**Multiple Testing: To Correct or Not to Correct?**

In our manuscript, we hypothesized that exposure to prenatal (lifetime) maternal depression would lead to higher levels of lymphocyte proliferation (LP), pro-inflammatory cytokines (IL-6, TNF-alpha) and IL-13, and lower levels of anti-inflammatory cytokines (IL-10) produced by CBMC collected at delivery.

In our primary analyses, we did not use a statistical method to analyze patterns of response. Instead, we had specific *a priori* hypotheses and examined the results in light of each specific hypothesis. According to our reasoning, these results cannot simply be considered multiple independent tests on a common family of outcomes. Instead, we evaluated 4 sets of cytokines, each in medium (unstimulated) along with 3 stimulated assays. We could consider assays across similar types of cytokines (i.e., pro-inflammatory) and different forms of stimulation (e.g., PHA, BLA, DER) to comprise families of outcomes. However we would not consider all cytokines in medium and across all types of treatment to comprise a single family of outcomes. Thus, for the purposes of considering issues related to multiple testing, we considered cytokines separately as follows:

1. We considered lymphocyte proliferation as an independent outcome with three treatments: stimulated with PHA, BLA, DER. Had one of these been significant we might have applied a multiple testing corrections but this was unnecessary, as all tests were null.
2. We considered TNF-alpha, IL-6, and IL-13 as related pro-inflammatory outcomes each with four treatments: medium alone; stimulated with PHA, BLA, DER. However, medium alone may look quite different from stimulated treatments, and therefore we consider this as a separate test of the hypothesis. IL-13, because of an excessive proportion of samples with assay values below the limit of detection, was not included in analyses. While we could consider the remaining 6 separate stimulated assays as a family of tests, this was unnecessary as the uncorrected p-values for all stimulated assays with these two cytokines were non-significant.
3. We considered IL-10 as an anti-inflammatory outcome with four treatments: medium alone; stimulated with PHA, BLA, DER. Again, because medium along may look quite different from stimulated treatments, we consider this as a separate test of the hypothesis.

It is worth noting that decreasing type I errors (e.g. by using techniques such as Bonferroni adjustments) inflates type II errors (the probability of accepting the null hypothesis when the alternative is true) (1). Thus, we feel that it makes sense to focus on the patterns, magnitude, and consistency of our results, in the context of the biology and *a priori* hypotheses (2). If multiple testing were an issue, we might expect to see a significant p-value for one of the tests purely by chance. We would not, however, expect to see a consistent pattern in the responses across a family of tests. Figure 1 in the manuscript shows that all three associations for maternal depression with stimulated IL-10 are negative, albeit only one reaches statistical significance. Given the somewhat limited sample size and consistent findings across all stimulated assays, this suggests the finding is small but reliable.

However, in order to test the robustness of our finding, we evaluated our finding after correcting for multiple testing across the three stimulated IL-10 treatments. We used the FDR (false discovery rate) method, which balances the risk of committing a type I and a type II error. We set the false discovery rate to 10% (a common threshold, meaning that 10% of significant tests might be truly null). Setting the false discovery rate provides a balance between the likelihood of committing a Type 1 (false positive) or a Type II (false negative) error. After correcting for multiple testing across the three stimulated treatments (3), the p-value for Bla g2 remained significant. Thus, we believe our interpretation of the findings, that maternal depression is associated with one of the cytokine outcomes considered (IL-10, anti-inflammatory) is valid.

References

1. Perneger TV. What’s wrong with Bonferroni adjustments. BMJ: British Medical Journal. 1998;316:1236.

2. Braun JM, Gennings C, Hauser R, Webster TFJEhp. What can epidemiological studies tell us about the impact of chemical mixtures on human health? Environmental health perspectives. 2016;124:A6.

3. McDonald JH. Handbook of Biological Statistics (3rd ed.). . Sparky House Publishing, Baltimore, Maryland.254-60.