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@glagow.ac.uk)
Naveed Sattar (Naveed.Sattar@glasgow.ac.uk)
Kate Tilling (Kate.Tilling@bristol.ac.uk)
Lucilla Poston (lucilla.poston@kcl.ac.uk)

Version: 4 Date: 31 Oct 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 his remaining two comments and provide a response to these below.
Responses to Reviewer #4:

Have only two remaining concerns.

In your response 4, you say:

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.

1. Removing outliers among a dependent variable and applying some standard method to the remaining data results in an inaccurate estimate of the standard error. This is a serious problem regardless of the sample size. Details are in the Wilcox book mentioned in my last report. Technically sound methods for dealing with this are described in this book as well.

2. I often encounter situations where 10% or more of the data are outliers based on a boxplot or the MAD-median rule. It is unclear why removing 1% suffices.

RESPONSE

We have dealt with both of these two points together and they refer to the same broad area.

Yes, we agree that removal of outliers may mean that observations are no longer independent. For this reason, in our previous sensitivity analyses, we used the bootstrap to estimate standard errors (as outlined in a recent paper by the reviewer that we now cite (reference [11] in the supplementary material) for all models (except the MLM, due to computational issues) where we removed outliers. We clarify this in the supplementary material and cite Rand Wilcox recent paper.

We have now re-done the paired t-test (and other analyses) using the MAD as the scalar comparison (i.e. dividing values by the MAD at the initial timepoint), as this is recommended as a robust deviation measure by the reviewer. Though we point out that the correlation between the MAD and the SD in our data set (across the 158 metabolites) was > 0.98, and the results do not change in any material way.

We agree that careful considerations are required when removing outliers and we should probably have been clearer about the considerations we had taken in this regard. Namely, that when examining outliers, it is very important to distinguish between true outliers (erroneous results) and influential observations which are far from the average values on either exposure or outcome. Here, the exposure is a binary variable and thus we are examining whether there were a few influential observations which are at extreme values for a metabolite, given the group they are in. We would not expect there to be a large number of these. 1% seemed a conservative estimate, and was borne out by examining the proportions of values exceeding the MAD*3.5
cutoff as recommended by Ramsey and Ramsey (now cited – reference [12] in supplementary material. Removing real observations that are far from the average can result in selection bias, therefore care is required in decided how to remove outliers, and that is the approach we took here. However, we have now MAD-median rule to select outliers as suggested by Rand Wilcox, which in our dataset means a \(2.24\times\text{MAD}\) threshold. For the vast majority of the metabolites this removes < 1\% of the top and bottom as outliers and results are not different from the original (1\%) threshold that we selected.

We prefer to consider the analyses removing "outliers" as sensitivity analyses - because most of these "outliers" are likely to be true observations, not errors. Thus we are not aiming to test hypotheses using datasets with "outliers" removed.

We have also added to the sensitivity analyses quantile regression (median and upper IQR) for the difference between initial and final value, with and without outliers.

With these additional analyses changes are:

In the main paper

‘eight’ (referring to the number of sensitivity analyses) is changed to ‘14’ in the methods page 12, line 270 and also page 12, line273.

The following has been added to the results section:

“…. (a full set of results for all of these analyses can be found in Supplementary File 2, with some additional discussion in Supplementary Material (sText).” [Page 16; lines 358-359]

In the supplementary material

The description of the additional sensitivity analyses has changed to address all of the above points:

“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). Furthermore, when removing outliers (see below) bootstrapping will ensure that the standard error is correct.[11]

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). Furthermore, when removing outliers (see below) bootstrapping will ensure that the standard error is correct.

5. A paired t-test (final measure-first measure), using bootstrapping to obtain standard errors, and with the Median Absolute Deviation (MAD) used as the scale comparison rather than the SD or IQR, as one reviewer[11] suggested this was a more robust measure of deviation, though in our analyses the correlation between the MAD and SD was > 0.98 across all metabolites and the two gave virtually identical results.

Repeated 1-5 above but first removing all outliers (top and bottom) at each timepoint and by treatment group. We initially removed the top and bottom 1% in these analyses having considered the following: (i) it is important to only remove measures that are highly likely to be erroneous and not true values that are markedly different to the mean for the study population, as the latter may introduce selection bias; (ii) in this study we were exploring a binary exposure (randomisation to intervention or standard treatment) and would not expect a high proportion of erroneous results at the extremes to influence our findings; and (iii) when we used the MAD*3.5 threshold for defining outliers as suggested by Ramsey and Ramsey,[12] for all of metabolites <1% were above this threshold (see final spreadsheet in Supplementary File 2). However, one of the reviewers[11] preferred that we use the MAD-median rule applied to our data, which, gave a threshold of MAD*2.24. For the vast majority of metabolites this also resulted in fewer than 1% at top and bottom being removed.

Results across these 10-sensitivity analyses (the listed 5, with and without removal of the top and bottom MAD*2.24 defined outliers) were very similar with correlations between each other and with the main results all > 0.9. Removal of outliers did not notably alter any results.

Additionally, we undertook four further sensitivity analyse; median quantile regression and 75th centile quantile regression, both with and without outliers removed based on the MAD*2.24 rule. The overall pattern of results were similar to those of the main analyses and all other sensitivity
analyses, with some evidence of larger differences in the upper quartile of metabolites than around the middle 50% (see Supplementary File 2).

All results for these sensitivity analyses are shown in Supplementary File 2.

Multilevel models of change over time allow all participants with at least one measure to be included in analyses under the assumption that data are missing at random. The statistical term ‘missing at random (MAR)’ differs from missing completely at random (MCAR) in that it does not mean missingness is independent of all other characteristics. It means that conditional on the covariates included in the model (here age, parity, ethnicity, BMI and study centre) and the observed repeat measurements, the missing repeat measurements are not systematically different to those observed. This means that the effect of the intervention in those with some missing metabolic profile data is the same as in those with complete data at all three-time points conditional on the covariables included in the model. The MAR assumption would also be necessary if we restricted analyses to only those will all three repeat measurements (i.e. a complete case analysis). Whilst we cannot directly assess this assumption we feel that it is unlikely to be violated given most women had all three repeats, with just 16% having only one measure and that loss to follow-up in the trial was minimal and similar in both arms. The standard errors, and hence 95% confidence intervals, in these models take account of the greater random error of predicted levels at any time in those with just one or two of the repeat measurements.” [Pages 5-6; lines 137-200]

A new revised Supplementary File 2 with all of the updated analyses has been uploaded.

We look forward to hearing further from you

Yours Sincerely,

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