_____
From: em.jaacap.0.4acd65.1d6fed7f@editorialmanager.com <em.jaacap.0.4acd65.1d6fed7f@editorialmanager.com> on behalf of JAACAP <em@editorialmanager.com>
Sent: Thursday, 28 April 2016 7:04 AM
To: Jureidini, Jon (Health)
Subject: JAACAP-D-16-00138 Decision

CC: kpumphrey@jaacap.org

****
In This Email:
1. Editor's Decision Letter
2. Reviewer Comments
****
Apr 27, 2016

RE: JAACAP-D-16-00138, "Study 329 Continuation Phase: Safety and Efficacy of Paroxetine and Imipramine in Extended Treatment of Adolescent Major Depression"

Dear Dr. Jureidini,

Thank you for the opportunity to consider your manuscript. I append below the comments from five peer reviewers. I regret that based on these critiques, and on my own careful consideration of your paper, I cannot accept it for publication.

I certainly hope that the feedback that follows is useful as you consider next steps and seek an alternative outlet for your findings, and that this decision will not dissuade you from submitting again to the Journal.

Please accept my best wishes for success in your future work.

Sincerely,
Andrés Martin, MD, MPH
Editor-in-Chief
Journal of the American Academy of Child and Adolescent Psychiatry

****
Reviewer Comments:

Reviewer #1:
The current report describes the outcome of subjects originally participating in a well-known, controversial randomized controlled trial (RCT) comparing the efficacy of imipramine, paroxetine, and placebo. As with much of the literature surrounding prior publications on this study, it is difficult to separate the contribution to clinical science, on the one hand, from the contribution to the literature on the history of psychopharmacology, on the other. Clearly, the current report has some contributions to add for both topics, but the contribution is probably more substantial in terms of historical significance than current clinical thinking. This is due to major flaws with the data reported here.

In general, I do think that this paper is well written, and I also think that it provides meaningful additions to the literature. The data are clearly novel. However, as noted above, there are also pretty significant problems with the data that create major issues for the paper. One major problem here is that the participation rates in this post-randomization phase are really unacceptably low; moreover, the reasons for discontinuation are somewhat unusual (e.g., "...shortage of study medication supplies...") and applicable to an unclear proportion of the subjects (e.g., "some"). Another major issue is that the main efficacy measure, the HAM-D, has major flaws that have led it to be avoided as a primary efficacy measure in studies examining the treatment of adolescent depression over the past 15 years. With a few relatively modest exceptions, I do think that the authors generally handle these major flaws in a reasonable way. They are flaws nonetheless that diminish this reviewer's enthusiasm for the report and that limit the conclusions that one can draw about the study.

Beyond these major issues, I will list a series of minor issues in the order in which they appear in the paper.

1. Abstract: In describing the trial, the introduction of the abstract does not clearly separate descriptions and data related to the randomized controlled phase of the study, reported previously, from the continuation phase, reported here. Due to high and possibly non-random dropout, the continuation phase has "lost" random assignment. This is not stated with sufficient clarity. Moreover, due to this flaw in the original design, the study cannot speak to efficacy, as is incorrectly implied in the discussion of the abstract.

2. Please spell out CSR on page 3, line 9.

3. The discussion of procedures for tapering on page 5 raises significant concerns. In light of lack of clarity on these procedures, it is possible that inappropriately rapid tapering could have contributed to adverse event profiles in the two active groups. This point should be made in the discussion.

4. The data analytic approach seems outdated. Procedures that utilize so-called "mixed models" should be employed to more appropriately handle missing data. Moreover, I would like to see an analysis of continuous HAM-D scores over time; the categorical definitions of outcomes create many problems here. I am fine with the reporting of the categorical outcomes, as they currently appear. However, these category-based analyses should be supplemented with appropriate analyses of continuous variables. Of note, the slope in all three groups in Figure 1 after the acute phase looks to be zero. This suggests a maintenance of clinical effects in all three groups. Of course, this could reflect the non-random entry of subjects predisposed to a "good outcome", a point that should be raised in the discussion. Nonetheless, the low level of symptoms here, relatively to findings in more recent studies examining maintenance of antidepressant response is worthy of some comment.

5. Page 17, last paragraph. There are problems with the discussion of the discontinuation design. It is not clear the degree to which appropriate procedures were followed for a discontinuation design study here. As a result, the question remains open about whether or not this is the most appropriate future step forward. Other studies, admittedly done with long half-life antidepressants like fluoxetine, suggest that the discontinuation design has major advantages. The FDA also has been a major advocate of this design. As a result, the discussion here should be more nuanced.

Reviewer #2:
This is a RIAT (restoring invisible and abandoned trials [Doshi et al 2013]) report on a continuation phase study of two antidepressant medications. The goals of RIAT (although this reviewer could find no external review of the unique value of the RIAT methodology) are stated as improving the conduct of trials and the quality of information gleaned from clinical research, and correct bias in the published literature. This publication attempts to address selected problems with an unreported target study. However, the presentation is incomplete; confusing and ambiguous; and overall fails to reach the RIAT goal of advancing scientific knowledge.

In the introduction (page 2, lines 7-15) the authors write that "In 2013, in response to concerns about selective reporting of outcomes of randomized controlled trials, an international group of researchers called on funders and investigators of abandoned (unpublished) or misreported trials to publish undisclosed outcomes or correct misleading publications. [1]" The goal of a RIAT publication should be to break the cycle of selective and misleading evidence in the medical literature. However, the 329 Continuation Phase (329-CP) study suffers critical design limitations that prevent any meaningful inference from the collected data. Moreover, this RIAT report fails to adequately follow RIAT publication guidelines that further obfuscate a reasonable understanding of the 329-CP results and this RIAT-initiative report does not help clinicians, researchers or the public better understand the risks or benefits of paroxetine and imipramine in the extended treatment of adolescent
major depression.

(A). Sample Size and Statistical power considerations

1). The sample size of 49 (paroxetine), 39 (imipramine) and 31 (placebo) is rather small (and the sample

size difference between paroxetine and placebo of 40% is rather large) to enter into a continuation phase RCT and have sufficient power to detect differences in relapse rates; which makes the paper limited in impact regardless of the outcome, especially when there are no preliminary data or theoretical basis to expect a difference in preventing relapse between paroxetine and imipramine. The authors should include a flow diagram starting with the number of people randomized in the initial trial, and indicate the flow of patients through (and out from) the various stages of the trial, continuation phase, and taper phase. Also, each table and figure should include sample size information sufficient for the reader to understand which of the samples indicated in the flow diagram are being described (or ideally work back to the original N randomized).

2). Both RIAT and the CONSORT guidelines require the documentation of sample size determination in a manuscript reporting on a trial. The RIAT guideline checklist included with this manuscript indicates that information is on page 5, but I do not see it there. For this report, what is needed is a recitation of the rationale for choosing the sample size for this 329-CP study. As a RIAT republication or rescue document, this report is an opportunity for describing whether the sample size is adequate for addressing the aims, goals or hypotheses of the 329-CP study. RIAT-initiative authors have an opportunity -- and responsibility -- to critique and provide guidance to researchers and industry on the importance of adequate statistical power, and to remind clinicians the extent to which inferences can be drawn from null results from under-powered studies, and confidence in parameter estimates derived from small sample size studies. To address this limitation, the authors should:
(a) explicitly describe the sample size determination (if any) for the 329-CP study and the parent study, and
(b) provide confidence intervals for all inferential statistics (e.g., proportions).
The authors combine the adverse events from the continuation and taper phases, which is problematic given the differences in design of these two phases.

Limited and selected inferential statistics:
(A) This RIAT report suffers from a logical inconsistency. If it is reasonable to review and improve upon the classification of AEs from original study documents, it should also be reasonable to improve upon other and all aspects of the 329-CP study, including the statistical data analysis of the CP study results. It is potentially misleading to report selected and limited statistics. The authors can address this limitation by developing and conducting a rigorous data analysis of the data they have collected, one that uses methods capable of addressing the selection effects to the CP phase and provides confidence intervals for implied effects of drug treatment.

(B) The GLM/ANOVA analysis reported in the methods (page 11 line 32-37: We applied ANOVA testing (generalized linear model) using a model including effects of site, treatment, and site × treatment interaction as per the Study 329 protocol). This may have been the study 329 protocol, but the handling of time is ambiguous. As from the figure time-wise results are presented, one might guess (I should not have to guess) that the analyses were repeated at each time point. This is an extremely inefficient use of the repeated measures data. More sophisticated methods are called for.

(3) No support for conclusions drawn in this RIAT report:
The authors make inferences that the 329 project CSR was not prepared to. For example:

(page 1 line 21) "The continuation phase did not offer support for longer-term efficacy of either paroxetine or imipramine"

(page 1 line 24) "Relapse and adverse events on both active drugs open up the risks of a prescribing cascade"

(page 1 line 26) "The previously largely unrecognized hazards of the taper phase ..."

This report does not provide evidence that there are excess hazards associated with the taper phase of the active treatment arms versus the control arm. In the absence of confidence intervals on estimated effects and test of hypotheses, the results are ambiguous with respect to long-term efficacy of drugs, rate of relapse and adverse events (AEs), and other hazards of drug administration.

(4) Fundamental problems with 329-CP as originally conceived propagated by this report.
The 329-CP is not a randomized clinical trial. CP data are based on people identified as "responders" in the

parent RCT. This stratification on a post-treatment factor removes the strength and basis for causal inference inherit when assignment is randomized. Moreover, because being a "responder" is likely an outcome of both drug treatment and initial severity, even if there is no association of drug treatment and initial severity in the population, in this selected sample we can induce such an association. This is called selection bias or collider bias.

Under the assumption that persons with more severe depression at baseline are more likely to be classified as a responder, we would expect a positive correlation between drug treatment and depression severity among the responder sub-sample. Assuming AEs are more common (or severe) among persons with more severe baseline depression, we would therefore expect to see a positive correlation between AEs and drug treatment. But this is spurious and driven by the selection bias inherent in studying only responders.

Under the assumption that persons with more severe depression at baseline are less likely to be classified as a responder, we would expect a negative correlation between drug treatment and depression severity among the responder sub-sample. Assuming AEs are more common (or severe) among persons with more severe baseline depression, we would therefore expect to see a negative correlation between AEs and drug treatment. But, this also is spurious and driven by the selection bias inherent in studying only responders.

Bottom line is the apparent correlation between drug use and AEs could be due to bias inherent in the study design. A sophisticated data analysis plan is needed to answer this question.

Efficacy Endpoints:
5). Concerning the analysis of relapse, there are other ambiguities that cloud interpretation of the data. Specifically, evaluable data for determining relapse within each of the three arms should be based on the six-month observation of participants who enter the continuation as "responders." Exactly who these participants are, who enters the continuation via what inclusion criteria and when, is not readily apparent. In discussing the Taper Phase (see ms. page 5), they: (a), remark on participants who did not agree to a taper phase; (b), for some they impute length of the taper phase, but don't state why length of the taper is crucial for the analysis of relapse, nor do they give any justification for having "assumed an average taper phase of 2 weeks."(c), they include in the taper phase those who taper in the acute phase as well as the continuation phase, but we don't know if these are participants who have relapsed, who are dropped due to non-compliance, or who no longer
wish to participate. Which, and how many, participants actually provide evaluable data for a continuation phase study analysis of relapse is impossible to discern. What we can surmise from one of the tables is that the numbers available are too few for any meaningful statistical estimate of continuation efficacy, which the authors acknowledge.

6). The protocol defined response in two ways: a 50% reduction of HAM-D score from baseline, and, more restrictively, a HAM-D score no greater than 8. However, it appears the authors' chose to define relapse using the more conservative definition only; their rationale for doing so is not clear. Nor do the authors explain why their application of a more "conservative remission criteria for their analyses of relapse will reduce the number of relapses." But more serious is that they also added to the relapse definition a participant who dropped, or was dropped, from the protocol following a "suicide-related event," even when the preceding HAM-D score was 8 or under. The rationale for amending the relapse criterion is rather arbitrary and not substantiated. Moreover, it is well known that such events can arise in youth for any number of reasons; a loss of drug effect is surely one, but for clarity a separate classification for these drops, or relapses, should be separately
coded. Such a multi-level classification then allows for a reporting of multiple, continuation phase outcomes, one including protocol drops, another restricted to the return of threshold level symptoms after acute phase response Also unclear is if the authors classified a participant who develops symptoms during the taper as a continuation phase relapse.

7). Page 6-Suicide related events-It is well known in the field that the definition of "a suicide related event" is both complex and controversial, yet the authors make no attempt to give their definition or methodology of determining suicide related events. Coding of adverse events---page 10- "The main coding challenges arose in relation to suicide related events in the acute phase; these are covered in Le Noury et al [3]." Given the importance of suicide related events, to refer the readers to another manuscript to learn how these authors created their definition of such events in this manuscript make it very difficult for the reader to

elucidate what they have done in this manuscript.

8). The authors state that "The protocol called for the percentage of patients withdrawing because of lack of efficacy to be evaluated once at the end of the continuation phase for each patient. We have included in this category those patients whose final HAM-D scores were consistent with a lack of efficacy, even if the stated reason for withdrawal was non-compliance or protocol violation or adverse events other than suicide-related events." This is very problematic given that the definition of non-compliance in this study was taking less than 80% or more than 120% of study capsules- and a patient missing two consecutive sessions. If patients are not taking the study medication that should not be considered lack of efficacy. Another ambiguity is found on ms. page 7, where they discuss withdrawals due to lack of efficacy, based on a one-time evaluation at the end of the continuation phase. It is unclear if this includes all observations of inefficacy; specifically, does this
mean withdrawals observed during the acute phase as well? I am unclear how these data are considered in the analysis.

9). Analysis of safety data-Page 11. "As the acute and continuation phases are of very different duration, and a significant number of patients dropped out in the course of the continuation phase, a simple listing of the adverse events from each phase risks misleading. We have therefore presented the total number of events but also estimated the rate at which events occurred by duration of exposure"---- There is no scientific basis (or explanation given) for how the authors estimated the rate at which events occurred by duration of exposure

10). Table 1 is described as depicting attrition due to non-response, drop-out, and relapse. This is not readily understood by reading the table headings. Table 2, and tables 4-13, overwhelm the reader with data; they are best provided to interested reader upon request.

11). Table 3 is difficult to interpret. First, no rationale is given for reporting results of the acute phase; how they derive a measure of relapse during the acute phase is unclear as the main outcome is response/lack of response. Perhaps they are coding as relapse an acute phase score less than 50% of baseline (or a HAM-D of 8 or less), but this is unorthodox given the accepted meaning of relapse in clinical trial research.

Reviewer #3:
The authors are to be commended for undertaking such a laborious task to make the results available for the extension of what started as a large RCT. Unfortunately, the sample was considerably smaller by the end of the extension. Still, something useful may be gleaned.

You make the point that the data may be analyzed in different ways and interpreted differently. In fact, you respect that point so much that you say you will just present the facts and let readers interpret them. However, the mass of data (13 tables and 3 graphs) is so daunting that it is unlikely most readers would probe deeply enough. It seems desirable to offer them some contextual guidance to aid interpretation.

You implicitly recognize the fact that the original acute trial is an important part of the context by including some elements of it (e.g., taper phase AEs). You also published a previous article in which you purportedly debunked the acute trial, drawing exactly the opposite conclusions from the original authors: not efficacious, with an increase in harms rather than well-tolerated and effective. This amazing feat set me to reading the original article and your re-analysis of the acute trial as well as the exchange of letters, besides the current manuscript. That homework brought the following realizations:

1. The conclusions hinge on what aspect of the data one attends to. E.g., The original authors chose to disregard their originally chosen 2 primary outcomes and instead attend to 4 secondary outcomes that most clinicians would agree validly reflect depression outcome, concluding effectiveness (poor science but clinical common sense). You chose to disregard those secondary outcomes that showed benefit even though one of them (Ham-D <8) was closely related to one of the original primary outcomes and another (CGI-I) was one of the pre-specified secondary outcomes, and instead rigorously followed the primary outcomes specified in the protocol (accurate science, but clinically obtuse).

2. You were critical of the original report well before you saw the data and you set out to analyze it as a

misreported trial; in other words, you were on a mission. This is OK in itself; clinical science advances by having open debates about alternative interpretations, which can generate testable hypotheses, but:

3. Both you and the original authors appear to have conflicts of interest that may have colored your choice of which data to attend to. They may have wanted to report a positive trial and appear to have had some research funding from the manufacturer of one of the drugs. You, on the other hand, profited (and continue to profit) from forensic consulting fees. It should be noted that any financial benefit the original authors had is long gone and was probably gone by the time of their publication (although professional pride may still be an incentive to defend the original conclusions), while you stand to profit in the future from more forensic work.

4. Your interpretation of the safety data benefits from hindsight but risks the hazards of second-guessing the judgment of clinicians who were on the scene. Since the trial was carried out, we now have a better appreciation of suicide risk than was extant then, and you were able to apply that updated knowledge. On the other hand, it is risky (and not necessarily the best science) to reject what the original clinicians decided about attribution based solely on paper records without seeing the patient.

5. Since everyone had supportive psychotherapy, the placebo group is actually a minimalist psychotherapy group, and the two active arms (paroxetine and imipramine) are drug+psychotherapy groups.

6. The most salient feature of both the acute trial and the extension (as well as the transition from acute to extension) is the high attrition rate (275>190>113>43), suggesting that the endpoint result is not representative of the original sample. The attrition could be interpreted either that the treatment was unsatisfactory or that the study procedures were not patient-friendly enough to retain participants or that it reflects the episodic nature of depression in youth.

7. It appears In Fig. 1 that only the better responders were selected (or self-selected) for the extension, so we should expect some regression to the mean during the extension, which would contribute to the relapse rate.

8. The biggest risk for adverse events appears to be at discontinuation (the taper phase), which could be interpreted either as a dependency risk or as evidence that the medication was doing some good and symptoms erupted when it was discontinued, even by tapering. Regardless, there was self-selection in those who agreed to the taper phase, so we don't know how broadly the taper-phase AE risk applies.

9. You are essentially attempting to conclude the null hypothesis for effectiveness, which requires more power than found in this sample. The original authors were on shaky grounds rejecting the null hypothesis after both primary outcomes failed, but you are on the San Andreas fault in attempting to conclude the null hypothesis when there was benefit on some measures. Probably the most valid conclusion would be that the study failed to show benefit by the pre-specified primary outcome measures, leaving effectiveness in doubt. You seem to have more solid grounds on the potential harms.

Those 9 points should be highlighted as "considerations for interpretation of the findings" to save readers having to repeat the homework I already did.

Figure 1 could be very informative. Unfortunately, it is difficult to follow in its present format, where all 3 conditions use the same symbols (circles) and shade of grey for data points. It looks like it was a color graph changed to B&W. Please use different symbols and lines (solid, dashed, dotted). It appears that psychotherapy plus a pill (whether active or placebo) and the natural course of depression got the mean below a clinical level (over 60% improvement), and selection of responders started all 3 groups at a broad-normal score of 5. After that, it appears that one group made further improvement while 2 groups maintained their improvement. It would be important to state the effect size and p value of the one group compared to the other 2. But this needs to be a LOCF graph; just graphing those who remained at each data point can be misleading.

Showing AEs per 100 weeks exposure is a good strategy. Figures 2 and 3 might be combined for more effect, putting the total and severe columns side by side, illustrating the greater severity of paroxetine AEs. If you add AEs from all phases, is the difference between paroxetine and imipramine significant? Similarly, suicidal and other behavioral AEs could be on the same graph side by side for easy comparison. The figure

legends should clarify whether behavioral/suicidal AEs are also included in the total AEs on the first figure. Again, the legend should indicate whether the difference between paroxetine and imipramine is significant for total suicide and other behavioral AEs.

This could be an interesting and provocative article, and you have clearly put a lot of work into mining the details, but you need to provide its context more clearly for readers.

Reviewer #4:
There are a number of problematic issues with the current analysis and presentation:

In describing relapses in the continuation phase the authors say "Across the study ." and give relapse percentages that use a denominator of all responders -- even those that didn't go into the continuation phase. Since the rate of response was higher in the active treatment groups and the rate of entering extension was higher than in the medication then the placebo and those in the placebo group in the continuation phase dropped out on average more and sooner this results in a misleading comparison of percentages that are not at all comparable.

In general, combining side effects, response, relapse, etc. in the acute titration phase with the continuation phase is problematic. The acute phase includes titration and is likely to include more week-to-week variation in mood and side effects. During the continuation phase there are different reasons for discontinuation from study, including those not involving lack of improvement or side effects.

Raw counts of side effects are also not comparable, since more youth on paroxetine, N=49, than placebo, N=31 entered the continuation and those on placebo dropped out of the continuation phase on average sooner. At a minimum the reader needs rates and confidence intervals, not counts. Also since dropout from the continuation phase wasn't the same for all three groups, presentation of rates per person-month of exposure in each of the three treatment groups are needed in the body of the paper (not just in an appendix).

Confidence intervals must be provided for the side effect rates. Otherwise the reader is unable to make any conclusions about whether side effects are meaningfully greater in an active treatment group than in the placebo group.

Classifying "akathisia", "aggravated depression", "Abnormal dreams", and "depersonalization", as a potential suicidal event (per Appendix 3) is highly questionable.

The analysis of treatment effects in the continuation phase (ANOVA including site, treatment and site x treatment) looking at week by week statistical differences is not sufficiently described and it is not clear whether the graph represents available data at that point (not including missing data or data for dropouts by LOCF imputation) or if the graphs use a LOCF approach. Simply considering the multiple time points as independent as was seemingly done in this analysis would not be correct then or now (but lack of detail precludes certainty as to what exactly was done). The relapse rates need to be calculated solely during the continuation phase on those who entered the continuation phase. The correct analysis would be to use a life-table approach given the very few subjects who were in the full continuation phase and the greatly different rate of discontinuation in the three different treatment groups. Basically on any relapse/death type analysis one cares not primarily
about the rate at some fixed time in the future but on the trajectory of conversions over time-relapsing next month is better than relapsing this month.

It is not usual to include site as a covariate unless there are statistically site effects-you use up too many degrees of freedom and bias the analysis strongly to failing to find a difference.

Reviewer #5 (Statistics and Methodology):
The authors are to be commended for their thorough analysis of the continuation and taper phases of the original study 329. It is clear that there has been a complete analysis of the unpublished data from the

study, and that many aspects of those data have been considered for the first time. This paper could be a valuable addition to the otherwise incomplete reports.

As noted in the abstract and elsewhere, there is no indicated need for formal statistical analysis. The only complex method used, GLM, is applied to the visualization - see figure 1 that is given without p-values, etc.

My general impression is that the paper is unnecessarily wordy and many passages could be edited to more concise form. For example, the entire section on endpoints could be reduced for clarity and brevity. As they now read, the tables are hard to understand. In most cases, they are in CONSORT-diagram form, i.e. the levels are nested. Perhaps inclusion of broad dividers could draw out the structure. In general, the tables seem to be a tedious listing of data. I offer no alternative; perhaps this is the best that can be done.

Similarly, the graphs do not offer much. I am generally a strong advocate of graphic reporting, but these do not seem to do anything.

In summary, this may be an important paper to have on record, but the presentation needs attention. Detail of methods may not be needed, data organization could be improved, and the graphs should be reconsidered.

Lesser editorial comments:
The title would be more informative if "Study 329..." appeared at the end.

CSR should be defined at least once.

In the abstract, taper-phase is mentioned in results, but not defined in methods.

There are a few typos - e.g., page 6 form vs. from, criteria vs. criterion. Careful proofing before publication.
****