Dear Dr Loder

Thank you for the exhaustive editorial and peer review of our important paper. It is the better for it. We have completed our response to reviewers and revised the manuscript accordingly. You will find that we have complied with most that you and the reviewers have requested of us.

With regard to CRFs, we agree that we should not have reported any inferences beyond those forms that we have actually examined, and have corrected this. We have also realised that the term "audit" is misleading, and we no longer use it.

Mindful of the difficulties for even a sophisticated set of reviewers to apprehend the unique challenges posed by our task, we would like to offer that Drs Nardo, LeNoury and Healy meet with you in person to demonstrate the operation of the portal, show the Study 329 Rewrite website, and the fact that it contains many of the extra details that the reviewers have asked for regarding the background to the original study, and demonstrate how the data from the appendices is transcribed into our own spreadsheets for the BMJ and/or Study 329 Rewrite website.

There remain three points of potential disagreement: the use of "more modern techniques" to analyse efficacy outcomes; the desirability of completing the analysis of CRF's; and whether we took adequate steps to ensure independence in analysing adverse events. We defend our position on these three points below, and request that our defence be published as part of our response to the reviewers alongside the reviewers' reports.

1. *Efficacy analysis*

When it comes to efficacy, the point behind trials and statistical testing is to thoroughly test a manufacturer's claims because the well-being of vulnerable people is at stake. Trials are designed to weed out bogus claims.

The contract under which GSK provided us with access to data specified that we were to follow the SKB protocol. This seemed appropriate. For a trial to be considered misleading, it should be by the standards that the initial researchers worked under, not "by modern standards". And we wished to avoid real or perceived bias in the way that we analysed the data.

Using the protocol methodology, we could find no hint of efficacy. It is not our place to adopt ever more sophisticated methods to find hints of efficacy when such hints could have highly dangerous consequences. (If only to defend themselves against a New York State Fraud charge or save themselves a $3 billion Dept of Justice fine as reported by BMJ in 2012, if more "modern" methods of data imputation could have in any way retrieved this study, one imagines GSK would have done so.)

When we started in 2013, we had only the data tables published on the Internet. Analysis involved first transcribing those numbers into the computer applying the various algorithms spread throughout the Full Study Report Acute. We did multiple passes by hand and by using OCR software.

When we later gained access to the GSK Secure Portal, we started over using the provided csv [Comma Separated Values] files, Open Office Spreadsheets, and the R Statistical package on GSK's remote desktop. While that may seem easier, it was much harder primarily because of the small "space" to open the number of spreadsheets required to transform the data into files that were suitable for the R analysis. Matching the algorithms was equally difficult and, again, multiple passes were involved. That is not a complaint. As difficult as the process was, it meant that we actually looked at the data repeatedly rather than simply undertook a pre-programmed extraction by the SAS software. (We think that Dr Nardo is the first clinician to actually look at those numbers. The named authors of the original article did not. And we think Rosemary Oakes and Jim McCafferty are in that same group. We know Sally Laden didn't look at them from her deposition, and Martin Keller doesn't "like numbers".)

While we began using more modern methods of imputation and statistical analysis in our initial take on the data before access to the portal and GSK's analytic programmes, over time and with much back and forth, we ended up deciding that the choice of analytic technique was a potential source of bias (our own bias), and that in the efficacy analysis, we should wherever possible stick to the methodology prescribed by the original SKB *a priori* protocol. Anyone can pull out different techniques *post hoc* and come up with an analysis that favors some more desired outcome, and we didn't want to do that. We appreciated the reviewer's comments, but the ones suggesting "other" methods might have been helpful for an initial submission of a clinical trial, but this is something else and those suggestions seemed somewhat off the point of what we were attempting to do in a RIAT reanalysis of an already published study.

One BMJ reviewer helpfully found a discrepancy between the original publication and our submission that ultimately pointed to an algorithm we had simply missed in the volume of pages: because adolescents don't show up like clockwork, there were many instances where there were two values assigned to a given week as they appeared "off schedule." The value picked was always the later value. We had seen that rule in the CSR, and done it in our original analysis, but we missed that for week 8 (the one that mattered most), that window was specified as running for 17 days instead of 7, from day 53 to day 70. We had used days 53-60. So we repeated the whole data extraction with multiple passes under GSK's algorithm, and the corrected files are in the resubmission.

In addition, on the complete redo we found something else. We had originally found and successfully corrected a "glitch" in a csv file that threw things well out of kilter (an errant carriage return can play much havoc in csv files). But on the redo, we found that more subtle glitches had done the same thing in three (of the seven) other files, again successfully repaired.

One other "algorithm" point. The K-SADS-L data was collected every other week, so many values were dropped because they were in odd weeks, making the analysis of those data questionable. Some LOCF values use week two data carried forward when there are many

later values dropped from the study because they came in odd weeks.

So in our resubmission, we have added a box making our choice of analytic methods explicit. If it was in their *a priori* Protocol we did it. If it was unclear or absent in the Protocol, we chose the most standard tests (and we made those choices *before* doing the analyses). For example, as we say, we did not agree that showing only pairwise comparisons rather than the omnibus ANOVA was justified. That was not in the Protocol or in our reference materials. But in response to the reviewers' comments, we added a table to the Appendix with pairwise values.

The Protocol doesn't specify a method for doing pairwise comparisons. R language has a function for just that purpose [pairwise.t.test]. The alternative was to do a full ANOVA analysis on each pair. We chose the latter for two reasons. First, the pairwise function was built to correct for multiple tests, and we would've had to turn off that correction and standard deviation pooling to fit SKB's ignoring correction for multiple variables. But more importantly, it didn't take the Effect of SITE into account, an integral part of the SKB specification. So we chose the ANOVA. It gave slightly different results, more favorable towards the SKB analysis, and actually put the HAM-D RESPONSE into the significant range in the Observed Case [OC], we learned after running the numbers later both ways. But we believe we are bound by the same *a priori* rule as SKB/GSK and stuck with the ANOVA.

In the published paper, SKB added four outcome variables a few months before breaking the blind. They were the only statistically significant variables. They averred that because they were added before the blind, they were equal to the *a priori* declared outcome variables and in the public subpoenaed documents, they claimed that they didn't need correction because they were mentioned before *the blind was broken*. First, GSK could provide no evidence of a "plan of analysis." Further, there's nothing about correction for multiple variables that is time dependent. And finally, at best the late comers could only be considered "exploratory variables," not the essence of a claim of efficacy. As we say in the paper, these reported add-ons can be easily refuted. Any rational correction for multiple outcome variables evaporates significance except for the HAM-D Depressed Mood item. We did not consider them in our analysis as explained in the new box, and in this case as in others we did not apply needed corrections to stick to the original Protocol and because it would be a *post hoc* technique chosen "after the results known," and we didn't want to do that in this RIAT reanalysis. The point has been thoroughly covered in previous publications.

2. *Examining the remaining CRFs*

While, as noted above, we have come to fully accept the reviewers' suggestion that we remove the extrapolated figures, we think it better that you do not require us to complete an analysis of the CRFs. When this Rewrite began, we had no expectation that we would get access to the CRFs. As our correspondence with GSK posted in your rapid responses shows, GSK were initially resistant to making the CRFs available. We did negotiate access to Appendix H (77,000 pages of CRFs, compared to approximately 5,500 pages in appendices A-G combined).

But this falls very far short of giving decent access to the data. In addition to the sheer volume, the conditions under which GSK granted access were so restrictive that GSK's expectation may have been that we would be doing something similar to what FDA do when

they audit the books, no more than dip into the occasional record to confirm that, for instance, individual patients existed.

We may have used the word audit because FDA's review of records like this is somewhat like the review undertaken by a financial auditor – a one in twelve or one in twenty or less sampling. In this study we believe FDA audited the records of 12 of the 275 patients. No one knows how thorough their review was – the primary task is to confirm the patient exists. What we undertook was more like a Data Safety Monitoring Board function – reviewing in detail 34% of the sample, 93 records, 8.5 times the number of records FDA reviewed.

We may not have been sufficiently explicit about the work required to examine in detail even this many CRFs. What we have done has been heroic. It isn't possible to train up a team of independent periscopologists to do what has been done, and anyway GSK were only prepared to grant a limited number of licenses in the first instance. Even if reaching out to some others had been possible it is not clear that anyone would have been prepared to give over 6 months of their life to a thankless, monotonous, unpaid task. In the course of doing some of this work, one of us lost not just a researcher who couldn't take any more of it but also because of current financial constraints an entire research post and its associated funding.

As events have transpired, our RIAT article has come to be about more than restoring Study 329. Once GSK granted access to the CRFs, the article has something very important to say about data access. It is unclear whether such access will ever be repeated; there is no commitment on the part of GSK to do this again for other groups.

Apparently "completing" the evaluation of AEs direct from the CRFs risks doing a number of things.

1. giving the impression that the dataset was complete when there were at least 1000 pages missing in the one third of the records we have reviewed
2. giving the impression that using GSK's periscope is a reasonable approach to data access. In fact, neither we nor our readers really do have access to the CRFs in a meaningful way when the periscope means that it takes about 10 hours to analyse each record form
3. distracting from what may be a major contribution to scientific debate – that there is no such thing as complete analysis, and conclusions from a trial are provisional and subject to improvement by others having equal access to the data.

It emerges from this research that a scientific article should always be a work in progress rather than be expected to authoritatively settle debate. Leaving the CRF analysis incomplete forces our preferred interpretation on the reader. We believe it is important that the adverse event profile of both paroxetine and ultra-high dose imipramine, never before or since subject to valid testing in children, is provisional. We want to invite others, including GSK, to engage with this study.

*3. Potential Bias in Coding*

As noted in our submitted paper, the original protocol for Study 329 makes no mention of how AEs from this trial would be coded. Of the total of 1411 AEs, we were in fact blind when it came to MedDRA coding in all bar six (0.005%) of these AEs.

But blind coding is irrelevant. The blinding that counts is whether the clinician was blind to the drug the child was on when s/he deemed that child to be having an AE and used the clinical descriptors that now appear on the records.

After that blind act, GSK coded these events. They may not have coded them blind. We used a much better coding system and coded blind. We did so because we anticipated the lack of understanding of readers who were not familiar with coding – we did not do so because the paper was methodologically stronger as a result of coding blind.

GSK may have been biased when coding. We may have been also. But coding is not something that is right because it is done by unbiased people. It benefits from having as many people as possible have an input and can be helped in some instances by people coming to the data with a bias, including the bias of knowing what the drug is.

It would be feasible for GSK after publication to get MedDRA trained coders to look at the codes we have applied and they would almost certainly in some instances be able to propose codes they will claim are preferable to the ones we have used. Even though their exercise in this area may not be undertaken blind and might be highly biased, we expect that in a small number of instances, if they offer the basis for their suggestions, especially if they can appeal to extra materials written in the margins of a CRF for instance, we might concede that their choices improve on some of ours.

Across all codings, we would expect disinterested coders to rate our efforts more highly than GSKs. We have certainly done better than GSK did originally in this study where there are some very clear breaches of good coding practice. We also think that even if there are a small number of events where GSK's suggested coding improves on ours, this is not likely to make a meaningful difference to the overall profile of adverse events.

But the key point is this. If GSK engage in such an exercise, they will demonstrate the benefits of data access. Once there is data access, there is nothing to be gained by investigators (in this case us) being biased. We have a huge incentive to be genuine.

We expect others having access to the data might find that we have made mistakes. Where BMJ might not welcome GSK or surrogates drawing attention to mistakes or biases, we would in fact welcome it. This is what science is supposed to be. We do not expect it to reveal systematic bias but if it does, that will be interesting in its own right.

If you publish this covering letter written in response to the reviews, GSK will be put in an interesting position. If they don't already realise it, every effort on their part to draw attention to mistakes we have made on the basis of publicly available data that others can view will draw more attention to the benefits of data access and the role of companies in denying access.

It is important to be open about what this means for BMJ. We acknowledge your anxiety that once issues have been raised by reviewers, publication of these reviews might leave you

vulnerable, but we do not think there is any way for BMJ to be bullet proof on the issue of adverse events.

We urge you nevertheless to publish the reviews as they stand and to wait and see if GSK (who are after all in the best position to carry out all kinds of analyses) respond by adopting the analyses proposed by the reviewers, and if so, what the outcomes are and what the scientific community would make of anything GSK offered in this area.

In conclusion, we offer you a heavily thought out attempt that may provide a basis for setting a first set of standards for future RIAT efforts. In the case of an already published article, RIAT is not intended to be a conduit for criticism and bickering, but rather a serious and thorough analysis of the results of a study in a manner that aims at opening up rather than closing down debate.