PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Impact of gestational age on child intelligence, attention, and executive function at age five: a cohort study
AUTHORS	Sejer, Emilie Pi; Bruun, Frederik; Slavensky, Julie Anna; Mortensen, Erik; Schiøler Kesmodel, Ulrik

VERSION 1 – REVIEW

REVIEWER	Abdool Yasseen University of Toronto, Dalla Lana School of Public Health, Epidemiology Division, Toronto, Canada
REVIEW RETURNED	30-Jan-2019

GENERAL COMMENTS	Comments: Abstract: - "follow up study" is not a study design, please specify if this is a longitudinal retrospective cohort or a prospective cohort. It seems like the former. Also since samples of the birth registry are taken, this should also be included in the study design description (e.g., population based retrospective sample …). - The study setting should be hospital based. - Participants should be both the mother-child dyad unit as the authors are investigating maternal and infant characteristics. Introduction: - Line 22, what proportion of term births receive care at centers specialized for children with disabilities? It would be good to have a comparator. Methods: - The study cohort is the result of sampling from a larger population. Please state what type of sampling was used, random or stratified sampling. If stratified, please specify what variables were stratified for. - Line 43, it is difficult to conceive that there is no considerable difference between participants and non-participants, please elaborate on the variables used to compare these groups, from reference #13 (e.g., age, sex, immigrant status, etc…). - I agree with the exclusion of high risk pregnancies and those that are likelier to not be neurophysiologically tested. However the authors should include a sentence to justify these exclusions. (e.g., multi-fetal pregnancies et al. were excluded since they represent a fundamentally different group of individuals that may not be representative of the norm…) - Please provide the reader with some information on how gestational age was determined (e.g., date of last menses, Ultrasound estimate, or a combination?)

	- GA is a continuous measure that the authors have categorized it into well-known preterm groups. However, they failed to categorize early term births (37 and 38 weeks), which may allow for a better contrast between the preterm groups and full term (i.e., 39 and 40 weeks). I would suggest a sensitivity analysis where the early term group is removed, and a direct comparison between preterm and full term is made. This may improve the contrast and produce results more in line with the author's conclusions. - While important for etiologic studies, DAGs are only useful in conceptualizing the exposure-outcome relationship in the presence of a limited number of possible co-variates. When there are greater than 5 co-variates, the whole process gets convoluted. Please remove reference to using DAGs, as I believe it is enough to say that we adjusted for clinically selected and available co-variates and investigated the possibility of confounding. Alternatively, if the authors decide to keep the DAG, they would have to include more details on the specific relationships listed in supplementary figure 1. I believe this would detract from the focal point of the paper and I would advise against doing so. Statistical analysis: - Please refer to the regression models used as "multivariable linear regression", and not "multiple linear regression analyses". - As stated previously, I would like to see a sensitivity analysis of the relationship between late preterm and full term (i.e., excluding early term births). I believe this may produce better results and further support the conclusions. - Please comment on the possibility for co-linearlity between co-variates included in the model. NB: Checking the variance inflation factor might help identify redundant variables, which may improve model fit. - The sample size of the study is not huge, please provide a post hoc power analysis to educate the reader on what margin of error is expected, given the achieved sample size. - With the line: "Inserting a potential intermediate factor…" I assume that the authors introduced a cross-product interaction term into the model. However, this by itself does not mean that the variable acts as a mediator. Please specify if counterfactual modeling was used, or alternatively the Baron-Kenny approach. Mediation is a difficult analysis to justify, and it might be better to drop this analysis rather than attempt to explain. It would be sufficient to say that variable interactions, suggestive of mediation were observed.

REVIEWER	Julian Mutz Social Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom.
REVIEW RETURNED	01-Feb-2019

GENERAL COMMENTS	Review of bmjopen-2019-028982 The main objective of this study was to examine the associations between gestational age (GA)/preterm birth and cognitive outcomes (IQ, attention, executive function) in children at the age of 5 using data from the Lifestyle During Pregnancy Study. The study takes into account a range of potential confounders, for instance maternal IQ, maternal age at birth and alcohol consumption during pregnancy.

	While this is an interesting study, there are several important limitations that limit my confidence in these findings. I have made several suggestions which I hope the authors will find useful in improving their work. Major revisions (1) In their analysis of GA as a categorical explanatory variable, only 8 mother-child pairs met the criteria for “very to moderately” preterm birth. This is a major limitation and should be more thoroughly discussed, especially in the abstract when the authors state that “GA has a crucial role in determining cognitive…”. Substantially more discussion of the implications of the low number of subjects in the “very to moderately” preterm group is needed. No power analyses have been reported, although the authors state several times that their study had limited power (to detect what effect size?). (2) The uncertainty of the effect sizes is illustrated by very wide confidence intervals. For example, the mean difference in performance IQ between the “very or moderately” preterm group and the “at term” group ranges from -21.9 to -1.5. How should readers interpret these findings? Can we be confident that this is not a chance finding? (3) The authors state in their introduction (P6 l.26) that it is important to examine the associations between preterm birth/low GA and cognitive outcomes to advise women at risk of preterm delivery, and “to give informed predictions about the future”. However, I am not convinced that the current study with its limitations provides strong evidence that can be used to inform policy. Moreover, no findings in the present report address the question of resuscitation that is mentioned as a rationale for studying low GA (P6 l.33). (4) The authors state that part of the effect of low GA on low IQ is mediated through low birthweight (e.g. Abstract l.53). However, appropriate models for mediation analysis have not been included in this work. (5) How should readers interpret the finding that simultaneously removing maternal IQ and parental education from the regression analyses results in non-statistically significant associations between GA and cognitive outcomes? Please include the results of these analyses (and for any other findings for which the authors report “data not shown”); otherwise, there is no way of verifying the validity of these findings. If these results cannot be included in the main body of the text, I suggest that these be included in the supplementary material. (6) How should the finding that adding birthweight to the model removes the statistically significant association between GA and IQ? Have the authors examined the association between GA and birthweight? Considering that the authors highlight this analysis as one of their main findings, I suggest that appropriate models for mediation analysis should be used. (7) P13 l.31: the authors state that they find “a statistically significant effect of very or moderately preterm birth on IQ …, even when adjusting…”. I do not agree with the language here. In fact, these associations only seem to be statistically significant in the adjusted models. How should readers interpret these findings? (8) Further justification is needed for why the authors decided to include “moderate to very” preterm birth as a group with 8 subjects but not include “very or extremely” preterm birth. (9) While I understand the appeal to make these findings clinically easy to interpret by using subgroups, I am wondering whether it is perhaps more sensible to focus on the continuous GA analyses?

	Minor revisions Abstract P2 l.12: consider replacing “34-<37” with “34 to <37” to improve legibility. This is worth updating throughout the main body of the text and in all tables. P2 l.15: I am wondering whether it would be more accurate (and transparent) to refer to the “moderate or very preterm” group with a range in weeks rather than a cut-off value, i.e. “xx-33” instead of “<34”. P2 l.20: for the study design “prospective cohort study” seems more specific than “follow-up study”. P2 l.28: I suggest that the authors aim for a more consistent way of reporting the variables that they included as potential confounders. Throughout their manuscript they use different labels, for example “family and background factors” or “socioeconomic confounders”. P2 l.46: please also include the number of subjects for the at term group. P2 l.56: I suggest refraining from using language such as “proved…” in this context. P2 l.56: it is not obvious what the authors mean when they refer to “weak confounders”. P3 l.10: consider replacing “significantly lower” with “substantially lower” if these differences are practically relevant differences. In the case of statistically significant differences, please state this as such. Article summary P4 l.18: consider replacing “exhaustive” with “extensive”. P4 l.27: the statement that there are “shortcomings in the data” is too vague. Please try to be more specific. P4 l.32: please state the number of subjects born preterm as well as the corresponding proportion (with both numerator and denominator). A brief comment on the implications of this limitation would be appropriate here too. Material and methods P7 l.38: please clarify for whom the neuropsychological tests were performed (i.e. mothers, children or both). This is not obvious from the current wording. P8 l.7: further details on the continuous exposure variable (i.e. GA in months) would be useful to distinguish that these are two separate analyses. P8 l.33: could the authors clarify whether this approach has previously been used in similar investigations. P9 l.42: please add the relevant website or citation for the software. P9 l.47: Danish Medical Birth Registry should not be abbreviated as this is no commonly known acronym. P10 1st paragraph: could the authors please specify from which exposure groups these subjects were excluded. P10 l.27: further details are needed regarding how categorical variables were added to the models. P10 l.30: please indicate maternal age at birth. P10 l.34: the description of the educational level variable (total duration in years) should be presented here instead of further below. P11 l.11: please add citation for robust standard errors. Results P12 l.23: “this trend diminished” – I do not find this language helpful here. Perhaps state “we found no evidence of… after controlling for…”.

	P12 l.31: see previous comment. Perhaps state that you did not find evidence of statistically significant associations. P12 l.33: please clarify whether “significant” refers to statistical and/or practical significance here. P12 l.46: I suggest that the authors provide further comment on the results of their analysis of GA as a continuous variable. P13 l.4: please also see earlier comments. Could the authors please clarify whether “significant” refers to statistical and/or practical significance. Discussion P14 l.7: the authors highlight the large sample size of their cohort as one of the main strengths of this work. While I agree that the cohort itself is large, a major limitation of their study is the small number of subjects that were in the group of “moderately or very” preterm birth, which is the main explanatory variable in their model. P14 l.19: please add a copy of the study protocol to the supplementary material. P14 l.34: substantially more information on power analyses (which have not been presented at all) are necessary here. P15 l.6: please see previous comments regarding “significance”. P15 l.41: I suggest the authors refrain from using language such as “borderline statistically significant”. Conclusion P16 l.40: considering that very limited details have been provided regarding the mediation analysis, further work is needed for this to be included as one of the main conclusions of the present work. Please see relevant comments above. P16 l.51: I do not agree that based on the current findings we can conclude that low GA has a crucial role in determining cognitive abilities. A more careful evaluation of their findings in light of uncertainty may be more appropriate here. Table 1 Consider replacing “gender” with “sex”. Table 2 In the table legend the authors state the number of participants for whom data are available. These numbers ought to be reported for each term group.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

Thank you for your assessment of the association between preterm birth and cognitive function. This is a wonderful area of research that needs population based data to address research questions. Below are a few of my main concerns with the methodology.

As with any population based study, careful thought must be given to who is included and excluded, and the choice of how to analyze the data. The authors need to add more details about the sampling and selection process of those included in the study. A reference was made to a previous study, however some fundamental details should be included in the current paper.

As it is difficult to tease apart the complex relationships and factors that influence cognitive development, it is indeed a strength of the study to have so many co-variates to explore.

However, with parametric methods such as regression analysis, it is often of equal importance to consider how these relationships affect the analysis methodology.

The use of a mediation analysis needs to be either dropped or elaborated more. I would suggest dropping this analysis, due to the complexity involved.

For these reasons as well as those listed in my comments, I suggest that the manuscript in its current version would require modifications to the analysis and interpretation, before it is ready for publication. I believe it is an important topic, and given that the authors can address my suggestions/edits, I believe it is publishable.
I hope these comments are helpful,
Reviewer
Comments:
Abstract:
- "follow up study" is not a study design, please specify if this is a longitudinal retrospective cohort or a prospective cohort. It seems like the former. Also since samples of the birth registry are taken, this should also be included in the study design description (e.g., population based retrospective sample…).
We thank the reviewer for the comment. In fact, given the prospective information on GA from the Danish Medical Birth Register, we find it most reasonable to consider the study a prospective cohort study – which is also suggested by reviewer 2. We have changed the abstract to:
"Design: Population-based prospective cohort study."
- The study setting should be hospital based.

We thank the reviewer for the comment. However, the study setting is population-based and not hospital-based. The participants in this cohort were identified through their general practitioner throughout Denmark during the study period, and all contacts related to the original data collection in the Danish National Birth Cohort and later in the Lifestyle During Pregnancy Study (the follow-up of the children at age 5) were done without involvement of hospitals.
- Participants should be both the mother-child dyad unit as the authors are investigating maternal and infant characteristics.

Revised according to the remarks of the reviewer. Thank you. The text now states:
"A cohort of 1776 children and their mothers."
Introduction:
- Line 22, what proportion of term births receive care at centers specialized for children with disabilities? It would be good to have a comparator.

We thank the reviewer for the comment, and we have now addressed this matter in the introduction:
"A study showed that at age five 10% of children born preterm still received care in centres specialised for children with disabilities compared to 2% of children born at term (odds ratio 7.9 [95% CI; 3.5 to 18.0]).5"
Methods:
- The study cohort is the result of sampling from a larger population. Please state what type of sampling was used, random or stratified sampling. If stratified, please specify what variables were stratified for.

We thank the reviewer for the comment and have revised the manuscript accordingly:
"Participants were sampled in strata defined by the prenatal maternal average alcohol intake with oversampling of women reporting a relatively high alcohol intake or binge drinking episodes during pregnancy.12 13"
- Line 43, it is difficult to conceive that there is no considerable difference between participants and non-participants, please elaborate on the variables used to compare these groups, from reference #13 (e.g., age, sex, immigrant status, etc…).

We thank the reviewer for the comment and have revised the manuscript accordingly:

“There were no considerable differences between the participants and non-participants with regard to maternal age, body mass index, parity, marital status, prenatal smoking and alcohol consumption, child sex, birthweight, and gestational age at birth.13”
- I agree with the exclusion of high risk pregnancies and those that are likelier to not be neurophysiologically tested. However the authors should include a sentence to justify these exclusions. (e.g., multi-fetal pregnancies et al. were excluded since they represent a fundamentally different group of individuals that may not be representative of the norm…)

We thank the reviewer for the comment and have revised the manuscript accordingly:
“Exclusion criteria were multiple pregnancies and congenital diseases with a large risk of mental retardation (the diagnostic term used at the time of data collection), as they represent a fundamentally different group of individuals that may not be representative of the norm”
- Please provide the reader with some information on how gestational age was determined (e.g., date of last menses, Ultrasound estimate, or a combination?)

We thank the reviewer for the comment and have revised the manuscript accordingly:
“Information on GA was obtained from the Danish Medical Birth Register and determined by ultrasound, while date of last menses was only used to determine GA in very few cases where an ultrasound estimate was not available.”
- GA is a continuous measure that the authors have categorized it into well-known preterm groups. However, they failed to categorize early term births (37 and 38 weeks), which may allow for a better contrast between the preterm groups and full term (i.e., 39 and 40 weeks). I would suggest a sensitivity analysis where the early term group is removed, and a direct comparison between preterm and full term is made. This may improve the contrast and produce results more in line with the author’s conclusions.

We thank the reviewer for the comment. We have conducted the requested post hoc analyses with a direct comparison between the very or moderately preterm group and the full term group (GA ≥ 39 weeks), and the late preterm group and the full term group (GA ≥ 39 weeks), respectively. These analyses produced essentially the same results as the already conducted analyses where children born at term with GA 37-38 weeks were included in the term group. Therefore, the risk of cognitive impairment did not differ within the term group, and removing the early term births (GA 37-38 weeks) did not alter the results.
We have now addressed this in the results section of the manuscript:
“In a post hoc analysis, we excluded the early term births (GA 37-38) and made a direct comparison between the very or moderately preterm group and the term group with GA ≥ 39 weeks (n=1443), and the late preterm group and the term group (GA ≥ 39 weeks), respectively (see supplementary table 2). In these analyses, the results did not change notably for any of the outcomes.”
- While important for etiologic studies, DAGs are only useful in conceptualizing the exposure-outcome relationship in the presence of a limited number of possible covariates. When there are greater than 5 co-variates, the whole process gets convoluted. Please remove reference to using DAGs, as I believe it is enough to say that we adjusted for clinically selected and available co-variates and investigated the possibility of confounding. Alternatively, if the authors decide to keep the DAG, they would have to include more details on the specific relationships listed in supplementary figure 1. I believe this would detract from the focal point of the paper and I would advise against doing so.

We thank the reviewer for the comment. When planning the study we decided to use DAGs to identify relevant covariates. Therefore, we believe that it should be included in our methods section, as it describes our actual approach, which was defined a priori. However, because of the concern that the specific DAG could detract from the focal point of the paper, we agree to remove the DAG (supplementary figure 1) from our paper.
Statistical analysis:

- Please refer to the regression models used as “multivariable linear regression”, and not “multiple linear regression analyses”.

Revised according to the remarks of the reviewer. Thank you.
- As stated previously, I would like to see a sensitivity analysis of the relationship between late preterm and full term (i.e., excluding early term births). I believe this may produce better results and further support the conclusions.

We thank the reviewer for the comment. Please see our response to the request above.
- Please comment on the possibility for co-linearlity between co-variates included in the model. NB: Checking the variance inflation factor might help identify redundant variables, which may improve model fit.

We thank the reviewer for the comment and have now addressed this in the methods (‘statistical analyses’) section:
“We investigated the possibility for collinearity between covariates and found no evidence of this, as the variance inflation factor never exceeded a value of 2 for any of the covariates in the regression models.”
- The sample size of the study is not huge, please provide a post hoc power analysis to educate the reader on what margin of error is expected, given the achieved sample size.
We thank the reviewer for the comment. We have provided a post hoc power analysis in the discussion section:
“A post hoc power analysis showed that analyses comparing very or moderately preterm birth (n=8) with birth at term (n=1728) had a power of 0.48, 0.28, and 0.59 for FIQ, VIQ, and PIQ outcomes, respectively.”
- With the line: “Inserting a potential intermediate factor…” I assume that the authors introduced a cross-product interaction term into the model. However, this by itself does not mean that the variable acts as a mediator. Please specify if counterfactual modeling was used, or alternatively the Baron-Kenny approach. Mediation is a difficult analysis to justify, and it might be better to drop this analysis rather than attempt to explain. It would be sufficient to say that variable interactions, suggestive of mediation were observed.
We thank the reviewer for the comment. None of the above mentioned approaches were used. As suggested by the reviewer, we have chosen to state that results for models with birthweight could suggest mediation. In the discussion, the text now states:
“The inclusion of birthweight in the regression analyses for IQ outcomes attenuated the associations for the very or moderately preterm group, and for the late preterm group, the associations completely vanished. This could be suggestive of mediation and underlines the importance of looking at GA relatively to birthweight when investigating effects of preterm birth, though our results for the very to moderately preterm children indicate that there may be cognitive effects of GA which are independent of birthweight, perhaps reflecting effects of very low GA on brain development.”

Reviewer 2
The main objective of this study was to examine the associations between gestational age (GA)/preterm birth and cognitive outcomes (IQ, attention, executive function) in children at the age of 5 using data from the Lifestyle During Pregnancy Study. The study takes into account a range of potential confounders, for instance maternal IQ, maternal age at birth and alcohol consumption during pregnancy. While this is an interesting study, there are several important limitations that limit my confidence in these findings. I have made several suggestions which I hope the authors will find useful in improving their work.
Major revisions
(1) In their analysis of GA as a categorical explanatory variable, only 8 mother-child pairs met the criteria for “very to moderately” preterm birth. This is a major limitation and should be more thoroughly

discussed, especially in the abstract when the authors state that “GA has a crucial role in determining cognitive…”. Substantially more discussion of the implications of the low number of subjects in the “very to moderately” preterm group is needed. No power analyses have been reported, although the authors state several times that their study had limited power (to detect what effect size?).

We thank the reviewer for the comment. We have toned down our conclusions and mentioned this limitation in the abstract as well. It will seem quite obvious to most readers that with only eight observations in one subgroup, the power is limited. A post hoc power analysis has been added to the discussion section.

(2) The uncertainty of the effect sizes is illustrated by very wide confidence intervals. For example, the mean difference in performance IQ between the “very or moderately” preterm group and the “at term” group ranges from -21.9 to -1.5. How should readers interpret these findings? Can we be confident that this is not a chance finding?

We thank the reviewer for the comment. Chance may always be a potential explanation for any finding. This is exactly why 95% CIs are provided (rather than just a p-value), as it tells the reader that we are 95% certain/confident that the true value in the background population lies somewhere between the two extremes. So in the example mentioned by the reviewer we are 95% certain/confident that the true difference in IQ between the two groups is in the CI interval which does not include 0, and thus is statistically significant at the 5% level. We assume this understanding and interpretation of a CI to be common knowledge among readers of BMJ Open. If the editor prefers, p-values can be added to the table.

(3) The authors state in their introduction (P6 l.26) that it is important to examine the associations between preterm birth/low GA and cognitive outcomes to advise women at risk of preterm delivery, and “to give informed predictions about the future”. However, I am not convinced that the current study with its limitations provides strong evidence that can be used to inform policy. Moreover, no findings in the present report address the question of resuscitation that is mentioned as a rationale for studying low GA (P6 l.33).

We thank the reviewer for the comment and agree that decisions regarding resuscitation and inform policies on this important matter should not be based on a single study alone, but rather on all the evidence that exists in the literature on the subject. We believe that this study contributes with important knowledge that improves the overall understanding of this matter. We have addressed this further in the discussion:

“Despite the limitations, especially the low number of preterm births, we believe that this study contributes with important knowledge that together with existing evidence in the literature may improve the clinicians’ ability to advise women at risk of preterm delivery and give informed predictions about the future.”

(4) The authors state that part of the effect of low GA on low IQ is mediated through low birthweight (e.g. Abstract l.53). However, appropriate models for mediation analysis have not been included in this work.

We thank the reviewer for the comment. The sentence, which the reviewer is referring to, has now been removed. No formal mediation analyses were carried out, however, we introduced birthweight in our regression models to see how it would affect the results.

As suggested by reviewer 1, we have now stated that in this study, results suggestive of mediation were observed. This has been revised throughout the manuscript. In the discussion, the text now states:

“The inclusion of birthweight in the regression analyses for IQ outcomes attenuated the associations for the very or moderately preterm group, and the results were no longer statistically significant. For the late preterm group, the associations completely vanished. This could be suggestive of mediation and underlines the importance of looking at GA relatively to birthweight when investigating effects of preterm birth, though our results for the very to moderately preterm children indicate that there may be cognitive effects of GA which are independent of birthweight, perhaps reflecting effects of very low GA on brain development.”

(5) How should readers interpret the finding that simultaneously removing maternal IQ and parental education from the regression analyses results in non-statistically significant associations between GA and cognitive outcomes? Please include the results of these analyses (and for any other findings for which the authors report “data not shown”); otherwise, there is no way of verifying the validity of these findings. If these results cannot be included in the main body of the text, I suggest that these be included in the supplementary material.
We thank the reviewer for the comment. As specified in our methods section, to evaluate the importance of including maternal intelligence and parental education in the assessment of an association between GA and cognitive outcomes, analyses were also conducted without these two variables. Although removing these variables from the regressions produced wider CIs indicating that they explain substantial parts of the variance in the IQ outcomes, the estimates of the association did not change notably, suggesting they are only weak confounders of the association of GA with offspring IQ and have no significant association with GA. The text now states:
“Although maternal IQ and parental education accounted for much of the variance in child IQ in this dataset,7 these two factors should only be considered weak confounders with no significant association with GA, as removing these variables from our analyses did not alter the associations notably. However, removal of the variables produced wider CIs confirming that they explain substantial parts of the variance.”
We have included the results of the analyses (previously not shown) as a supplementary table (1).
(6) How should the finding that adding birthweight to the model removes the statistically significant association between GA and IQ? Have the authors examined the association between GA and birthweight? Considering that the authors highlight this analysis as one of their main findings, I suggest that appropriate models for mediation analysis should be used.
We thank the reviewer for the comment. We have addressed this issue above (major revision #4). The fact that the GA effects on IQ outcomes vanish when birthweight is introduced in the regression analyses, suggests that the GA effect may in part be mediated though birthweight. As no formal mediation analyses have been carried out, this conclusion has been toned down in the manuscript.
(7) P13 l.31: the authors state that they find “a statistically significant effect of very or moderately preterm birth on IQ …, even when adjusting…”. I do not agree with the language here. In fact, these associations only seem to be statistically significant in the adjusted models. How should readers interpret these findings?
Revised according to the remarks of the reviewer. Thank you. The text now states:
“We found a statistically significant effect of very or moderately preterm birth on IQ and teacher-assessed executive function when adjusting for potential confounders.”
The associations probably become statistically significant in the adjusted analyses, as we adjust for multiple variables that have proven to be predictors of our outcomes. Therefore, the ME error is reduced and the estimates are more likely to become statistically significant.
(8) Further justification is needed for why the authors decided to include “moderate to very” preterm birth as a group with 8 subjects but not include “very or extremely” preterm birth.
We thank the reviewer for the comment. We decided not to further divide the 8 subjects with GA <34 weeks into subgroups of very or extreme prematurity, as we thought it would not make sensible results. This matter is addressed in the ‘strengths and limitations’ section in the discussion.
(9) While I understand the appeal to make these findings clinically easy to interpret by using subgroups, I am wondering whether it is perhaps more sensible to focus on the continuous GA analyses?
We thank the reviewer for the comment. The analyses with GA as a continuous variable produced essentially the same results as the analyses with GA as a categorical variable, and we have now elaborated the results of the continuous GA analyses in the results section. However, we chose to focus on the results of the categorical GA analyses, as they clinically are easier to interpret.
Minor revisions
Abstract

P2 l.12: consider replacing “34-<37” with “34 to <37” to improve legibility. This is worth updating throughout the main body of the text and in all tables.
We thank the reviewer for the comment and have revised the manuscript accordingly.
P2 l.15: I am wondering whether it would be more accurate (and transparent) to refer to the “moderate or very preterm” group with a range in weeks rather than a cut-off value, i.e. “xx-33” instead of “<34”.
We thank the reviewer for the comment. However, we think it would be less accurate as it can be unclear whether 33 mean 33+0, 33+1, 33+6, or something else. Therefore, we have not changed the classification.
P2 l.20: for the study design “prospective cohort study” seems more specific than “follow-up study”.
We thank the reviewer for the comment. We have changed it to: Population-based prospective cohort study.
P2 l.28: I suggest that the authors aim for a more consistent way of reporting the variables that they included as potential confounders. Throughout their manuscript they use different labels, for example “family and background factors” or “socioeconomic confounders”.
Revised according to the remarks of the reviewer. Thank you. Generally, we have changed it to “family and background factors”.
P2 l.46: please also include the number of subjects for the at term group.
Revised according to the remarks of the reviewer. Thank you.
P2 l.56: I suggest refraining from using language such as “proved…” in this context.
Revised according to the remarks of the reviewer. “Proved” has been changed to “were”.
P2 l.56: it is not obvious what the authors mean when they refer to “weak confounders”.
We thank the reviewer for the comment. We have tried to elaborate in the ‘main findings’ section in the discussion:
“Although maternal IQ and parental education accounted for much of the variance in child IQ in this dataset,7 these two factors should only be considered weak confounders with no significant association with GA, as removing these variables from our analyses did not alter the associations notably.”
P3 l.10: consider replacing “significantly lower” with “substantially lower” if these differences are practically relevant differences. In the case of statistically significant differences, please state this as such.
We thank the reviewer for the comment. We have changed the wording to “substantially lower”. These results are also statistically significant which is shown in the ‘results’ section of the abstract.
Article summary
P4 l.18: consider replacing “exhaustive” with “extensive”.
Revised according to the remarks of the reviewer. Thank you.
P4 l.27: the statement that there are “shortcomings in the data” is too vague. Please try to be more specific.
Revised according to the remarks of the reviewer. Thank you. The text now states:
“Robust standard errors were used to account for the sample design, possible deviations from normality and variance homogeneity.”
P4 l.32: please state the number of subjects born preterm as well as the corresponding proportion (with both numerator and denominator). A brief comment on the implications of this limitation would be appropriate here too.
Revised according to the remarks of the reviewer. Thank you. The text now states:
“The proportion of children born preterm in this study population was small (48 out of 1776), which limited our power to detect any true differences as significant.”
Material and methods
P7 l.38: please clarify for whom the neuropsychological tests were performed (i.e. mothers, children or both). This is not obvious from the current wording.
Revised according to the remarks of the reviewer. Thank you. The text now states:
“Out of the sampled mother and child pairs, 1776 children had neuropsychological tests performed at age five and had information on GA available, and thus were included in our analyses.”

P8 l.7: further details on the continuous exposure variable (i.e. GA in months) would be useful to distinguish that these are two separate analyses.
Revised according to the remarks of the reviewer. Thank you. The text now states:
“We used GA as 1) a continuous variable (days) and 2) a categorical variable, comparing late preterm birth (34 to <37 completed weeks of gestation) and very to moderately preterm birth (GA<34 weeks) with birth at term (GA ≥37 weeks), respectively.”
P8 l.33: could the authors clarify whether this approach has previously been used in similar investigations.
We thank the reviewer for the comment. This approach has been used in all previous publications based on this sample. Prorating for example is described in most Wechsler manuals.
P9 l.42: please add the relevant website or citation for the software.
Revised according to the remarks of the reviewer. Thank you.
P9 l.47: Danish Medical Birth Registry should not be abbreviated as this is no commonly known acronym.
Revised according to the remarks of the reviewer. Thank you.
P10 1st paragraph: could the authors please specify from which exposure groups these subjects were excluded.
We thank the reviewer for the comment and have revised the manuscript accordingly:
“This resulted in removal of three birthweight observations (one from the term group and two from the late preterm group) that exceeded our threshold when evaluated according to Danish standards.21 Moreover, we removed one unrealistic body mass index of 13.9 kg/m2 and one observation of average alcohol intake of 36 drinks/week during pregnancy (from the term group).”
As stated, we excluded two birthweight observations from the late preterm group (the rest of the excluded observations were from the term group). However, birthweight was not part of our main multivariable linear regressions, and therefore it did not affect the number of preterm subjects included in these analyses.
P10 l.27: further details are needed regarding how categorical variables were added to the models.
We thank the reviewer for the comment. We have revised the manuscript accordingly:
“We created dummy variables from the categorical variables before inserting them in the regression models.”
P10 l.30: please indicate maternal age at birth.
Revised according to the remarks of the reviewer. Thank you.
P10 l.34: the description of the educational level variable (total duration in years) should be presented here instead of further below.
Revised according to the remarks of the reviewer. Thank you.
P11 l.11: please add citation for robust standard errors.
Revised according to the remarks of the reviewer. Thank you.
Results
P12 l.23: “this trend diminished” – I do not find this language helpful here. Perhaps state “we found no evidence of… after controlling for…”.
Revised according to the remarks of the reviewer. Thank you. The text now states:
“Among the late preterm children, a tendency towards lower IQs was evident in the unadjusted analyses, but we found no statistically significant differences after adjusting for potential confounders.”
P12 l.31: see previous comment. Perhaps state that you did not find evidence of statistically significant associations.
Revised according to the remarks of the reviewer. Thank you. The text now states:
“For the attention measures, the mean differences were small, and we did not find evidence of statistically significant associations.”
P12 l.33: please clarify whether “significant” refers to statistical and/or practical significance here.
Revised according to the remarks of the reviewer. Thank you.
P12 l.46: I suggest that the authors provide further comment on the results of their analysis of GA as a continuous variable.

Revised according to the remarks of the reviewer. Thank you. We have added the following text to the results section:
“Analyses with GA as a continuous variable did not alter the conclusions substantially (see table 3). We found a statistically significant increase in FIQ of 0.08 points [95% CI; 0.01, 0.15] per increase in GA (in days) in the adjusted analysis. Similar estimates were seen in the analyses of VIQ and PIQ, however we found no statistically significant associations in the adjusted analyses. For teacher-assessed executive function, we found a statistically significant decrease in GEC and MI of -0.07 points [95% CI; -0.14, -0-01] per increase in GA (in days) indicating better executive function with increasing GA, however these estimates also became insignificant when adjusting for potential confounders.”
However, we are concerned that the elaboration of both the results of GA as a continuous and categorical variable will make the interpretation more difficult for the reader, as the results are much alike.
P13 l.4: please also see earlier comments. Could the authors please clarify whether “significant” refers to statistical and/or practical significance.
Revised according to the remarks of the reviewer. Thank you.
Discussion
P14 l.7: the authors highlight the large sample size of their cohort as one of the main strengths of this work. While I agree that the cohort itself is large, a major limitation of their study is the small number of subjects that were in the group of “moderately or very” preterm birth, which is the main explanatory variable in their model.
We thank the reviewer for the comment. We address this major limitation of having only a small number of children born preterm included in our study in the discussion, and we present possible explanations and implications.
P14 l.19: please add a copy of the study protocol to the supplementary material.
We thank the reviewer for the comment. Study protocols for both the original LDPS and the current study are available from the authors. There is also a study design article available in the references (reference #12).
P14 l.34: substantially more information on power analyses (which have not been presented at all) are necessary here.
We thank the reviewer for the comment. We have provided a post hoc power analysis in the discussion section:
“A post hoc power analysis showed that analyses comparing very or moderately preterm birth (n=8) with birth at term (n=1728) had a power of 0.48, 0.28, and 0.59 for FIQ, VIQ, and PIQ outcomes, respectively.”
P15 l.6: please see previous comments regarding “significance”.
Revised accordingly. Thank you.
P15 l.41: I suggest the authors refrain from using language such as “borderline statistically significant”.
Revised according to the remarks of the reviewer. Thank you. The text now states:
“When assessing attention measures, we only found one statistically significant result, which might be because of chance alone.”
Conclusion
P16 l.40: considering that very limited details have been provided regarding the mediation analysis, further work is needed for this to be included as one of the main conclusions of the present work. Please see relevant comments above.
We thank the reviewer for the comment. We have addressed this issue above (major revision #4 and #6).
P16 l.51: I do not agree that based on the current findings we can conclude that low GA has a crucial role in determining cognitive abilities. A more careful evaluation of their findings in light of uncertainty may be more appropriate here.
Revised according to the remarks of the reviewer. Thank you. The text now states:

“Therefore, GA has an essential role in determining cognitive abilities independent of maternal IQ and parental educational level.”
Table 1
Consider replacing “gender” with “sex”.
Revised according to the remarks of the reviewer. Thank you.
Table 2
In the table legend the authors state the number of participants for whom data are available. These numbers ought to be reported for each term group.
We thank the reviewer for the comment. We have now reported these numbers in the legends of tables 2 and 3. We have only presented information on missing data for preterm births, as these are the most interesting numbers and it would fill up too much space to also mention the missing data for term births. The number of missing term births can be calculated by subtracting the missing preterm births from the difference between 1776 (overall number of participants) and the actual number of cases in each analysis.

VERSION 2 – REVIEW

REVIEWER	Abdool S. Yasseen University of Toronto
REVIEW RETURNED	29-Jun-2019

GENERAL COMMENTS	Many thanks for addressing my concerns.

REVIEWER	Julian Mutz Social Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London
REVIEW RETURNED	11-Jul-2019

GENERAL COMMENTS	Review of bmjopen-2019-028982.R1 The authors have addressed a substantial number of my comments on the initial submission of their work. Nevertheless, I have some reservations about the appropriateness of the analyses involving subgroups of gestational age, and the conclusions that can be drawn from these findings. While the authors have stated in response to my original comment 1 that they have toned down their conclusions, the wording of the main conclusion has not changed in any meaningful way. The authors now state that “GA has an essential role in determining cognitive abilities” instead of “GA has a crucial role in determining cognitive abilities”. With respect to my original comment 2, the authors state that “[t]his is exactly why 95% CIs are provided (rather than just a p-value), as it tells the reader that we are 95% certain/confident that the true value in the background population lies somewhere between the two extremes. So in the example mentioned by the reviewer we are 95% certain/confident that the true difference in IQ between the two groups is in the CI interval which does not include 0, and thus is statistically significant at the 5% level. We assume this understanding and interpretation of a CI to be common knowledge among readers of BMJ Open.” The point that I wanted to make is that these confidence intervals are very wide and that a comment on the uncertainty surrounding these estimates would be appropriate. Please also refer to Greenland et al. (2016) regarding misinterpretations of confidence intervals.

	Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology. 2016 Apr 1;31(4):337-50. In response to my original comment 9, the authors state that "The analyses with GA as a continuous variable produced essentially the same results as the analyses with GA as a categorical variable". I doubt that this is an accurate representation / summary of their findings as there seem to be several discrepancies between the analyses of GA as continuous and categorical variable: "Similar estimates were seen in the analyses of VIQ and PIQ, however we found no statistically significant associations in the adjusted analyses. For teacher-assessed executive function, we found a statistically significant decrease in GEC and MI of -0.07 points [95% CI; -0.14, -0-01] per increase in GA (in days) indicating better executive function with increasing GA, however these estimates also became insignificant when adjusting for potential confounders." I also suggest rewording "insignificant" in the last sentence. Minor comments Article summary With respect to the statement "limited our power to detect any true differences as significant", I suggest removing "as significant". Materials and methods My previous comment "Danish Medical Birth Registry should not be abbreviated as this is no commonly known acronym" still needs to be addressed. Results The authors state "However, when these variables were removed simultaneously from the regression, most estimates became insignificant due to wider CIs." Please revise the wording "insignificant". Finally, it is worth mentioning that the post hoc analysis excluding early term births was requested by one of the reviewers.

VERSION 2 – AUTHOR RESPONSE

Review of bmjopen-2019-028982.R1
The authors have addressed a substantial number of my comments on the initial submission of their work. Nevertheless, I have some reservations about the appropriateness of the analyses involving subgroups of gestational age, and the conclusions that can be drawn from these findings.

While the authors have stated in response to my original comment 1 that they have toned down their conclusions, the wording of the main conclusion has not changed in any

meaningful way. The authors now state that “GA has an essential role in determining cognitive abilities” instead of “GA has a crucial role in determining cognitive abilities”.
We thank the reviewer for the comment. In the abstract and the conclusion, we have now changed the wording to:

“GA may play an important role in determining cognitive abilities independent of maternal IQ and parental educational level.”

With respect to my original comment 2, the authors state that “[t]his is exactly why 95% CIs are provided (rather than just a p-value), as it tells the reader that we are 95% certain/confident that the true value in the background population lies somewhere between the two extremes. So in the example mentioned by the reviewer we are 95% certain/confident that the true difference in IQ between the two groups is in the CI interval which does not include 0, and thus is statistically significant at the 5% level. We assume this understanding and interpretation of a CI to be common knowledge among readers of BMJ Open.” The point that I wanted to make is that these confidence intervals are very wide and that a comment on the uncertainty surrounding these estimates would be appropriate. Please also refer to Greenland et al. (2016) regarding misinterpretations of confidence intervals.
Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology. 2016 Apr 1;31(4):337-50.
We thank the reviewer for the comment and agree that some of the confidence intervals are quite wide. We have addressed this in the discussion, and the text states:

“Generally, the effect estimates are subject to some uncertainty illustrated by the wide CIs.[24]”
In response to my original comment 9, the authors state that “The analyses with GA as a continuous variable produced essentially the same results as the analyses with GA as a categorical variable”. I doubt that this is an accurate representation / summary of their findings as there seem to be several discrepancies between the analyses of GA as continuous and categorical variable: “Similar estimates were seen in the analyses of VIQ and PIQ, however we found no statistically significant associations in the adjusted analyses. For teacher-assessed executive function, we found a statistically significant decrease in GEC and MI of -0.07 points [95% CI; -0.14, -0-01] per increase in GA (in days) indicating better executive function with increasing GA, however these estimates also became insignificant when adjusting for potential confounders.” I also suggest rewording “insignificant” in the last sentence.

We thank the reviewer for the comment. We agree that there are some discrepancies between the analyses of GA as a continuous and categorical variable, and we have modified the text accordingly. The text now states:

“In analyses with GA as a continuous variable (see table 3), we found a statistically significant increase in FIQ of 0.08 points [95% CI; 0.01, 0.15] per increase in GA (in days) in the adjusted analysis. Similar estimates were seen in the analyses of VIQ and PIQ. However, we found no statistically significant associations in the adjusted analyses. For teacher-assessed executive function, we found a statistically significant decrease in GEC and MI of -0.07 points [95% CI; -0.14, -0-01] per increase in GA (in days) indicating better executive function with increasing GA, however these estimates also became statistically non-significant when adjusting for potential confounders.”

Minor comments

Article summary

With respect to the statement “limited our power to detect any true differences as significant”, I suggest removing “as significant”.

Revised according to the remarks of the reviewer. Thank you.

Materials and methods

My previous comment “Danish Medical Birth Registry should not be abbreviated as this is no commonly known acronym” still needs to be addressed.

Revised according to the remarks of the reviewer. Thank you.

Results

The authors state “However, when these variables were removed simultaneously from the regression, most estimates became insignificant due to wider CIs.” Please revise the wording “insignificant”.

We thank the reviewer for the comment. We have reworded the sentence, and the text now states:

“However, when these variables were removed simultaneously from the regression, most estimates became statistically non-significant due to wider CIs.”

Finally, it is worth mentioning that the post hoc analysis excluding early term births was requested by one of the reviewers.
We thank the reviewer for the comment. We believe that the current wording in the manuscript is appropriate. However, if the editor thinks that we should emphasize that the post hoc analyses was requested by one of the reviewers, we will of course accommodate this.

VERSION 3 – REVIEW

REVIEWER	Julian Mutz Social Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, United Kingdom
REVIEW RETURNED	02-Aug-2019

GENERAL COMMENTS	Review of bmjopen-2019-028982.R2 The authors have sufficiently addressed the majority of my comments. I have no further concerns.