## PEER REVIEW HISTORY

BMJ Paediatrics Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Puberty, sex, and behavior in neurodevelopmental disorders versus typical development |
| **AUTHORS** | Penner, Melanie<br>Dupuis, Annie<br>Arnold, Paul<br>Ayub, Muhammad<br>Crosbie, Jennifer<br>Georgiades, Stelios<br>Kelley, Elizabeth<br>Nicolson, Robert<br>Schachar, Russell<br>Anagnostou, Evdokia |

## VERSION 1 – REVIEW

| | |
|---|---|
| **REVIEWER** | Reviewer name: Koyeli Sengupta<br>Institution and Country: United Kingdom of Great Britain and Northern Ireland<br>Competing interests: None |
| **REVIEW RETURNED** | 27-Mar-2022 |

| | |
|---|---|
| **GENERAL COMMENTS** | Dear Authors,<br>It has been a privilege to review the article titled Puberty, Sex and behavior in neurodevelopmental disorders versus typical development submitted to BMJ pediatrics Open.(manuscript ID bmjpo- -2022-001469) and I express my thanks to the editors for providing me with this opportunity.<br>I agree with the authors that parents and caregivers frequently link puberty to worsening behavioral concerns, and this article is a welcome step to examine the evidence.<br>The article is written very clearly, avoids unnecessary jargon, and can potentially help primary-care physicians universally with pointers for anticipatory guidance.<br>Please find my suggestions below:<br><br>Major comment:<br>The participant's self-assessment /caregiver assessment on the Tanner scale is a vital measure in this study. Therefore, it would strengthen the article if the authors provided information about the tool itself. For example, were the graphic representation of the different pubertal stages' actual photographs/pictures or drawings? (Evidence of good-quality graphical material as a factor of better correlation in self-assessment of pubertal stages was suggested by work performed in Hong Kong by Chan et al.,2008).<br>Secondly, please include information on the validity and reliability of the self-report version utilized. Unfortunately, the current citations (#26 and 27 in the Reference list) do not seem to indicate this information clearly.<br><br>Minor comments:<br>Page 8, lines9-10- Authors mention more able participants provided informed Consent. In my understanding, Consent may only be given by individuals who have reached the legal age of Consent (in the |

| | |
|---|---|
| | U.S. this is typically 18 years old). Assent is the agreement of someone unable to give legal Consent to participate in the activity. Kindly confirm if minor participants provided consent/assent.<br>Page 8, lines 15- Kindly mention if there was an upper limit for age for inclusion in the study.<br>Page 8, lines 17-<br>Tanner's stages are also widely referred to as Sexual Maturity Rating (SMR). Kindly include both names for readers familiar with the term SMR.<br>Page 9, line 33-<br>i. It is helpful to know that the authors ensured that the CBCL and Tanners staging self-assessments were contemporaneous. To readers unfamiliar with the OBI-POND database and protocols, the point at which the participants received a diagnosis is unclear.<br>ii. Secondly, would the authors kindly provide some more details on what point participants/caregivers completed the CBCL and Tanners? For example, was it the same time as the initial diagnosis?<br>iii. Participants who completed the Tanners self-report themselves- were they assisted/ supervised by parents or was this done in complete privacy? Would presence or absence of parental guidance/supervision impact the validity of self-report?<br>Page 13, lines 24-26: Discussion section-<br>Authors might also want to consider existing literature suggesting decreased accuracy of self-assessment of SMR in younger children (Peng et al., 2018; Rasmussen et al., 2015). Would that have any implications on this study's finding of lower levels of externalizing behaviors in early pubertal males compared to pre-pubertal males?<br>Page 14: In the Limitations section:<br>i. Lines 31-36: The authors mention that "Both self-report and caregiver-report of Tanner PH stage have been shown to have good reliability in typically<br>developing females,35 though self-report in males is less accurate." However, other authors have commented on the intrasubject variation in the synchronous appearance of pubic hair and breast/genitalia stages, suggesting that it may be misleading to expect stage synchronization in as many as 50% of normal children, across sex. (Cameron N, Bogin B. Human growth and development. 2nd ed.London, UK: Elsevier; 2012.)<br>ii. Would authors like to comment on the possible implications of the discrepancy in sample sizes of the two groups being compared, NDD and typical?<br><br>Page 15, Lines 10-11:<br>Conclusion- It would be advisable to refrain from commenting on the trend of externalizing behaviors in females, as it was not statistically significant.<br><br>Title of the article-<br>Puberty is a period that encompasses overlapping components: hormonal, physical, and cognitive. However, the measure of puberty in this study focused specifically on SMR and did not include other aspects like hormonal levels. Given that, I would urge the authors to consider specifying "Pubertal development / Tanner's Stage/ Pubertal stages in their title rather than the more generic term puberty which encompasses other factors as well. |

| | |
|---|---|
| **REVIEWER** | Reviewer name: Dr. Peter Flom<br>Institution and Country: Peter Flom Consulting New York, United States<br>Competing interests: None |
| **REVIEW RETURNED** | 19-Mar-2022 |

| | |
|---|---|
| **GENERAL** | I confine my remarks to statistical aspects of this paper. While the general approach is OK, I have quite a few issues to resolve before I can recommend publication. I was torn between requiring major or minor revision. |

| | |
|---|---|
| **COMMENTS** | General:<br>Unless there is a very good reason, it is a mistake to group 5 level variables into 3 levels. This increases type 1 and type 2 error.<br><br>Page 8<br>This way of creating a Tanner score seems problematic. It may be OK, but that would need more text to explain why it was done this way (and not simply a reference, unless this method of scoring Tanner is very common and "normal" in the literature).<br><br>Problems: Taking means (line 19) assumes that the 1 to 5 scores are interval or ratio data. This may be reasonable, but it may not. A median would not make this assumption. However, if the mean is appropriate, then categorizing the mean is not appropriate. If you have a sensible mean score, you should use that as the variable. But even if the Tanner stage is reasonable as 1 to 5 score, those should not be further grouped into a 3 level variable.<br><br>So: Best is to leave the score as a rational number, next best 1 to 5. And use the appropriate measure of central tendency.<br><br>p. 9 lines 9-12. If a three way interaction is suggested by theory, then it should be left in the model, regardless of its significance. First, tests of significance are highly dependent on sample size (although that is probably not a problem here). Second, it is also dependent on the reliabilty of the measures. This could be a problem here.<br><br>Third the effect size of the interaction is important. If theory says it should be big, and it is not, then that is important to note. Also see the statement about p values from the American Statistical Assoc. which is at https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjb-O3Hm9L2AhVLkIkEHdqjB3wQFnoECAwQAQ&url=https%3A%2F%2Fwww.amstat.org%2Fasa%2Ffiles%2Fpdfs%2Fp-valuestatement.pdf&usg=AOvVaw2ecBXaMmjRQv05cu8HJjS6 (if that doesn't work, just search for ASA and "p value statement" should find it).<br><br>Tables 2, 3 - First, why show the *predicted* values rather than the actual ones? The latter would be more informative. The regression equations can be shown in formulas, or in the usual kind of table for regression results (showing beta values and so on).<br><br>Second, you can then show the differences for each combination of sex and dx. This would let the reader draw conclusions about the size of the differences and whether they are meaningful (as opposed to significant).<br><br>Finally, while I can't say a figure is *required* it would be nice to have a figure representing figure 2 and one for figure 3 |

| | |
|---|---|
| **REVIEWER** | Reviewer name: Dr. Ann Neumeyer<br>Institution and Country: Massachusetts General Hospital, United States<br>Competing interests: None |
| **REVIEW RETURNED** | 01-Apr-2022 |

| | |
|---|---|
| **GENERAL COMMENTS** | Well written and interesting cross-sectional study of Canadian youth looking for associations between puberty, sex and behavior profile in children with NDD comparing to typical youth.<br><br>My largest concern with the study is the fact that the populations might be quite different with respect to IQ and communication and I don't know if the TD youth were similar with respect to IQ as the other populations, or if there is even a way to check this. Furthermore, do nonverbal children have more behaviors than do verbal adolescents? this might also be addressed. Quick google |

| | |
|---|---|
| | search revealed a manuscript which does also study behavior but not associated with pubertal stage. this study should be noted (guerrera et al Frontiers in Psychiatry 2019).<br>That said, the study needs to state this as a limitation of the study and that this should be addressed in future studies. |

**VERSION 1 – AUTHOR RESPONSE**

BMJ Paediatrics Open – Response to Reviews We would like to thank the reviewers for their thoughtful comments. We have adapted the paper based on their recommendations. Editor in Chief: Comment: Title add " a cross-sectional study" Response: Added. Comment: What this study adds. Delete the 1st sentence as it is Methods Response: Deleted. Comment: Discussion delete the 1st sentence. Journal policy is for authors to avoid describing their study as the first Response: Deleted Reviewer 1: Comment: Unless there is a very good reason, it is a mistake to group 5 level variables into 3 levels. This increases type 1 and type 2 error. Response: Thank you for this comment. Our research question was one of the association between puberty and behavior and not one of how each individual Tanner stage was associated with behavior. As such, there was significant clinical rationale for grouping the Tanner stages into pre-puberty, early puberty, and late puberty to better reflect the research aims. Data within some of the individual Tanner stages were sparse for some of the diagnostic groups, and these groupings allowed us to stay true to our clinical aim while still having sufficient power to carry out the analysis. Comment: Problems: Taking means (line 19) assumes that the 1 to 5 scores are interval or ratio data. This may be reasonable, but it may not. A median would not make this assumption. However, if the mean is appropriate, then categorizing the mean is not appropriate. If you have a sensible mean score, you should use that as the variable. But even if the Tanner stage is reasonable as 1 to 5 score, those should not be further grouped into a 3 level variable. So: Best is to leave the score as a rational number, next best 1 to 5. And use the appropriate measure of central tendency. Response: Thank you for this comment. There is not a clear consensus in the literature as to how to approach these decisions. Please see response above re: decisions to group various Tanner stages based on the clinical rationale for this question. There is no assumption that the effect of Tanner stage on behavior increases/decreases across stages and as such, effects were estimated with each Tanner stage treated as unordered categories, necessitating sufficiently large samples and precluding the estimation of effects for intermediate stages. We agree that the wording could have been clearer and have added the following to page 8: “SOG and PH ratings were combined into one categorical variable representing pubertal status. Where SOG and PH scores differed by 1, the lower score was used. When scores differed by 2, the intermediate score was used.” Comment: p. 9 lines 9-12. If a three way interaction is suggested by theory, then it should be left in the model, regardless of its significance. First, tests of significance are highly dependent on sample size (although that is probably not a problem here). Second, it is also dependent on the reliabilty of the measures. This could be a problem here. Third the effect size of the interaction is important. If theory says it should be big, and it is not, then that is important to note. Also see the statement about p values from the American Statistical Assoc. which is at [link]. Response: While we agree with the reviewer that the p-value should not be the only factor taken into consideration when evaluating an effect, it’s important to note that in both models, the 3-way interaction did not approach statistical significance: Internalizing Behavior $F(6,1019) = 1.28$, $p=0.3$; Externalizing Behavior $F(6,1019) = 0.59$, $p=0.7$. Keeping non-significant interactions in the model would have forced us to report a much large number of pairwise comparisons increasing the risk of type I error unless we adjusted our alpha, or failing to detect important effects that were identified thanks to the gain in power from pooling across categories where effects were not found to be significantly different. For this

reason, we have not included the interaction in the model. Comment: Tables 2, 3 - First, why show the *predicted* values rather than the actual ones? The latter would be more informative. The regression equations can be shown in formulas, or in the usual kind of table for regression results (showing beta values and so on). Second, you can then show the differences for each combination of sex and dx. This would let the reader draw conclusions about the size of the differences and whether they are meaningful (as opposed to significant). Response: Thank you for this helpful comment. We have provided further breakdown of the data in Table 1 to show how age, CBCL Internalizing, and CBCL Externalizing scores differed by diagnosis, sex, and pubertal stage. In the models, we made the decision to present results as predicted values rather than actual values to avoid overinterpretation of differences that are not statistically significant and in the hopes that this might be easier for a clinical audience to interpret. Comment: Finally, while I can't say a figure is *required* it would be nice to have a figure representing [table] 2 and one for [table] 3. Response: We have added these as Figure 1 (Internalizing behavior, corresponds to Table 2) and Figure 2 (Externalizing behavior, corresponds to Table 3). Reviewer 2: Comment: The participant's self-assessment /caregiver assessment on the Tanner scale is a vital measure in this study. Therefore, it would strengthen the article if the authors provided information about the tool itself. For example, were the graphic representation of the different pubertal stages' actual photographs/pictures or drawings? (Evidence of good-quality graphical material as a factor of better correlation in self-assessment of pubertal stages was suggested by work performed in Hong Kong by Chan et al.,2008). Secondly, please include information on the validity and reliability of the self-report version utilized. Unfortunately, the current citations (#26 and 27 in the Reference list) do not seem to indicate this information clearly. Response: We thank the reviewer for this comment. First, we have provided a better reference for the Tanner staging drawings that were used in this study, and specified on page 8 that they were drawings. We have incorporated the reference you shared (p.8): Drawings used for self-assessment in a HongKong sample showed substantial agreement for SOG and PH for females, with males having substantial agreement for PH and moderate agreement for SOG.27 Comment: Page 8, lines9-10- Authors mention more able participants provided informed Consent. In my understanding, Consent may only be given by individuals who have reached the legal age of Consent (in the U.S. this is typically 18 years old). Assent is the agreement of someone unable to give legal Consent to participate in the activity. Response: In Canada, informed consent can be provided irrespective of age, so long as the person is determined to be capable. Comment: Page 8, lines 15- Kindly mention if there was an upper limit for age for inclusion in the study. Response: Thank you – we have clarified on p. 7 that POND recruits participants up until age 21y11m. This study did not have a different upper age limit and for that reason, we have left our description of "age 8 and older" on p.8. Comment: Page 8, lines 17- Tanner's stages are also widely referred to as Sexual Maturity Rating (SMR). Kindly include both names for readers familiar with the term SMR. Response: We have added this after the first mention of the Tanner staging form on page 8: "Participants aged eight years or older (or their caregivers when research staff/caregivers felt that participants were not able) completed a Tanner staging form (also called Sexual Maturity Rating; SMR)…" Comment: Page 9, line 33- It is helpful to know that the authors ensured that the CBCL and Tanners staging self-assessments were contemporaneous. To readers unfamiliar with the OBI-POND database and protocols, the point at which the participants received a diagnosis is unclear. Response: Thank you for this comment. We have clarified that POND enrolls participants at any point after their diagnosis (p. 7, under Setting and Participants). As such, there was no defined period of time since diagnosis that the measures occurred. Comment: Secondly, would the authors kindly provide some more details on what point participants/caregivers completed the CBCL and Tanners? For example, was it the same time as the initial

diagnosis? Response: Thank you for this comment – we have specified on page 7 that CBCL and Tanner staging were completed at the time of enrollment. Comment: Participants who completed the Tanners self-report themselves- were they assisted/ supervised by parents or was this done in complete privacy? Would presence or absence of parental guidance/supervision impact the validity of self-report? Response: Because these participants were thought to be capable of self-report, there were no parental checking procedures included in the data collection. Research staff leave the participant alone in a room to complete this form. Comment: Page 13, lines 24-26: Discussion section- Authors might also want to consider existing literature suggesting decreased accuracy of self-assessment of SMR in younger children (Peng et al., 2018; Rasmussen et al., 2015). Would that have any implications on this study's finding of lower levels of externalizing behaviors in early pubertal males compared to pre-pubertal males? Response: In response to this comment, we compared the age of those in the early puberty stage by respondent and sex to determine if a bias towards reporting a higher Tanner stage could lead to a younger age at early puberty for males. Across both males and females, self-respondents were significantly older, not younger, than those with parent respondents. Mean age (sd) by sex and respondent, early puberty: Self Parent p Males 13.0 (1.7) 12.3 (1.9) .008 Females 12.7 (1.7) 11.8 (1.8) .015 Comment: Page 14: In the Limitations section: Lines 31-36: The authors mention that "Both self-report and caregiver-report of Tanner PH stage have been shown to have good reliability in typically developing females,35 though self-report in males is less accurate." However, other authors have commented on the intrasubject variation in the synchronous appearance of pubic hair and breast/genitalia stages, suggesting that it may be misleading to expect stage synchronization in as many as 50% of normal children, across sex. (Cameron N, Bogin B. Human growth and development. 2nd ed.London, UK: Elsevier; 2012.) Response: Thank you for this comment. We agree that intra-subject variation can exist. In order to account for this, we adopted a balanced approach where we allowed for two stages of variation within an individual but excluded those with more than this due to concerns about reliability of reporting. Comment: Would authors like to comment on the possible implications of the discrepancy in sample sizes of the two groups being compared, NDD and typical? Response: We agree that the sample size from the NDD diagnoses are larger than the TD groups, which is due to POND's focus on NDDs. We have analyzed and reported the NDD groups by diagnosis, which helps to bring about a bit more balance. Comment: Page 15, Lines 10-11: Conclusion- It would be advisable to refrain from commenting on the trend of externalizing behaviors in females, as it was not statistically significant. Response: Thank you for this comment – we have removed the statement about externalizing behaviors from the Conclusion. Comment: Title of the article- Puberty is a period that encompasses overlapping components: hormonal, physical, and cognitive. However, the measure of puberty in this study focused specifically on SMR and did not include other aspects like hormonal levels. Given that, I would urge the authors to consider specifying "Pubertal development / Tanner's Stage/ Pubertal stages in their title rather than the more generic term puberty which encompasses other factors as well. Response: This is an excellent point – we have changed the title to "Pubertal stage, sex, and behavior in neurodevelopmental disorders versus typical development" Reviewer 3 Comment: My largest concern with the study is the fact that the populations might be quite different with respect to IQ and communication and I don't know if the TD youth were similar with respect to IQ as the other populations, or if there is even a way to check this. Furthermore, do nonverbal children have more behaviors than do verbal adolescents? this might also be addressed. Quick google search revealed a manuscript which does also study behavior but not associated with pubertal stage. this study should be noted (guerrera et al Frontiers in Psychiatry 2019). That said, the study needs to state this as a limitation of the study and that this should be addressed in future studies. Response: Thank

you for this comment. We agree that IQ and communication skills are factors that could influence behavioral trajectories through puberty but were beyond the scope of this study. We have added this to the Limitations section: “Longitudinal studies measuring pubertal stage and behavior are needed to optimally disentangle the effects of age and puberty, and should include factors such as

**VERSION 2 – REVIEW**

| | |
|---|---|
| **REVIEWER** | Reviewer name: Koyeli Sengupta<br>Institution and Country: United Kingdom of Great Britain and Northern Ireland<br>Competing interests: None |
| **REVIEW RETURNED** | 28-Apr-2022 |

| | |
|---|---|
| **GENERAL COMMENTS** | In my opinion, the authors have addressed and responded adequately to all my queries and comments.<br>Thank you. |

| | |
|---|---|
| **REVIEWER** | Reviewer name: Dr. Peter Flom<br>Institution and Country: Peter Flom Consulting New York, United States<br>Competing interests: None |
| **REVIEW RETURNED** | 18-Apr-2022 |

| | |
|---|---|
| **GENERAL COMMENTS** | The authors have addressed most of my concerns. My remaining issue is this one:<br>I wrote<br><br>Comment: p. 9 lines 9-12. If a three way interaction is suggested by theory, then it should be left in the model, regardless of its significance. First, tests of significance are highly dependent on sample size (although that is probably not a problem here).<br><br>Second, it is also dependent on the reliabilty of the measures. This could be a problem here. Third the effect size of the interaction is important. If theory says it should be big, and it is not, then that is important to note. Also see the statement about p values<br>from the American Statistical Assoc. which is at [link].<br><br>The authors replied:<br><br>Response: While we agree with the reviewer that the p-value should not be the only factor taken into consideration when evaluating an effect, it’s important to note that in both models, the 3-way interaction did not approach statistical significance: Internalizing Behavior F(6,1019) = 1.28, p=0.3;Externalizing Behavior F(6,1019) = 0.59, p=0.7. Keeping non-significant interactions in t he model would have forced us to report a much large number of pairwise comparisons increasing the risk of type I error<br>unless we adjusted our alpha, or failing to detect important effects that were identified thanks to the gain in power from pooling across categories where effects were not found to be significantly different.<br><br>For this reason, we have not included the interaction in the model.<br><br>My response. You do not need to test all pairwise comparisons, if theory suggests that only some are important . You can simply report the predicted values and SEs. The authors' response does not address the idea of showing that an effect, which was thought to be large, is small.<br><br>Another option is to first show results with the interaction and then |

| | |
|---|---|
| | say something like "While Smith and Jones suggested that this three way interaction would be large, we found only a small and non-significant effect (see Table XXX)" We therefore first report these results, and then proceed with a simpler model, removing these interactions"<br><br>I think that's a good compromise.<br><br>Peter Flom |

## VERSION 2 – AUTHOR RESPONSE

We would like to thank the reviewers for their continued input to strengthening the paper. In response to Reviewer 1, we have now included the results of the full model in a Supplementary Table, where the CBCL score differences are reported by pubertal stage, sex, and diagnosis. We have made minor cuts to other wording in order to still be within the journal's word limit.