UNIVERSITY OF OXFORD

NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY,
& MUSCULOSKELETAL SCIENCE (NDORMS)

BOTNAR CENTRE
NUFFIELD ORTHOPAEDIC CENTRE
HEADINGTON • OXFORD • UK
OX3 7LD
Telephone: 01865 737643
Fax: 01865 227671

14th June 2018

Dear Dr Loder,

Thank you very much for your further consideration of our manuscript. We discussed your recommendations in detail and believe we have now addressed these in this latest revision.

In this latest version, we removed the statement that the difference between treatment groups exceeds the pre-specified minimum clinically important difference from the abstract. In addition, whenever the difference in outcome scores between groups is discussed, this is now followed by a reference to the fact the lower boundary of the confidence interval is less than the minimum clinically important difference. We also made further changes to ensure our interpretation of the study results is more balanced, and in particular, that a proportion of patients will not achieve a clinically important improvement after arthroscopic hip surgery.

Although outside of the trial statistical analysis plan, we have now included further subgroup analysis and explored the differential treatment effect of age and baseline Hip Outcome Score as linear continuous variables in the primary analysis model (non-linear fits were also explored). In addition, we explored the differential treatment effect of age and baseline Hip Outcome Score as categorical variables, and the results of this analysis are displayed as forest plots in the supplementary figures.

Our detailed description of the missing data imputation has now been moved from the supplementary data to the main manuscript. A detailed description of our sensitivity analysis is

NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY,
& MUSCULOSKELETAL SCIENCE (NDORMS)

BOTNAR CENTRE
NUFFIELD ORTHOPAEDIC CENTRE
HEADINGTON • OXFORD • UK
OX3 7LD
Telephone: 01865 737643
Fax: 01865 227671

contained within supplementary figure 2, and this could also be moved to main manuscript at your request.

We hope that our responses satisfactorily address your recommendations and we would be entirely agreeable to making further changes.

Yours sincerely,

Antony Palmer MA BMBCh DPhil FRCS (Tr and Orth)

NIHR Academic Clinical Lecturer in Trauma and Orthopaedics

Responses to Reviewer Comments:

1) I did not understand the following sensitivity analysis "Primary analysis was also repeated with the baseline 'expectation' HOS ADL as a covariate" – what is the expectation of the baseline value and how was it calculated? Why is this needed over and above a simple adjustment for observed baseline value?

Participants were asked to complete an additional HOS ADL score at baseline to express their expected outcome from treatment. At the conception of the trial, we hypothesised that patient expectation may influence treatment outcome, hence included this 'expectation HOS ADL' score as a covariate, as per our study protocol. A reviewer also requested that we explore the effect of patient expectation on treatment outcome. We did not find an association between patient expectation and treatment effect.

2) As mentioned before, the pre-defined subgroup analyses categorise age and baseline value at arbitrary cut-points. This loses information, and renders the investigation meaningless. E.g. those at age 39 cannot really be that different from those age 41, and yet the cut-point of 40 is used. Indeed, this age analysis identifies a difference in those above and below aged 40 – but how can we interpret this? The treatment effect does not truly jump when moving from 39 to 40 year old. Better would be to consider age and other continuous variables their original continuous scale, and even allow a spline (non-linear) function.

I asked the authors about this upon first revision, and they do not revise the manuscript in this regard. They do explain to me that they did do some other analyses (e.g. with age as linear), but do not include them. However "they would be agreeable to including this data at your request". I would like this to be addressed please. They argue against including age as linear saying that "We have concerns over the robustness of this observation due to the limited data in older age groups, the dataset not being sufficiently powered for this analysis, and the likelihood that the effect of age may not be adequately modelled by a linear relationship"

- however, I do not find these arguments as concerning as choosing the 40-year cut-point (which is meaningless as mentioned, and loses about 1/3 of the power when age is modelled as continuous). There are many references on the need to handle continuous variables as continuous, such as refs 1 2 An excellent blog on this subject is here: http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous

We acknowledge the points made with respect to the loss of power by categorising a continuous variable. In response to this comment, we thoroughly explored the appropriateness of performing this additional subgroup analysis amongst the statistical team at the Centre for Statistics in Medicine and Oxford Clinical Trials Unit. The consensus opinion was that detailed subgroup analysis, including the use of continuous covariates, lies outside the scope of this manuscript and does not change the clinical message.

In addition to our concerns over the robustness of this additional analysis, further reasons why we have not performed the analysis are that i) it would deviate from our pre-specified statistical analysis plan for the primary publication of this trial ii) the intended purpose of our exploratory subgroup analysis was hypothesis generation rather than guiding clinical practice.

The exploratory categorical subgroup analysis was based on the minimisation thresholds used for randomisation and was intended to guide future analysis (ideally in larger pooled datasets with adequate power). We further discuss the caveats of our categorical analysis in the manuscript.

3) The authors note that the observed difference 'exceeded the MCID'. I would not use the MCID abbreviation for minimally-important clinical difference. Moreover, the CI for the true treatment effect (6.4 to 13.6) contains values below 9 (the defined MCID), and so – if the authors truly want to make a strong statement about whether the effect is above the value of 9 – it would seem important to also state that some of the evidence is in accordance with a value below 9. This is often not mentioned, however.

Therefore, there is a tension in the current manuscript about whether there is strong evidence of an important treatment effect here, or actually whether it is inconclusive.
The authors often draw attention to the effect being greater than 9, but do not draw attention to the CI containing values < 9.

This issue was raised in the first review. But it remains prominent (e.g. see abstract, results, and start of discussion) in many places without a more balanced view. The authors have added a note in their conclusion at the end of the Discussion saying: "However, further research is required to identify patients most likely to benefit from intervention given a significant proportion of patients did not achieve a clinically important improvement". Though potentially true, this is not the same issue. Variability in patient responses may indeed make some patients respond with a smaller improvements than other patients. However, the main focus of the original MCID is on whether the overall (mean) effect (across all patients) is at least 9. Therefore, the main question is whether the mean effect is clinically important, let alone whether all patients would always get a big improvement themselves.

The further discussion about whether all patients achieve a clinically important benefit is a further complication. In their response it is stated that "However, only 51% patients receiving arthroscopic surgery exceeded this MCID (assuming the MCID of 9 points is valid for our cohort). " – but I do not see how 9 is derived for an individual patient. The MCID relates to a 9 mean point improvement in one group compared to another. But at the patient level, there is no direct comparator. So we cannot say this patient did 9 better than if they had been in another group. So, do the researchers refer here to a patient improving by at least 9 to be important (without comparison). But why is 9 points relevant as both an absolute change (for the individual) and a difference (in the mean difference in one group compared to another)? More clarity is needed.

May I suggest that the authors revisit all their discussion about clinically important differences in the paper. I understand why MCID are useful in sample size calculations, but I find the current emphasis on this in the results and interpretation to be quite confusing and I sense other Editors do too.

We acknowledge that confidence intervals overlap the pre-specified minimally clinically important difference of nine points and have made further modifications to the manuscript to ensure a more balanced view. In the abstract, our statement that the increase in HOS ADL exceeds nine points directly follows the stated confidence intervals, hence here we have not modified this text.

We also appreciate the confusion with respect to our use of the term MCID. The minimally clinically important difference of nine points (between groups) (Martin et al 2008) was used for our power calculation and assessment of treatment effect (primary outcome measure).

However, with respect to the clinical translation of our trial findings, when making treatment decisions, what is more valuable is to estimate the relative proportion of patients that receive a clinically important change (within the individual).

An estimate for minimum clinically important change (within the individual) for HOS ADL was not available at the time of study design, but has since been estimated at five points, although the minimum detectable change (within the individual) was nine points (Kemp et al. 2013).

We used a threshold of nine points (minimum detectable change and a clinically important change) to evaluate the proportion of patients who reported a clinically important change in symptoms (Kemp et al 2013).

Our current revisions aim to resolve this confusion and justify our approach and interpretation.

4) Given the large missing data (e.g. only 80% were complete in the physio group), it would seem important to have a clear sub-section in the results about the findings when missing data were handled through imputation and other approaches. That is, at the moment it is just mentioned in a brief sentence and the supp material that missing data analyses did not change the findings. But this needs more prominence I feel. Related point is that more details are needed about exactly how the multiple imputation was done in the methods, and also to explain what the 'rctmiss command in Stata' is actually doing/ assuming. The response gives good details, but not the revision. May I also ask whether the outcome values were included in the imputation model, as recommended (e.g. see 3)

Multiple imputation using chained equations was performed in line with best recommended practice, using Stata's 'mi impute chained'. A linear regression model was used to impute missing outcomes for the HOS at 8 months post randomisation. Variables in the model included all covariates to be included in the analysis model (baseline HOS, age (continuous) and gender. In addition, other variables that were thought to be predictive of the outcome were also included: lateral centre edge angle, maximum alpha angle, Kellgren-Lawrence grade, and baseline HADS scores. Imputations were run separately by treatment arm. Imputations were based on a predictive mean matching approach, choosing, at random, one of the five values with the closest predicted scores.

Small amounts of missing data in the lateral centre edge angle, maximum alpha angle, Kellgren-Lawrence grade were also imputed based on the same variables, including the outcome variable using a multiple imputation by chained equations approach.

Other outcome data was not included in this imputation for the following reasons:

- *The HOS, as the primary outcome had less missing data than other outcome variables. Therefore no additional information was to be gained from including other, less completely observed outcome data at this time point.*
- *The 5 month follow-up was also less completely observed than the 8 month follow-up. Data collection for later follow-up points had not completed at the time of the analysis. Again, including these less completely observed variables would not have added information to the imputation model.*

As the analysis based on multiple imputation makes missing at random assumptions very similar to those made in the complete cases primary analysis, we did not expect this analysis to produce very different results.

The 'rctmiss' command facilitates missing not at random sensitivity analysis. For this study, 'rctmiss' was used to investigate scenarios in which participants with missing data in each treatment arm in turn were assumed to have outcomes that were, on average, up to 9 points worse than those with available data. We have now clarified the analysis for missing data in the methods section and added additional information and emphasis on the robustness of the trial conclusions to the results section and supplementary data.

5) The results per centre in the response are interesting, and the forest plot would be welcome addition to the supplementary material.

We have added this forest plot to the supplementary data section.