroxetine.” I would nuance, “with possible dependence”. | We disagree. ‘Consistent with’ is already a qualifier, so adding ‘possible’ would be tautological | |
| Reviewer: 2 (Peter Doshi) | | |
| 1. Organizational issue. I think that in general the authors do not need to mention the Keller et al. publication in the Methods or Results sections of this RIAT manuscript. The misreporting of study 329 in the Keller manuscript has been well documented by the authors elsewhere. The primary purpose of this manuscript, as I see it, is on presenting an honest and accurate report of the study 329 results than it is to further document misreporting of Keller et al. If additional aspects of misreporting in Keller et al. were discovered in the process of RIATing study 329, this is important and I think the authors can include this information, but I think it would be best to keep this to the Introduction and Discussion sections. | Agreed. | Keller references removed from results; modified version included in para 4-6 of *Discussion*. Old tables 6 & 8 incorporated into new table x in *Discussion*. |
| 2. Audit of non-random sample of AEs. The RIAT authors carried out an audit of the adverse event section of case report forms (CRFs) for a non-random sample of 93 of the total 275 trial participants. The authors are very clear throughout the manuscript to indicate that this was a non-random sample. It would have been better of course if 100% of CRFs were audited, but given the number of hours it took to audit 93 (approx. 1000 hours they say in the text), a | Addressed above. Note this reviewer recognised the impracticality of auditing all cases:<br>It would have been better of course if 100% of CRFs were audited, but given the number of hours it took to audit 93 (approx. 1000 hours they say in the text), a full audit likely only will happen if another group picks up the baton. | Changes as detailed above. |

| | | |
|---|---|---|
| full audit likely only will happen if another group picks up the baton. I think the authors are correct to include analyses and tables that show the pre-audit and post-audit tallies of AEs. However I do not think it wise for the authors to extrapolate and present estimates, based on findings from their non-random sample, of the number of additional AEs they would have discovered had they been able to audit all 275 CRFs. (This might be OK if it were a random sample but it is not.) But here in particular, I do not think it wise because my impression of the non-random sample – of all participants that withdrew from the study (85) plus 8 children known to have become suicidal – is that it is a sample more likely to have problems in the transfer of information from CRF to CSR. | | |
| 3. I didn't see a COI statement for the authors in any of the manuscript and appendix files? | We did submit them, but they didn't get into the PDF for some reason | See above; have added COI statements to main manuscript |
| 4. Methods. Can the authors explain why they chose to follow the 1994 protocol instead of the 1993 or 1996 versions? Which version of the protocol was the last version before participant recruitment began in April 1994? Which versions do the authors have the full text for? | | See Loder, query 4 |
| 5. Methods. "Where relevant, we have referred to these variations." What does this mean? | Agree this is confusing and have clarified | Now reads:<br>'Furthermore, the CSR reported some procedures that varied from those specified in the protocol, and we have noted variations wherever they were considered significant.' |
| 6. Methods/Participants. "The protocol called for 300 subjects, but this was reduced to 275." | | See Loder, query 23, Naudet, query 8 |

| | | |
|---|---|---|
| Can this be clarified? So the 1993 protocol called for 300 subjects but this was revised to 275 in the 1994 protocol? | | |
| 7. Methods/sources of harms data. "Roughly 1000 pages were missing from the CRFs audited." Can the authors explain how they knew pages were missing and can conclude this? (e.g. numbered pages indicating missing pages etc.) Were all 93 participants whose CRFs were audited missing the same pages/sections? Also, did they alert GSK to this and if so what was GSK's response? | | See Loder, query 25. |
| 8. Methods/coding of AEs. In the paragraph beginning, "Classifying a problem…" can the authors clarify if MedDRA puts 'sore through' in the central nervous system bucket? | MedDRA has particular problems with sore throat – as any coding system would.<br>There are options for it to go into the infectious, gastric, respiratory and nervous system SOC.<br>We have looked at all instances blind to the study drug and allocated it to nervous system, respiratory and infectious respectively and to the surprise of at least one of us (DH) the results did not pan out as expected – a clear preponderance of nervous system problems on paroxetine. | |
| 9. Box 1. "Most recoding issues were clear-cut." What is meant by 'clear-cut'? | MEDDRA is a more straightforward process less open to bias than using ADECS.<br>in almost all instances the clinical descriptions were sufficiently clear that most coders would come to the same MEDDRA code | Now reads<br>'Most recoding was straightforward. ' |
| 10. Competing interests statement appears missing. The authors say "as attached" but I could not find the attachment. | Clarified above | |
| 11. Methods/analysis of harms data. The authors chose to analyze MedDRA SOC classes psychiatric, cardiovascular, gastrointestinal, respiratory and place all other AEs in "other". | These categories were specifically chosen to correspond with the Keller paper 'Table 3', in order to help with any comparisons. They presented data using the categories: 'Cardio-vascular', 'Digestive', 'Nervous', 'Respiratory' and | |

| | | |
|---|---|---|
| After looking at the results tables, these look like reasonable choices to me, but can the authors include a sentence that explains how they made this choice? | 'Other'. | |
| 12. Methods/patient withdrawal. In the paragraph beginning "The CSR states that the primary reason…" it mentions "CSR Appendix G". Can the authors say here briefly what Appendix G contains? | 329 DEP Appendix G Case Report Form Tabulations by Patient Intent-to-Treat Population [2073 pages]: demographics, presenting conditions, concomitant medication, adverse experiences, vital signs, laboratory data | We think it would add too many words for not enough gain to fully explain what each appendix we refer to contains |
| 13. Methods/blinding. Could the authors also mention whether they reviewed the Certificates of Analysis for the study medications to double-check whether they appeared to have been correctly formulated to ensure blinding? | We have reviewed the Certificates of Analysis for the study. The study pills themselves differed, though all were provided as over-encapsulated bluish green tablets. No information was available regarding blinding success. As described in our manuscript: "Paroxetine was supplied as film-coated, capsule-shaped yellow (10 mg) and pink (20 mg) tablets. Imipramine (50 mg) was bought commercially and supplied as green film-coated round 50mg tablets. 'Paroxetine placebos' matched the paroxetine 20 mg tablets, and 'imipramine placebos' matched the imipramine tablets. All tablets were over-encapsulated in bluish-green capsules to preserve blinding." | |
| 14. Methods/statistical methods. The authors write, "We followed the methodology of the a priori 1994 study protocol." Why is the 1994 protocol labeled "a priori"? Was it the last version prior to participant enrollment? | Correct | |
| 15. Methods/statistical methods. In the paragraph beginning "The primary efficacy variable", there are two sentences with the phrase "primary efficacy variable". I suppose this is a reflection of the trial having two outcomes prespecified as "primary"? | As noted above, we phrased this poorly and have corrected it | See Naudet, query 14 |
| 16. Discussion. Does the following text refer to | This refers primarily to the CSR, which deviated from the | Now reads: |

| | | |
|---|---|---|
| Keller et al. or the CSR: "The authors/sponsors departed from protocol by performing pairwise comparisons of two of the three groups when the omnibus ANOVA showed no significance in either the continuous or dichotomous variables." This should be clarified. If this refers to the CSR, then to some extent there is a discovery among the RIAT authors that they have found reporting bias within the CSR itself, and I think this is an important finding which they should highlight as such. | protocol. This was uncritically accepted by Keller et al. | 'The authors/sponsors departed from their study protocol in the CSR itself by performing pairwise comparisons of two of the three groups when the omnibus ANOVA showed no significance in either the continuous or dichotomous variables.' |
| 17. Box 3. "The inability to access all CRFs may have introduced some error." Not sure what is meant by this. Are the authors talking about their inability due to time/resources to audit everything? Is this a reference to the difficult to use portal for accessing the study data? Or is this a reference to the approximately 1000 pages that were missing from the CRFs that GSK made available through their portal? | Agreed | Now reads:<br>'Time and resources prevented access to all CRFs because of the difficulties in using the portal for accessing the study data and because significant data were missing.' |
| 18. RIATAR. Why are some items so long? For example, so many sources are given for Funding (#25). | Because there are so many potential ambiguities and contradictions we thought it important to disclose all possible sources of data; better to be thorough than readable | |
| 19. Abtract/Results. Suggest changing, if appropriate, "for any measure" to "for any primary or secondary [efficacy] outcome." | Agreed, Changed | Now reads:<br>'The responses to paroxetine and imipramine were not statistically or clinically significantly different from placebo for any pre-specified primary or secondary efficacy outcome.' |
| 20. Background. "RIAT publication of Study 329 which was funded by…" Change to "RIAT | Agreed | Reworded as suggested |

| | | |
|---|---|---|
| publication of Study 329. The original study was funded by…" | | |
| 21. Background. "On 14 June 2013, the RIAT researchers notified GSK that Keller et al. appeared … Study 329." This refers to a letter I sent GSK. We did not specifically mention study 329 in this email. In order to make the sentence accurate, I suggest rewording: "On 14 June 2013, the RIAT researchers asked GSK whether it had any intention to restore any of the trials it sponsored." | Agreed | Reworded as suggested |
| 22. Similarly, change "GSK did not signal any intent to publish a corrected version of the article." to "GSK did not signal any intent to publish a corrected version of any of its trials." | Agreed | Reworded as suggested |
| 23. Methods/Secondary Efficacy Variables. "We could not find any document that provided any scientific rationale for these post-hoc changes…" Did you find any "non-scientific" rationale? If not, perhaps delete "scientific". | We stick by the use of the term scientific, because although it is outside the scope of this paper, the story of 329 is replete with nonscientific (mostly marketing-based) rationales | |
| 24. Methods/Outcomes. The headings 1. Principal Endpoints for Efficacy and 2. Principal Endpoints for Harms. I think this is slightly confusing with the language of "primary" and "secondary" efficacy variables. How about just labeling the sections "Efficacy Endpoints" and "Harms Endpoints"? | Agreed | Reworded as suggested |
| 25. Methods/Harms. I think the "(p. 18)" at the end of the quoted paragraph is a typo as it is also stated above. | Agreed | Reworded as suggested |
| 26. Box 1. "At the week 6 visit … GSK…" Do the authors mean SKB? | Yes | We have altered all references from GSK to SKB where appropriate |

| | | |
|---|---|---|
| 27. A variety of terms are used to represent the provenance of AE data e.g. “CSR recoded” and “CRF audit” from table 7, “AEs in Appendix D” from table 9, and “AEs reported (CSR check)” in table 12. I wonder if better terms can be used to make the meaning more transparent. Perhaps some variant of “SKB/GSK coded”, “RIAT recoded”, and “RIAT recoded plus CRF audit”? Another thought is to use terms like ADECS and MedDRA e.g. “SKB/GSK ADECS coded”, “RIAT MedDRA recoded”, and “RIAT MedDRA recoded plus CRF audit discovered additional AEs”. I realize that some of my proposed titles are long and won’t fit will in the space of a tight table, but my suggestion to remove the Keller columns as well as the “CRF estimated” i.e. extrapolated AE count columns from the Results section will hopefully free up some space. | The proposals are good | Have adopted this terminology in tables |
| 28. Results/Discontinuations. “Consort” should be “CONSORT”. | Corrected | CONSORT |
| 29. Results/Discontinuations. “GSK regarded these patients as participants in the continuation phase…” Should this be SKB? | As above | |
| 30. Box 2/section 8. “… because it became clear that the blind had been broken…” Can you just be clear whose blind you are talking about? I.e. I think this is SKB’s blind, but I’m not 100% sure as part of the RIATers recoding happened blind while other parts did not. | This has been clarified | Now reads:<br>'because it became clear that the blind had been broken in several cases before relatedness was adjudicated by the original investigators' |
| 31. Discussion section/two paragraphs before Conclusion. “… analysis of adverse events requires access to individual patient level data | Agreed | Now reads:<br>'analysis of adverse events requires access to individual patient level |

| | | |
|---|---|---|
| (CRFs)." I would reword the ending to "…requires access to individual patient level data in the form of CRFs." | | data in the form of CRFs.' |
| 32. Conclusion. "Study 329 showed no advantage … on any of the specified parameters." Would using the word "pre-specified" be better than "specified"? | Agreed | Added 'pre-' |
| 33. Methods/Interventions. "These imipramine doses are high for adolescents. In the six comparator studies submitted by SKB as part of their 1991 Approval NDA for paroxetine in adults, the mean imipramine dose overall was 140mg, with a mean endpoint dose of 170mg.[14]" I think this should go to the Discussion section unless it was part of the original methods. | Agree with move | In accordance with Naudet, query 2, moved to Results below Table 4 |
| 34. Methods/Source of harms data. Suggest moving the following to Results: "Of the eleven paroxetine patients with AEs designated as serious, nine discontinued because of AEs. A large number of other patients discontinued because of AEs that were not regarded as serious, or for lack of efficacy or protocol violations (see Figure 1). None of these latter discontinuations led to patient narratives." | This is in fact a methodological issue as it pertains to availability of data. Have clarified that by rewording: | Now reads:<br>'Additional information was available from the summary narratives in the body of the CSR for patients who had AEs that were designated as serious or led to withdrawal. (Of the eleven paroxetine patients with AEs designated as serious, nine discontinued because of AEs.) However, the large number of other patients discontinued because of AEs that were not regarded as serious, or for lack of efficacy or protocol violations (see Figure 1), did not generate patient narratives.' |

| | | |
|---|---|---|
| 35. Box 1 looks like it belongs in Results, not Methods. | Similarly we think this speaks to methodological difficulties | |
| 36. Table 8 is great, but perhaps should go in the Discussion? | While we had thought it might be inappropriate to have tables in the discussion, we are OK with this. | Modified to include comparison of total psychiatric AEs. |
| Reviewer: 3 (Hilde PA van der Aa) | | |
| 1. The authors followed the methodology as stated in the pre-specified protocol of 1994, in which proposed statistical approaches or statistical assumptions were not justified. Outdated techniques were used to analyse the data, leading to more uncertainty. I would recommend authors to (also) include modern techniques of data-analysis or at least mention this 'limitation' in the discussion part of the paper:<br>- One of the limitations of this trial is the large number of dropouts. Therefore, a linear mixed models approach to analyse the data with a maximum likelihood assumption is better suited to estimate effects than the chosen ANOVA and GLM.<br>- If authors, however, do decide to use ANOVA and GLM multiple imputation would be a better way to handle missing data than the currently used LOCF, see for example the paper by Beunckens C, Molenberghs G, Kenward MG. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. Clinical Trial, 2005; 2: 379-86. | | See new box 1 |

| | | |
|---|---|---|
| 2. - Authors described that they did not correct for attrition and non-compliance in the sample size calculation. In addition, they also did not correct for the different strata in their sample (12 centres included). This should also be reported. | <pre>> # HAM-D WEEK 8 LOCF<br>> hamd.lm <- lm(WK8LOCF ~ TRXNAME*SITE, hamd)<br>> anova(hamd.lm)<br>Analysis of Variance Table<br><br>Response: WK8LOCF<br>Df Sum Sq Mean Sq F value Pr(>F)<br>TRXNAME 2 126.5 63.246 1.1887 0.30645<br>SITE 11 969.0 88.094 1.6557 0.08457<br>TRXNAME:SITE 22 884.2 40.189 0.7553 0.77810<br>Residuals 235 12503.8 53.208<br>> hamd.lm <- lm(WK8LOCF ~ TRXNAME+SITE, hamd)<br>> anova(hamd.lm)<br>Analysis of Variance Table<br><br>Response: WK8LOCF<br>Df Sum Sq Mean Sq F value Pr(>F)<br>TRXNAME 2 126.5 63.246 1.2141 0.29868<br>SITE 11 969.0 88.094 1.6911 0.07551<br>Residuals 257 13388.0 52.093<br>> lsmeans(hamd.lm,~TRXNAME)<br>TRXNAME lsmean SE df lower.CL upper.CL<br>IMIPRAMINE -9.082130 0.8020837 257 -10.66162 -7.502636<br>PAROXETINE -10.532636 0.8043016 257 -12.11650 -8.948776<br>PLACEBO -8.945358 0.8204876 257 -10.56109 -7.329623<br><br>Results are averaged over the levels of: SITE<br>Confidence level used: 0.95</pre><br>everything we did takes the effect of the sites into account [LSMean, ANOVA, $X^2$] – see above | in the new box:<br>'The Protocol called for ANOVA testing [GLM] for continuous variables using a model that included the effects of SITE, TREATMENT, and SITE x TREATMENT interaction, with the latter dropped if p>0.10. Logistical Regression [chi Square 2x3] was prescribed for categorical variables under the same model.' |
| 3. Limitations of the current study should be described in more detail. | | See response to Hetrick, query 13 |
| 4. The limitations of the statistical analysis (as mentioned above) should be mentioned. | | See new box 1 |
| 5. the authors state that 'The inability to access all CRFs may have introduced some error.' (page 27, line 25). This should be explained in more detail. | | This clause now deleted |
| 6. At the beginning of the discussion authors state to draw minimal conclusions regarding efficacy and harms, inviting others to offer their own analysis. I think this is a just conclusion based on previously mentioned limitations. However, this cautious approach of interpreting the results of the RIAT study should also be reflected in the conclusion part of the abstract | We believe that our conclusions in both the abstract and the discussion are fully justified by the data that we have presented | |

| | | |
|---|---|---|
| and the discussion. | | |
| 7. Throughout the whole paper authors describe the 'new study' compared to the 'old study' of Keller et al. This makes it difficult to read and to distinguish the methods used in the RIAT trial. Though it is important to report these differences, they might for instance be collected in boxes or reported in italic or combined in the methods section of the paper. | we address this by removing mention of Keller from the results section and simplifying tables | |
| 8. The abstract does not follow the standard style of 'The BMJ' for research articles: objectives, design, setting, participants, intervention, main outcomes, results and conclusion. | Agreed | See revised Abstract, which adheres to this format |
| Reviewer: 4 (Sarah Hetrick) | | |
| 1. It's hard to know exactly what should be in the background, or indeed what he objectives are or how a paper like this should be written up. On one hand it is simply the description of a trial, but on the other hand it has several important other objectives I think: first, to correct errors of the previous write-up; second, to highlight the issue of reporting bias. I am not 100% sure that the second objective was clearly articulated or achieved, and perhaps this is the objective of RIAT but not necessarily of this paper as such. My personal opinion is that more could be made of it in this paper (and perhaps this would address my concerns about originality made above) and that the background appears to indicate that that correcting errors and highlighting the issue of reporting bias is what the paper is about. | We have deliberately downplayed criticism of the Keller et al paper in terms of its reporting bias, partly because as been dealt with elsewhere, but also because we didn't want to distract from the straightforward re-presentation of Study329 according to the RIAT approach | No change made |

| | | |
|---|---|---|
| 2. Should the background include something about letter by Jon Jureidini and Martin Kellers response in 2003? This saw the correction of findings to a certain extent. | We don't think so, for similar reasons and because Keller's response in 2003 corrected nothing but trivialities in the initial reporting | |
| 3. I was interested to know whether the reader should just believe that the Keller 2001 paper was ghost written or whether there is some kind of proof of this? How did the authors find this out/know? | We have documented elsewhere that there is no doubt that this paper was ghostwritten. A reference to this paper is included in our introductory section. | |
| 4. In the fourth paragraph the authors refer to the RIAT statement, but I wasn't clear what this was? | Agree unclear | have changed 'statement' to 'recommendation' |
| 5. In the fifth paragraph the authors outline the objectives of the original study but don't state where these objectives were derived from? The Keller paper, or from the SKB reports? | We have corrected this to make it clear | Now reads:<br>'SKB's stated primary objective' |
| 6. It wasn't clear to me how patients were identified: obviously authors have stated that telephone screening was undertaken, but was this of a particular population? It also wasn't clear what happened during the screening phase that enabled investigators to know that symptoms were stable i.e. was the K-SADS or HAM-D administered twice over and at what time points. Was there a placebo lead-in phase? I think this information should be included. | It is not clear from the CSR or protocol how subjects were identified, or the particular population. In response to another reviewer comment, we have added a statement, which suggests that this was a clinical population (see Loder, query 22). | In the Methods section, after the sentence that ends with, 'A 7 to 10 day screening period was used to obtain past clinical records and to document that the depressive symptoms were stable', we have added the following:<br>'At the end of the screening period, only patients continuing to meet the inclusion criteria (DSM-III-R major depression and the HAM-D total score of 12 or greater) were randomized. There was no placebo lead-in phase.' |
| 7. Again, because the objectives were slightly unclear (or mixed?) I think the write-up is | No further details are available regarding allocation concealment or blinding in CSR or the protocol. | After the statement, 'The blind was to be broken only in the event of a |

| | | |
|---|---|---|
| missing some detail about the methods (if one of the objectives is to publish a sound write-up of this trial). This includes details about how allocation was concealed (i.e. states that patients were assigned treatment numbers in strict sequential order, but where the treatment numbers in sealed opaque envelops?), who was blinded and how i.e. from the write up we can assume that the patient and the person providing the pills to the patient were blinded, but were all the investigators, were the people giving the supportive counseling (who were these), was the statistician doing the analysis? | | serious AE that the investigator felt could not be adequately treated without knowing the identity of the study medication', we have added the following sentence: 'The identity of the study medication was not otherwise disclosed to the investigator or SKB staff associated with the study.' |
| 8. ITT analysis includes all those randomized not all those who receive at least one dose of medication and have at least one post-baseline efficacy assessment. | See also the protocol that defines the ITT for efficacy as we have analyzed it. | See new box 1 |
| 9. I wonder if the authors have thought about undertaking the analysis using more modern and robust methods of imputing the missing data e.g. multiple imputation? I know the authors have indicated that they have provided the data and that therefore others can undertake the analysis as they wish; and that there intentions were to analyse as per the original protocol. But it would be interesting to know what difference a more robust method of imputation makes to the outcomes. In the Cochrane systematic review, undertaking the analysis using LOCF vs OC data made little important difference to the outcomes. | Covered above | See new box 1 |
| 10. I do wonder if the authors should highlight the possible overestimation of the AE figures as | Our analysis unequivocally demonstrates significant harms from paroxetine, and this needs to be included in the | |

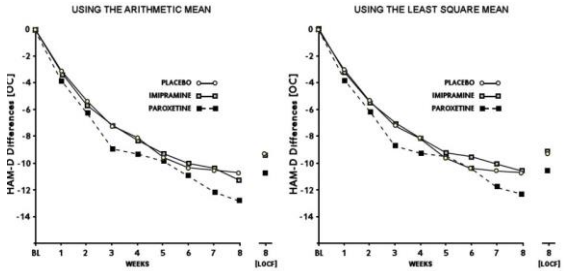
| | | |
|---|---|---|
| a limitation in the discussion and highlight this in the abstract; or I wondered if indeed, given the way in which AEs have been derived and that there is no analysis (and certainly no a priori planned analysis), that this finding should not be stated in the abstract at all. The abstract should perhaps be a clean reporting exactly according to the objectives and pre-planned analysis. | abstract. | |
| 11. Having said that (9), the results with regard to drop outs and AEs is long and complicated and includes long tables with a lot of information; it is hard to know whether readers will take much note or be able to follow it. I think following through each step is important i.e. the author shave tried to do some synthesis by pulling the AE's into groups). Whether further analysis or synthesis could be helpful is unclear; perhaps it is more useful for those undertaking systematic reviews and meta-analyses to think about what to do with this data. | | We have simplified our tables |
| 12. Some of the paper appears not to be finished e.g. there is a question mark after the dot point "Global Impression Scale" (pg 5) and xxx used to indicate some websites (pg 23-25). | See above comments | See Naudet, query 29 |
| 13. In Box 3, authors state that trial participants had relatively chronic depression; which may be true but isn't clearly reported in the results. I'm not entirely convinced that many adolescents have shorter durations of depression. Previous studies suggest that the duration depression might range from 6 to 9 months; but that up to | Keller et al.'s Table 1 reported that the mean duration of the depressive episode for the three groups was about 14 months (14, 14, 13), much more than the 8-week duration specified by the inclusion criteria.<br><br>The reference we cited (Lewinsohn et al 1994) reported a mean duration of 26.4 weeks, with a median of 8 weeks. | Now reads:<br>'The trial duration was only eight weeks. Participants had relatively chronic depression (mean duration more than one year), which would limit the generalizability of the results, particularly to primary |

| | | |
|---|---|---|
| 50% of children and adolescents can still be at 12 months, and 20 to 40% at 24 months (Kovacs, Feinberg et al. 1984; Birmaher, Ryan et al. 1996; Harrington 2001). The trials included in the Cochrane review demonstrated this with a large range of duration of current episode from 10 or 15 weeks to 100 or 108 weeks. | The references cited by Dr Hetrick mainly focused on tertiary clinical samples, e.g. Kovacs et al. (1984): 'Potential cases were accessed through the University of Pittsburgh's child psychiatric outpatient services and the ambulatory medical clinics of the Children's Hospital of Pittsburgh, and via a hospital-based private pediatric group' (p. 230). | care, because many cases of adolescent depression have shorter durations.[26] Generalizability to primary care would also be limited by the fact that participants were recruited via tertiary settings. |
| Reviewer: 5 (Ernest Berry) | | |
| 1. as someone used to the deciphering the world of acronyms in my own sphere, make a heartfelt plea to reduce their scope and volume: they are intimidating to the layperson attempting to understand medical information and trials and often bafffling to patients. | | We think a lay summary is going to be very important. |
| Reviewer: 6 (David henry) | | |
| 1. I believe that if this goes forward the revision should include a retrieval of all of the clinical report forms, masking of the CRFs to remove any clues as to the drug being taken and independent re-coding of the adverse event reports by individuals not previously involved in criticism and re-analysis of this trial. | As discussed above, this is unrealistic. See our comment to editor in relation to Loder, query 12. | |
| 2. I believe that at least one author has appeared on behalf plaintiffs taking legal action against the manufacturer. | Jon Jureidini has been retained as an expert by Baum Hedlund in a class action in relation to prescribing of paroxetine to children. He provided independent advice, and did not appear on behalf of anyone. DH has appeared on behalf of plaintiffs v GSK in adult cases – not pediatric. JLN and MN, who did the coalface work, have not appeared. | |
| 3. Beyond 'setting the record straight', which may be important in its own right, does the re-analysis of the trial contribute to our understanding of the efficacy and safety of | We think that it is evident that correcting the record about what Study329 showed about paroxetine makes an important contribution to our understanding of the efficacy and safety of these drugs. As demonstrated in Box 2, it also | |

| | | |
|---|---|---|
| these drugs in young people? | gives a whole new way to show how companies hide adverse events. | |
| 4. Were the authors the best people to conduct the re-analysis of Trial 329? While overseeing the work, should they have commissioned another group to carry out the more sensitive re-coding of outcomes? | See response to Loder, query 7. | |
| 5. Did the techniques used by the authors guard adequately against bias that might be introduced by their expectations, shaped by their previous experience of this study and related advocacy efforts? | See response to Loder, query 7.<br>We are happy for any bias we might have to be in the full light of day. We think that it actually positive to put on trial here – are scientific articles supposed to be bullet-proof, or there to be shot at? | |
| 6. Do the data and analytical methods support the conclusions of the authors? | If the reviewer thinks that we have failed to support the conclusions, could you please point out instances? | |
| 7. In the light of the current situation will the authors provide a clear indication of how they believe their re-analysis of trial 329 will further inform regulatory decisions and clinical practice? | If our work is taken seriously, we expect it will change perceptions of the clinical literature. At the least, adding a carefully analysed account of a major drug trial published in a major journal will inevitably inform regulatory decisions and clinical practice. | |
| 8. What additional value will this exercise provide – for instance for others performing restoration of other important trials? | As this reviewer points out, our paper maps out adverse event issues that others will need to take into account. We think one of the important messages for others contemplating restoration is the enormity of the workload if one is to go beyond the CSR. | |
| 9. If we accept a need for the re-analysis, have the authors taken adequate steps to manage their professional conflicts of interest? The guarantor of the study has been active over a number of years, has published at least one critique of this study, has corresponded quite vigorously with the authors of the original report, and the journal editor, and has acted on behalf of plaintiffs taking legal action against the | Our data, in so far as GSK allows it, will all be available for others to assess any bias. Also see multiple other comments on similar queries from other reviewers. Finally, the guarantor's negative position on the trial is an effect of his findings, not a cause. | |

| | | |
|---|---|---|
| manufacturer. There is nothing wrong with any of these activities. The concern here is that the authors have adopted such a strong negative position on the drug, and this trial, that they could suffer loss of face if the results of the re-analysis went against their original strongly held position. A number of the decisions that they made required judgements and I am not satisfied that they have taken adequate measures to avoid bias in making these. | | |
| 10. Will the authors provide more detail on the methods of blinding assessors of the written material that required subjective judgments? | The blinding applicable to the efficacy analysis occurred (or didn't) 20 years ago.<br>The task now is judging what is the most appropriate code – once we put all the data into the public domain, we leave ourselves open to criticism – which is a great incentive to come up with a reasonable coding. | |
| 11. How successful was blinding and did they consider asking a group independent of the study team to carry out this work with copies of reports from which key information (such as drug name) had been masked? | Blinding was complete for our initial recoding (99% of all). But in fact the critical blinding is whether the investigator was blind to the treatment being used at the time of the initial determination of an adverse event – this we assume was the case. | |
| 12. With more resources and time could they have retrieved all the clinical report forms in order to reclassify the adverse outcomes? | There will always be a lot of missing pages, and the challenges of having more than one person doing the primary work here should not be underestimated– it is never going to be possible to train up a cadre of people to be certified periscope operators (see Jureidini JN, Nardo JM. Inadequacy of remote desktop interface for independent reanalysis of data from drug trials. BMJ. 2014 Jul 9; doi: 10.1136/bmj.g4353)<br>So having more resources won't really do it – time would get wasted on training and reliability might fall | |
| 13. Regarding the analytical approach to the efficacy data, I understand the logic of what | | See new box and our reporting of pairwise comparisons (Appendix 2 |

| | | |
|---|---|---|
| they propose – which is to carry out ANOVA and only do pairwise analyses if the overall analysis reaches a statistical threshold. I am not a methodologist, but I feel this is excessively conservative in this case with two active treatments being compared with placebo. Based on prior evidence it was possible to construct separate hypotheses for each drug that would justify pairwise comparisons. As the authors say this is controversial, but their stance could add to the impression that they did not start this re-analysis from a position of equipoise. I believe that for the record they should present and interpret these analyses – as readers will try to do them anyway. | | Table i) |
| 14. The main differences revealed by the re-coding and re-analysis of Trial 329 are in the adverse event data. A crucial part of this is the audit of 93 cases with adverse outcomes. To quote the authors "This audit comprised all 85 participants identified in CSR Appendix H who were withdrawn from the study, along with 8 further participants who were known from prior inspection of the CSRs to have become suicidal. " As noted elsewhere this recoding was largely carried out un-blinded and the authors use crude multipliers to estimate possible numbers for the trial populations. As they say this scaling up from a non-representative sample may have over-estimated the numbers for key adverse outcomes. The exercise led to a relatively large increase in CNS adverse events with paroxetine - in particular suicidal ideation and suicide | See notes above; the audit was in fact blinded. As noted above, there is no approach that will elucidate adverse outcomes with complete accuracy, and increasing the apparent sophistication of the methodology may only mask the inadequacies of the ultimate outcome, subject as they are to misrecording and misinterpretation, for reasons that include unwitting bias. | extrapolations deleted |

| | | |
|---|---|---|
| attempt – plus depression worsening and aggression. As this is arguably the key finding of the re-analysis I think greater efforts should have been made to obtain all the CRFs and to mask them by manually screening out drug names or other clues in the text and to have the recoding carried out by people who were not involved directly in the study and had not been involved in previous efforts by the authors to discredit the trial. | | |
| 15. Considering efficacy, I think the authors should present their re-analyses alongside the originals. For the primary outcomes, as far as I can see for the HAMD >50% drop or <8 the authors results for imipramine and placebo are identical to those in the original report by Keller et al. However, the proportional response with paroxetine in the re-analysis (65.6%) in the LOCF analysis is slightly lower than the original figure (66.7%). They may have reclassified a responder – can that be clarified? | We have resolved this and it was our error. The subject in question got off schedule on week 5 [no value]. In our original analysis, we had NO for responder [HAMDRESP] at week 8. GSK had YES. That was the discrepancy it took so long to find. So this subject had a response in the waning hours of the Acute Study. We had coded NO because we went back and whatever algorithm SAS used to assign weeks, it was consistent, and it always picks the latest value in the assigned week. In all other ambiguous cases, our resolution was the same as theirs. | Figure altered accordingly.<br>See also the new box. |
| 16. For the drop in HAMD scores the authors have presented the LSMeans from their modeling, whereas I think the original paper presents the differences in arithmetic means. The differences are small and probably don't affect the overall conclusion, but this point should be clarified. | The outcome is the same whether LS or arithmetic means are used:<br>  <br><br>The Statistical methods APPENDIX A clearly states that all means are LSMeans. | Have noted in legend to Table 3: 'Using arithmetic means did not alter the findings.' |

| | | |
|---|---|---|
| 17. The authors have presented the adverse event data in a series of tables. These are quite clear (except for doubts about the total estimated numbers from the incomplete audit). However it feels like they are scattered across several tables. I think the authors should try to produce a summary table where the major outcomes – efficacy and adverse events are summarized from the original trial report and from their re-analyses are presented. | We have simplified our tables, but we believe that further attempts to integrate them will make them more difficult to comprehend. | - |
| 18. The overall report is long – 127 pages, of which 32 constitute the main trial report. The remainder can be handled as supplementary material and as the authors state may be valuable to others. But the main part of the report is quite long and tends to editorialise in almost every section. I think a more tightly written report that sticks to a description of what was done, what was found and how the findings differ from the original would be more readable | We have written the shortest paper that we have been able to. | - |