Author’s response to reviews

Title: The effect of a lifestyle intervention in obese pregnant women on change in gestational metabolic profiles: findings from the UK Pregnancies Better Eating and Activity Trial (UPBEAT) RCT.

Authors:

Debbie Lawlor (D.A.Lawlor@bristol.ac.uk)
Harriet Mills (harriet.mills@bristol.ac.uk)
Nashita Patel (nashita.r.patel@kcl.ac.uk)
Sara White (sara.white@kcl.ac.uk)
Dharmintra Pasupathy (dharmintra.pasupathy@kcl.ac.uk)
Annette Briley (annette.l.briley@kcl.ac.uk)
Diana Santos Ferreira (Diana.SantosFerreira@bristol.ac.uk)
Paul Seed (paul.seed@kcl.ac.uk)
Scott Nelson (Scott.Nelson@glasgow.ac.uk)
Naveed Sattar (Naveed.Sattar@glasgow.ac.uk)
Kate Tilling (Kate.Tilling@bristol.ac.uk)
Lucilla Poston (lucilla.poston@kcl.ac.uk)

Version: 3 Date: 26 Sep 2018

Author’s response to reviews:

Dear Diana Samuel (BMC-Medicine Editor)

Thank you for the opportunity to respond further to Reviewer #4’s comments.

We have addressed all of his comments and provide a point by point response below. In undertaking all of these analyses we noticed an error in one set of the previous results relating to the change in metabolites across pregnancy in standard deviation units (the results presented in Figure 2 and the final column of supplementary table 3. This was an honest typographical error in one line of code. We have repeated all of those results and updated results in the paper. The
effect sizes are smaller but patterns and overall conclusions for those results are unchanged. The main results for the effect of the RCT on metabolite change in pregnancy were not affected at all by this error and remain unchanged from all previous submissions (original and revisions). All analyses in the revised paper have now been checked independently by two statisticians with both having identical results for all tables and figures. In relation to this the revised paper in addition to the changes documented below in response to reviewer #4. The following changes have been made:

- Changes to some numerical results in the abstract page 5 between lines 101 and 105
- Changes to some numerical results in the text of the main manuscript page 14 between lines 310 and 321
- Changes to all results in the final column of supplementary table sTable 3 (Supplementary material pages 13-18)

All of the above changes can be observed with tracked changes on the marked revision 3 submission

- In addition, we have uploaded a new Figure 2 (all panels a to c) with the corrected results.

Responses to Reviewer #4:

1. You cite 23 to support the claim that non-normality is not an issue.

But this paper does not take into account differences in skewness, which is known to be a serious concern. It was done using only one skewed distribution. This is not very convincing when you look at other simulation studies dealing with robustness issues. Consider something simple: the paired t-test. It is known that differences in skewness from time 1 and time 2 can be a serious concern (e.g., Wilcox, 2017).

Apologies for not making this clear in my last report. A general pattern is that as we move toward more involved models, robustness becomes an increasingly serious concern when using any method based on means and variances.

Response: There is evidence that the fixed effects are robust to misspecification of the random effects. However, the standard errors of the fixed effects are less robust. We have amended this statement in the manuscript. (Page 12, Lines 263-264). We have also undertaken eight additional (to the GEE sensitivity previously undertaken in response to this reviewer’s request) sensitivity analyse that deal with skewness, heteroskedasticity and outliers (please see response to point 4 below).
2. Suggest not saying exact p-value, just say p-value. It is not exact although it might be close.

Response: We have made this change and removed the word ‘exact’ (Page 10, line 225)

3. What is a robust standard error? In the supplemental material this is not made clear, it simply refers to the main paper. If you are using a robust measure of location and its corresponding standard error, this would be very useful. But have the sense that this is not the case.

Response: The robust standard error here is the standard error calculated using the White-Huber sandwich estimator. So the standard error should be robust to e.g. heteroskedasticity, but is not a robust measure of location. Apologies if we were unclear about this in the previous response – we were using robust standard errors to avoid assumptions about heteroskedasticity, rather than a robust measure of the standard deviation (or any robust measure of location). We have added the following to the manuscript:

“The robust standard errors were calculated using the calculated using the White-Huber sandwich estimator and are robust to heteroskedasticity.” (page 12, lines 268-270)

4. In your response you state:

The standard deviation is only used to provide an appropriate scale comparison for the different metabolites, as they are all measured on such different scales.

Yes, but this does not deal with the lack of robustness associated with the standard deviation. So it remains unclear whether a robust measure of scale might be more informative.

In summary, the paper would benefit from a more robust look at the data, and the results would be more convincing if more effective methods were used to deal with non-normality. I am not saying that the results are wrong or misleading. But there is uncertainty about whether this is the case or whether importance features of the data are being missed.

Response We now also transformed all variables using the IQR as the measure of scale – i.e. by first scaling all measures by dividing them by the IQR at the first timepoint/1.34. The 1.34 multiplier is because if the variable were normally distributed then the IQR=approximately 1.34*SD.

We have then conducted the following sensitivity analyses:

1) Multilevel model using SD as the scale comparison. The parameter estimate should be unbiased but the standard errors may be affected by non-normality of the measures.
2) Multilevel model using IQR as the scale comparison. This analysis should be robust to non-normality in the original measure, and the parameter estimate should be unbiased but the standard errors may be affected by non-normality of the measures.

3) A paired t-test (final measure-first measure), using bootstrapping to obtain standard errors, and with the SD of the first measure (in the control group) used as the scale comparison. This analysis should be robust to non-normality in the differences (via bootstrapping).

4) A paired t-test (final measure-first measure), using bootstrapping to obtain standard errors, and with the IQR used as the scale comparison rather than the SD. This analysis should be robust to non-normality in the original measure (using the IQR rather than the SD), and also robust to non-normality in the differences (via bootstrapping).

Repeated 1-4 above but first removing 1% of all outliers (top and bottom) at each timepoint and by treatment group. This should examine sensitivity of the conclusions to outliers.

All sensitivity analyses give broadly similar results, with correlations between estimates from the different methods $>0.9$ There are differences in the selection of results as “significant”, which could potentially be due to incorrect distributional assumptions; influence of outliers; or because when the data are in truth Normally distributed, then the non-robust tests are more efficient.

We have added the following text to the main manuscript methods section:

“We undertook a set of eight further sensitivity analyses using different approaches to examine the sensitivity of our conclusions to possible heteroskedasticity, skewness and data outliers. Additional details of the linear spline model and its assumptions, including assumptions related to missing repeat measurements, and these additional eight sensitivity analyses are provided in supplementary text (sText).” (Page 12, lines 270-275)

We have added the following text to the supplementary material methods:

“We conducted the following additional sensitivity analyses to explore how robust our results were to heteroskedasticity, skewed distributions and outliers:

1) Multilevel model using SD as the scale comparison. The parameter estimate should be unbiased but the standard errors may be affected by non-normality of the measures.
2) Multilevel model using IQR as the scale comparison. This analysis should be robust to non-normality in the original measure, and the parameter estimate should be unbiased but the standard errors may be affected by non-normality of the measures.

3) A paired t-test (final measure-first measure), using bootstrapping to obtain standard errors, and with the SD of the first measure (in the control group) used as the scale comparison. This analysis should be robust to non-normality in the differences (via bootstrapping).

4) A paired t-test (final measure-first measure), using bootstrapping to obtain standard errors, and with the IQR used as the scale comparison rather than the SD. This analysis should be robust to non-normality in the original measure (using the IQR rather than the SD), and also robust to non-normality in the differences (via bootstrapping).

Repeated 1-4 above but first removing 1% of all outliers (top and bottom) at each timepoint and by treatment group. This should examine sensitivity of the conclusions to outliers.” (Supplementary material page 5, Lines 137-155)

To the in the main manuscript results:

“All additional sensitivity analyses give broadly similar results to our main analyses, the generalised estimating equation method and each other, with correlations between estimates from all of the different methods >0.9. There are differences in specific results that reach false discovery corrected statistical “significance”, with some p-values being slightly higher or lower in the sensitivity compared with our original methods. This could potentially be due to incorrect distributional assumptions, influence of outliers, or because when the data are in truth Normally distributed, and therefore the non-robust tests are more efficient. Significance tests for mean differences in extremely large and very large VLDL concentrations between the intervention and control arm, in particular were very similar across different methods (a full set of results for all of these analyses can be found in Supplementary File 2). (Pages 15-16, lines 350 to 360)

And to the main manuscript discussion:

“The consistency of findings between our main analyses, general estimating equations and eight additional sensitivity analyses using different methods address possible non-normality, heteroskedasticity, skewness and outlines, support the robustness of our modelling approach.” (Page 17, lines 386-389)
We thank the reviewer for their further comments and we apologise for the error that has now been identified and rectified in some of our earlier results.

We look forward to hearing further from you

Yours Sincerely,

Prof Deborah A Lawlor, on behalf of all co-authors