Reviewer's report

Title: General movements in preterm infants undergoing osteopathic manipulative treatment (OMT) : a randomised controlled trial.

Version data: May 2015.

Reviewer's report:

In general this ms. is much improved. However some major problems remain, the most serious of which concern the statistical analysis. There are two problems there: 1) the absence of a test which seems necessary to support one of the articles main claims, and 2) inadequate and confusing reporting of the tests that were conducted.

1) The only tests of outcome reported are for significance of change between baseline and endpoint within each group. Those tests are reported for both global GM assessment and for MOS. For the MOS no significant change is found in IG group, but a significant decline is found in the CG group. The text the repeatedly uses the phrase "compared to" to contrast the two groups with regard to those findings. However no test of the statistical significance of the comparison between group difference in MOS change scores is reported. In the absence of such a test, nothing can be said definitively as to the significance of that between groups difference. Any reader with even a little statistical knowledge would be surprised that a test comparing change in MOS between groups was not reported; it is so obviously necessary to support the claim that the text tries to make, i.e. that the between group difference in change scores is significant. For that reason I encourage the authors to perform such a test and to report it.

In this connection the multiple time points for data that are illustrated in Figure 2 suggest that a much more powerful test than a simple comparison of baseline to endpoint scores is possible. A longitudinal regression model would derive greater statistical power to determine the average rate of change in MOS for each group by using all of the data points collected at each time point. This would have greater power to detect a significant difference in rate of change in MOS between the two groups.

2) In addition to the above insufficiency of analysis, there are several confusions and omissions in the way in outcomes the tests which were conducted are reported. These are detailed below.

If these problems of statistical analysis and reporting were remedied, then the
ms. would be suited for publication. The study itself is actually quite interesting and innovative, and would be of significant interest to those concerned with neonatal health.

I encourage the authors again, as in my previous review, to consult an experienced biostatistician to conduct the omitted tests and to format the reporting of statistical results more adequately..

Major essential revisions:

Summary: In the "results" section p values are given, but the tests which produced those p values are not identified.

* The words "compared to" introduce confusion. To most scientific researchers they imply that a statistical test was employed to compare the reductions between the two groups. However the p values that are reported in this paragraph seem to have been produced by comparing baseline to endpoint measures within each group. No test for significance of the between group difference in changes is reported.

* Lines 164-65: "primary analysis" should be changed to "primary outcome." similarly "secondary analysis" should be changed to "secondary outcome"

* LInes 172-73 "The maximum composite score of 42 indicates the most optimal GM performance." This is confusing, you previously describe "the global GM" assessment, so it's easy for the reader to understand "GM" as an abbreviation for that assessment (In lines 209-214 you do use "GM" as an abbreviation for that score). It would be less confusing for the reader if there was a clear abbreviation for the "global GM assessment" (e.g. GGM, ) and another abbreviation for the Motor Optimality Score, (MOS). Then when you discuss the MOS outcomes, you should not use the abbreviation that indicates the global GM assessment.

* Line 178 "Statistical analysis" does not mention the use of Pearson Chi Sq test, although you report results of such tests in lines 211 and 213. On line 225-26 you report results from Cohen's kappa test.

Figure One does not conform to current standards for study flow diagrams. Please consult Schultz et al. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Trials, 2010 11:32 for guidance. For example - the study flow should include at the top the number of candidates screened, the number excluded and the most common reasons for exclusion, the points at which data were collected should be indicated, there should be a separate box at the bottom labeled "analysis" and giving the number of subjects in each treatment group whose data were included in analysis.

* Table 1 "Demographic characterstics" should include only rows 1-4: Gender, GA at birth, birth weight, GA at first video recording. Lines 5 and 6 are study outcomes and should be presented in a separate Table 2 titled "Study Outcomes." In that Table 2 the final column should not include conclusions such as "The GM Optimality Score of the IG remained unchanged over time." Those statements should appear in the Outcomes and Conclusions sections of the text. The final column should simply present the p values and labeling adequate for the reader to understand what test produced them. The sources of p values in
both Table 1 and 2 should be identified. Usually this is done in footnotes or legends to the table, e.g. "p values from 2 sided Fishers Exact test."

* Line 209-223: There are several omissions and confusions in how the statistical outcomes are presented here. Line 209 "The global GM judgment showed ..." I think you mean "the global GM judgement at baseline showed" "There was no difference between the two groups" again I think you mean "There was no difference between the two groups at baseline" In "The quality of GMs did not change during intervention" "during intervention" is confusing; better to say "The quality of GMs had not changed in either group by endpoint" Lines 212-13: The list "(IG: 2N, 9 PR 1 CS; CG: 5 N 7 PR 1 CS)" appears to be a list of values at endpoint. This is confusing because the immediately preceeding sentence refers to change between endpoint and baseline, but the list is not a list of change scores. It would be less confusing to provide another Table 3 that gives the counts for each rating category at baseline and at endpoint for each group.

* Lines 210-213 At the end of each list the counts for each catory of global GM you have "Pearson Chi square n.s." but you do not explain whether the Pearson test was a comparison between the two groups, or a comparison of counts at baseline with counts at endpoint. If you tested only between the two groups, then you cannot claim that there was no significant change between baseline and endpoint. In order to make that claim you have to test for significance of differences between baseline and endpoint.

* Lines 216-23: "the Motor Optimality Score showed a significant difference over time between IG and CG." As noted above, you do not report a test of the significance of the between group difference in MOS change. Without such a test you cannot claim that the difference in MOS change between the two groups was significant.

Discussion

* Lines 230ff "Overall, the MOS showed i) a significant difference over the time between IG and CG," but you still have not provided a test for that, so it cannot be claimed. "ii) a significant decrease in CG ..." .."and ii) the MOS remained unchanged in IG." Again the phrase "over the time" is confusing; better to use "baseline to endpoint" Also the reader should at this point in the text be referred to Table 2.

lines 261-63 "For this reason the application of OMT..." Should be changed to "This suggests that the application of OMT..."

* Figure Two As the text stands Figure 2 is unnecessarily confusing. The analysis presented uses only the baseline and 3 week endpoint data; the graphic display in Figure 2 of the data points and error bars for all the intervening assessments is needlessly confusing since they were not used the analysis. If you are not going to report a longitudinal analysis that uses all of those intermediate data points, then it would be much less confusing for Figure 2 to display only the baseline and endpoint values for each group, along with error bar for each of those. Figure 2 would then have only four data points and four error bars; much easier to understand.
Minor essential revisions
* Conclusion "we were able to show that healthy preterm infants" should be "We were able to show that a group of healthy preterm infants".
* Line 314 Again the phrase "compared to" is confusing. The text does not report test of the statistical significance of the difference in MOS changes between the two groups.
* Lines 317-18: "Future studies should include in the control arm a control for... " should be "Future studies should include a control for touch and presence..."
* The ms. would benefit from another round of editing for English. At many points articles and prepositions are missing or inappropriate; e.g. line 3 "as ideal form of osteopathic..." line 36 "however, stressful environment..." lines 59 "and general movements assessment..." line 279 "A high incidence of PR GMs....and do not" should be "A high incidence of PR GMs....does not"
* line 78 "NICU" some readers will not know what this means, better to spell it out on the first occurrence.

Discretionary revisions
* lines 290-92: "Whereas the score of the CG" should be "Whereas the MOS score for the CG..." Also relocate this statement to the beginning of the Discussion; that is where the main findings should be summarized.
* lines 65-68: "Assessment of GM have been.. to predict neuromotor development" might be better placed in the Methods section where the clinometric properties of measurement instruments are generally stated.