BMJ - Decision on Manuscript ID BMJ.2018.043530.R 3

Body: 02-Jul-2018

Dear Mr. Palmer,

Manuscript ID BMJ.2018.043530.R3 entitled "Arthroscopic Hip Surgery compared with Physiotherapy and Activity Modification for the Treatment of Symptomatic Femoroacetabular Impingement: A Multi-Centre Randomised Controlled Trial"

Thank you for sending us your revised paper. I'm afraid our statistician, Professor Riley, feels there is still quite a bit of work to be done. We'd like to try one more round of revision, after which if substantial concerns remain we will have to part ways.

You should also reference the other trials that have recently been published, e.g. FASHION, and discuss your trial in light of those results.

Very truly yours,

Elizabeth Loder, MD, MPH
eloder@bmj.com

*** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://mc.manuscriptcentral.com/bmj?URL_MASK=9c7d2fa0ff1c461e8f40379cedf18ba7

Reviewer: 1

Comments:
I thank the authors for their response to my comments, and I see they have also replied in detail to many other reviewers. So appreciate this was some task to go through. I think the paper is improved, but some concerns remain. Some fine-tuning is needed.

The message about whether the finding is actually clinically important is clearer (i.e. the CI contains values below an improvement that may be considered worthwhile), but perhaps needs to come across better in the conclusions of the abstract and the what this study adds.

My main concern is that I am still struggling with the subgroup analyses and the main subgroup results being based on continuous variables dichotomised. In their response (and indeed their discussion) there is a tension: on the one hand the subgroup results (and the difference response to treatment across individuals) get a lot of attention in the discussion (including in the what this study adds), but on the other hand the authors note the low power and the need for further research. I feel the paper would be far more straightforward if the focus was predominately on the overall treatment effect, and relatively less attention given to the subgroup results, for the following reasons (some of which repeat similar points).

1) In their response to me, they show better analyses where continuous variables are analysed as continuous. But they say 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 (other

fits were also explored), hence elected not to include this data in the manuscript" – rather the authors continue to focus on arbitrary cut-points in the paper, which I simply cannot agree with. I do not see any reason why assuming the treatment effect suddenly changes from aged 39 to 41 (i.e. at the 40 years cut-off) is plausible. Assuming a linear trend for age and interaction with treatment is far more plausible; moreover, if subgroups are really of interest that non-linear relationships could be examined. So I do not find the revision satisfactory in this regard.

2) Also, I am confused as the authors say in the response that they 'elected not to include this data in the manuscript' but in their revised methods section say "The differential treatment effect for age and baseline HOS ADL (as continuous variables) was further explored by adding an interaction term for treatment*age and treatment*baseline HOS ADL into the primary analysis model. Linear and non-linear effects (squared and cubic terms) for age and baseline HOS ADL were explored." So it actually is included?

Indeed, on page 20, the authors appear to have added in analysis of continuous variables on their continuous scale. However, in the text they only report the results for the effect of the variable on each treatment group separately, and then report the 'heterogeneity p-value'. The first relates to whether the variable (e.g. age) is a prognostic factor, and so this not relevant to the research question or subgroup effect. The 'test', I think, relates to a test of an interaction between the continuous variable and treatment effect. But, why the focus on a p-value? Rather they should present the interaction estimate and CI, and discuss the actual magnitude of DIFFERENCE in treatment effect for, e.g., 1-year change in age. I find this very confusing, and leads to problems with their main conclusions in the discussion and elsewhere, such as "Exploration of subgroups suggested arthroscopic surgery may lose superiority over physiotherapy in older patients".

3) The What this Study Adds box concludes: "Not all patients benefit from surgery and the decision to operate must be carefully evaluated by surgeons with specialist expertise in arthroscopic hip surgery." I think this statement arises because the authors report and discuss the amount of individuals that achieved at least a clinically significant improvement of 9. "Clinically important improvement within the individual, defined as an increase in HOS ADL greater than 9 points, was reported in 51% (95% CI 41 to 61%) of patients allocated to arthroscopic surgery and 32% (95% CI 22 to 42%) of patients allocated to physiotherapy". Although this is interesting, I do not think the authors can infer from this that some individuals are truly doing better than others.

The issue is measurement error. Individuals have random error about their final value. If they repeat again, they may get a different final value. Therefore someone just above a change of 9 may not be any different than someone below a change of 9, due to random error. A related issue is regression to the mean. Senn discusses how this problem, i.e. measurement error, may make it seem that some people respond better than others … but it is simply random error and thus erroneous personalised medicine inferences. See refs 1 2

4) The Results section should have sub-sections for summary of baseline characteristics, the main (overall) primary analyses, secondary outcomes, and then subgroup examination. This would help make the subgroup results less prominent.

5) In their response, the authors suggest that the continuous covariates were dichotomised so that subgroup results could be given on a forest plot and tests for heterogeneity made across subgroups. However, this approach to testing has

considerably lower power than calculating a p-value for the interaction term, when including a single interaction term in the model with the variable as continuous (e.g. Final score = intercept + baseline_score + age + treatment + age*treatment)

6) Even when presented, the authors still do not report the difference in subgroups (i.e. interaction estimate). E.g. "Pre-planned subgroup exploration suggested a potential differential treatment effect for the different age groups with greater superiority for arthroscopic surgery in patients under 40 years of age (effect size 13.45 (95% CI 6.45 to 20.45) for < 40 years and 1.73 (95% CI -7.44 to 10.91) for > 40 years)." – what is the actual difference and its CI? Where is the table of subgroup results and interaction estimates? Figure 4 basically shows wide CIs that overlap, and a p-value, and so is not very relevant.

7) A related point: There is an over-emphasis on 'significance', especially in the subgroup analysis. Better to just focus on the interaction estimate and its CI. E.g. the authors say "Eight-month post-randomisation secondary PROM scores including HOS Sports subscale, NAHS, OHS, iHOT, HAGOS, UCLA, PainDetect, EQ5D and HADS depression score were significantly higher in participants who received arthroscopic surgery compared with physiotherapy ($p < 0.05$) (Table 4). However, there was no significant difference in the HADS anxiety score between treatment groups ($p = 0.184$)." – clinical readers need to know the potential magnitude of the interaction effect, not whether the p-value is above or below the significance threshold.

Other points

In general, I find the discussion extremely long, with much discussion or elaboration about issues that were not identified as important or had low power to detect genuine effects or subgroups. This dilutes the paper. I suggest the BMJ clinical looks at this closely before acceptance and publication. An example is "Despite the limitation of performing multiple statistical tests, our results also suggest …"

- I read with interest the authors' explanation as to why the CI for the treatment effect is wider after imputation. I still do not follow this. I recognise that there is uncertainty in the imputed values, and this is accounted for through multiple imputations. But, after all, you are still adding in new individuals (even if their weight is lower due to the uncertainty in their values being accounted for), and so compared to a complete case analysis, there is still extra information being introduced. In a linear regression setting, I struggle to see why this leads to wider CIs. My only guess is that the residual variance in the model is estimated to be larger after imputation, which then leads to wider CI (indeed I think the SD of responses is largest after imputation). Anyway, this is a minor point, and I take the imputation results at face value, but it is rather curious.
- "Baseline demographic and clinical characteristics were well balanced across treatment groups (Table 1)." – as mentioned in my first review, this relates to the full set of patients, and actually those in the main analysis (At 8 months) were far fewer. Therefore, we also need to see the characteristics for those in the main analysis. The authors response is to include this in supplementary material S2 … but this table is only mentioned in the middle of the discussion, and not in the actual results. Can we have it in the main text instead please?
- Table 6 – what do the p-values represent? Again, we need effect estimates and CIs please. Perhaps an odds ratio or risk ratio?

In summary, I think the revision is improved, but the interpretation and presentation must be fine-tuned further (especially in regards the subgroup presentation and analysis) for the BMJ.

I hope these comments are ultimately useful for the authors, and improve the article and the BMJ readers. Best wishes, Richard Riley

Reference List

1. Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians? Stat Med 2009;28(26):3189-209.
2. Senn S. Individual response to treatment: is it a valid assumption? BMJ 2004;329(7472):966-8.

Additional Questions:
Please enter your name: Richard Riley

Job Title: Professor of Biostatistics

Institution: Keele University

Reimbursement for attending a symposium?: No

A fee for speaking?: No

A fee for organising education?: No

Funds for research?: No

Funds for a member of staff?: No

Fees for consulting?: No

Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: No

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No

If you have any competing interests <A HREF='http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/declaration-competing-interests'target='_new'> (please see BMJ policy) please declare them here: