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

26th April 2018

Dear Dr Loder,

Thank you very much for your consideration of our revised manuscript. Our study team discussed each of the comments at length. We have composed a response to each point in the text below and modified our manuscript accordingly.

As recommended, we placed particular emphasis on the statistics and revisited several aspects of our analysis in response to comments from Prof. Riley. We hope that we have addressed these comments satisfactorily, however, we would be entirely agreeable to further suggested modifications.

We also confirm that there has been no change to the author list.

Yours sincerely,

Antony Palmer MA BMBCh DPhil FRCS (Tr and Orth)

NIHR Academic Clinical Lecturer in Trauma and Orthopaedics

Report from The BMJ's manuscript committee meeting

Present: John Fletcher (chair); Richard Riley (statistician); Elizabeth Loder; Jose Merino; Wim Weber; George Roeggla; Tiago Villanueva; Daoxin Yin

*** Thank you for registering the trial prospectively. In revising the paper please make certain that all outcomes listed in the trial registry are reported in the paper, using the same terminology and in the same order as they appear in the trial registry. If any outcomes have been changed in the trial registry, please explain when and why this happened and provide results for all outcomes so that readers can judge the effect of the changes. If you present any outcomes that were not listed in the trial registry, please clearly flag them as post-hoc and please do not emphasise them in your interpretation of the trial results.**

Response 1: We ensured that our outcomes reflect the published protocol. Our protocol states that our primary outcome measure is 'change in HOS ADL' eight months post randomisation, which we assessed using outcome HOS ADL adjusted for baseline values in an ANCOVA model (please see Response 15). This methodology was agreed and signed-off in our Statistical Analysis Plan prior to unblinding the study data for analysis. Calculating 'Outcome HOS ADL' gives the same statistical significance as 'Change in HOS ADL'. (Senn S. Change from baseline and analysis of covariance revisited. Stat Med 2006).

In response to reviewer comments, we also included the proportions of patients achieving a Patient Acceptable Symptomatic State (please see Response 7) and specified that this calculation was performed in response to publications since FAIT study design, hence not included in the protocol.

Also included is our pre-specified analysis of the 'expectation HOS ADL', which was not found to influence study outcomes (please see Response 12). The proportion of patients who achieved their 'expectation' HOS ADL eight months post randomisation was 15% (95% CI 23-41%) for physiotherapy and 31% (95% CI 9-24%) for arthroscopic hip surgery.

Outcome measures for osteoarthritis (imaging and chemical biomarkers) and health economics will be reported after 36-month follow-up, as per our published protocol.

*** The lower bound of the confidence interval is compatible with a smaller difference between groups than the 9 points said to be clinically significant, so we thought the interpretation of the study should be more nuanced.**

Response 2: We have clarified that there is a statistically significant difference in treatment effect between groups, of a magnitude that suggests clinical benefit, although we acknowledge that the true difference may fall below the MCID. In response to reviewer comments, we offer a more balanced interpretation of the study results in the discussion and conclusion.

Our conclusion now reads:

'In conclusion, this study suggests that patients with symptomatic FAI referred to secondary or tertiary care achieve a greater improvement in patient reported outcomes with arthroscopic hip

surgery compared with non-operative measures. However, accurate patient selection is essential as up to a half of patients may not achieve a clinically important improvement.'

*** We were impressed with your careful attention to remote randomisation, with minimisation, after recruitment and baseline data input - this appears from table 1 to have worked well. High recruitment rate from those eligible reassures about generalisability as does the multicentre nature of the trial.**

*** Could you use the term "Arthroscopic surgery" rather than "arthroscopy"? Arthroscopy implies just a diagnostic look and see, at least in the minds of many of the editors who read the paper.**

Response 3: We have replaced the term 'arthroscopy' with 'arthroscopic surgery' throughout the manuscript.

*** Please see the separate comments of our statistician Professor Richard Riley.**

Response 4: Please see responses to Reviewer 5.

*** We do think that the comment of one of the reviewers needs far more attention than it currently receives: "The assessor blinded outcomes deliver no difference between the groups, while the self-reported (non-blinded) outcome (including the primary outcome) clearly are in favor of the arthroscopic intervention. I think this aspect absolutely needs more discussion".**

Response 5: We have now included quantitative group comparisons within the manuscript for secondary outcome measures (see tables below). Our reservation was the limited evidence that improved range of movement equates to improved clinical outcomes. During our feasibility work, we established that patient reported outcome measures (and osteoarthritis prevention) were most important to patients.

Whilst acknowledging the limitation of performing multiple statistical tests, we found a statistically significant improvement in range of hip flexion in patients receiving arthroscopic surgery compared with physiotherapy (difference in outcome adjusted for baseline 4.8 degrees (95% CI 0.5 to 9.1) p=0.03, but for no other movement. There was a corresponding statistically significant decrease in the proportion of patients experiencing pain on hip flexion in the arthroscopic surgery group compared with the physiotherapy group (p=0.012). This finding is consistent with a study of patients receiving arthroscopic hip surgery where hip flexion was the only hip movement independently associated with patient reported outcome measures (reduced hip flexion associated with worse scores) (Kemp et al. 2016).

Our results also suggest reduced pain on abduction and adduction in patients who underwent arthroscopic surgery compared with physiotherapy, perhaps representing less global irritability. We did not detect a difference between groups for the FAdIR impingement test, but fewer patients had a positive FAbER test in the arthroscopic surgery group.

The manuscript has been updated to reflect these findings and our interpretation, which is that blinded outcome measures also support the suggestion that arthroscopic surgery may be superior to physiotherapy for the treatment of symptomatic femoroacetabular impingement.

Table 6A : Range of Hip Movement at Baseline and Eight Month Post Randomisation Assessment						
	Physiotherapy Baseline	**Physiotherapy 8 Month Assessment**	**Arthroscopic Surgery Baseline**	**Arthroscopic Surgery 8 Month Assessment**	**Difference in ROM* Adjusted for Baseline**	**Statistical Significance**
Hip Flexion	95.7 (SD 19.1) Range 27, 126 n=107	99.7 (SD 17.5) Range 25, 130) n=85	96.9 (SD 15.8) Range 50, 130 n=111	105.8 (SD 16.3) Range 40, 138 n=96	4.8 (95% CI 0.5, 9.1)	p=0.030
Hip Extension	17.9 (SD 7.9) Range 5, 50 n=100	15.7 (SD 8.0) Range 0, 46) n=83	18.2 (SD 8.0) Range 0, 40 n=104	16.8 (SD 7.4) Range 0, 45) n=96	1.6 (95% CI -0.6, 3.8)	p=0.158
Hip Abduction	27.5 (SD 11.9) Range 5, 60 n=107	29.6 (SD 11.7) Range 5, 70 n=84	27.1 (SD 12.0) Range 5, 80 n=110	30.3 (SD 10.6) Range 8, 66 n=96	1.0 (95% CI -2.1, 4.1)	p=0.534
Hip Adduction	21.6 (SD 7.9) Range 5, 44 n=104	23.2 (SD 8.9) Range 5, 50 n=84	20.9 (SD 8.2) Range 5, 60 n=108	23.9 (SD 8.2) Range 9, 45 n=96	1.1 (95% CI -1.2, 3.5)	p=0.354
Hip Internal Rotation	24.0 (SD 11.2) Range 5, 55 n=107	28.9 (SD 11.2) Range 2, 55 n=84	24.9 (SD 11.2) Range 2, 56 n=110	30.8 (SD 10.6) Range 5, 69 n=96	1.4 (95% CI -1.6, 4.4)	p=0.366
Hip External Rotation	25.0 (SD 11.8) Range 5, 80 n=107	27.4 (SD 9.7) Range 8, 70 n=84	26.2 (SD 10.6) Range 7, 80 n=110	27.0 (SD 8.9) Range 10, 50 n=96	-1.1 (95% CI -3.6, 1.4)	p=0.383

*ROM = Range of Movement

Table 6B – Hip Assessment at Baseline and Eight Month Post Randomisation						
		Physiotherapy Baseline^	**Physiotherapy 8 Month Assessment^**	**Arthroscopic Surgery Baseline^**	**Arthroscopic Surgery 8 Month Assessment^**	**Statistical Significance$**
Pain on Flexion	Yes	77 (70.0%)	56 (50.9%)	80 (71.4%)	46 (41.1%)	p=0.012
	No	31 (28.2%)	29 (26.4%)	31 (27.7%)	51 (45.5%)	
	Not Available	2 (1.8%)	25 (22.7%)	1 (0.9%)	15 (13.4%)	
Pain on Extension	Yes	44 (40%)	24 (21.8%)	41 (36.6%)	18 (16.1%)	p=0.101
	No	61 (55.5%)	59 (53.6%)	67 (59.8%)	79 (70.5%)	
	Not Available	5 (4.5%)	27 (24.5%)	4 (3.6%)	15 (13.4%)	
Pain on Abduction	Yes	72 (65.5%)	48 (43.6%)	74 (66.1%)	41 (36.6%)	p=0.046
	No	36 (32.7%)	36 (32.7%)	38 (33.9%)	56 (50.0%)	
	Not Available	2 (1.8%)	26 (23.6%)	0 (0%)	15 (13.4%)	
Pain on Adduction	Yes	51 (46.4%)	39 (35.5%)	61 (54.5%)	30 (26.8%)	p=0.032
	No	55 (50.0%)	45 (40.9%)	50 (44.6%)	67 (59.8%)	
	Not Available	4 (3.6%)	26 (23.6%)	1 (0.9%)	15 (13.4%)	
Pain on Internal Rotation	Yes	78 (70.9%)	47 (42.7%)	77 (68.8%)	44 (39.3%)	p=0.155
	No	30 (27.3%)	37 (33.6%)	34 (30.4%)	53 (47.3%)	
	Not Available	2 (1.8%)	26 (23.6%)	1 (0.9%)	15 (13.4%)	
Pain on External Rotation	Yes	55 (50.0%)	33 (30.0%)	50 (44.6%)	30 (26.8%)	p=0.239
	No	53 (48.2%)	51 (46.4%)	61 (54.5%)	67 (59.8%)	
	Not Available	2 (1.8%)	26 (23.6%)	1 (0.9%)	15 (13.4%)	
FAdIR Test	Positive	95 (86.4%)	66 (60.0%)	103 (92%)	70 (62.5%)	p=0.378
	Negative	11 (10.0%)	18 (16.4%)	9 (8.0%)	26 (23.2%)	
	Not Available	4 (3.6%)	26 (23.6%)	0 (0%)	16 (14.3%)	
FAbER Test	Positive	89 (80.9%)	52 (47.3%)	91 (81.3%)	42 (37.5%)	p=0.015
	Negative	18 (16.4%)	32 (29.1%)	21 (18.8%)	54 (48.2%)	
	Not Available	3 (2.7%)	26 (23.6%)	0 (0%)	16 (14.3%)	

FAdIR Test = Pain on Flexion, Adduction, Internal Rotation. FAbER Test = Pain on Flexion, Abduction, External Rotation. ^Frequency (Percentage) $Chi Square

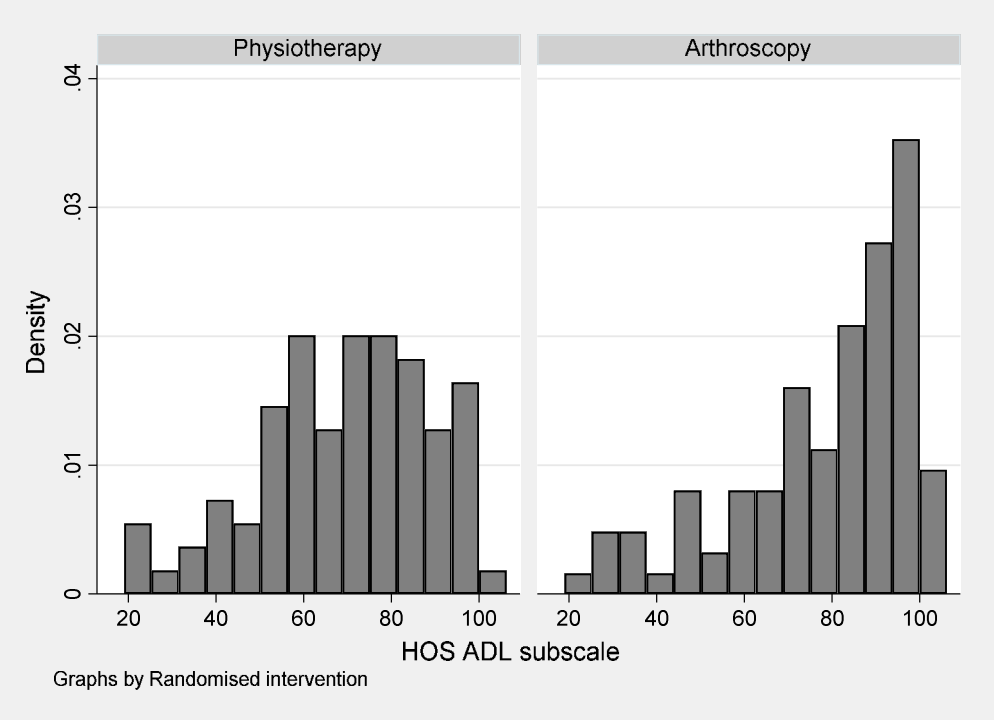
***The MICD for the HOS-ADL is 9 (Arthroscopy. 2016 Sep;32(9):1877-86). Here you found an average difference of 10. This is not a big difference; can you say something about placebo effects of this type of surgery?**

Response 6: We expanded our discussion of placebo effects to ensure our conclusions are appropriately balanced. In addition to the 'placebo effect', we have also emphasised the possible 'nocebo' effect and how these may, at least in part, explain the study findings. During study design we explored the option of a third placebo arm, however, this was not felt to be acceptable by the majority of patients and surgeons, and would have limited study recruitment. We reference the placebo surgical trials for shoulder and knee arthroscopy.

*** Another important outcome for this patient group is the PASS: Patient Acceptable Symptomatic State, and for HOS-ADL it is 87 (The American Journal of Sports Medicine Vol 43, Issue 8, pp. 1844 - 1849). It is not a primary outcome, but can you give us the proportions of patients in each group that reached this threshold?**

Response 7: We calculated the proportion of patients achieving a score above 87 on the HOS ADL eight months post randomisation as 48/100 (48%, 95% CI 38%-58%) for the arthroscopic surgery group and 17/88 (19%, 95% CI 12%-29%) for the physiotherapy group. The manuscript has been modified to include these values, that support the proposal that arthroscopic surgery may be superior to physiotherapy with respect to HOS ADL at eight months post randomisation.

	Physio	*Arthroscopy*
Exceed MCID (HOS ADL >9)	*32% (23%-42%)*	*51% (41%-61%)*
Exceed PASS (HOS ADL >87)	*19% (12%-29%)*	*48% (38%-58%)*



*** We would like more discussion of the physiotherapy intervention. Some editors thought this was a very weak comparison, and that the standard in clinical practice where they work is far more intensive and frequent physiotherapy. PT more. It's a weak comparison. The standard is 12 sessions a month and then it gets renewed. Very relevant issue. Causes a lot of pain and distress. I am mildly in favour of this.**

Response 8: There are currently no guidelines or evidence to support an optimal programme of physiotherapy for femoroactebular impingement syndrome. Our goal-based programme was developed based on the consensus opinion of the study team and published case series (no higher level evidence available).

Our study was conducted within the NHS setting in the UK, where physiotherapy provision is limited to a maximum of approximately six sessions of individual physiotherapy. A factor considered when deciding the intervention dose was to determine if a positive treatment effect could be seen at levels of intervention deliverable within current NHS resources. It was felt that increasing the duration of treatment and number of sessions markedly would reduce the generalisability of trial findings, increase NHS excess treatment costs, and is not supported by current evidence. Consultation with physiotherapy service managers also highlighted that increasing session numbers would likely reduce centre recruitment.

The overall package of treatment agreed for the study was a maximum of eight physiotherapist contacts. Exact session timings were agreed individually between participant and physiotherapist, but with the expectation that there would be more sessions towards the start of the treatment period (as would occur in routine clinical practice). Although a greater number and frequency of physiotherapy sessions may be offered in some other countries, or where funding is not from the NHS, there is no evidence to suggest these programmes are superior to our intervention. Our discussion now includes this justification.

We also provided further detail of the physiotherapy intervention, both for femoroacetabular impingement treatment and post arthroscopic surgery (please see Response 25).

*** Our patient editors say: "Thank you for the well written and thoughtful PPI declaration and dissemination plans."**

These comments are an attempt to summarise the discussions at the manuscript meeting. They are not an exact transcript.

In your response please provide, point by point, your replies to the comments made by the reviewers and the editors, explaining how you have dealt with them in the paper. If several reviewers have made the same point, you can group the comments and respond only once. If you do not agree with a reviewer suggestion, please explain why - you should not feel you have to do everything that is recommended if you have good reasons not to.

As mentioned above, the comments from Professor Riley should take precedence as he will be reviewing the revision.

Reviewer: 1

Recommendation:

Comments:

Overall, this an important study that is worldwide waited for. The study is very well performed according to all quality criteria and is written down very readable and clear. I only have a few comments.

1) **In the trial register the clinical examination tests (ROM, and impingement test) are mentioned as secondary outcomes. I therefore do not understand why the authors only reported them descriptively. I recommend to assess and report the between group difference for this and its 95% CI.**

Response 9: Please see Response 5. We have now performed a statistical assessment of these measures.

2) **The assessor blinded outcomes deliver no difference between the groups, while the self-reported (non-blinded) outcome (including the primary outcome) clearly are in favour of the arthroscopic intervention. I think this aspect absolutely needs more discussion.**

Response 10: Please see response 5. There are statistically significant differences between treatment groups with respect to blinded clinical assessments, although we have interpreted the clinical importance with caution.

3) **Continuing, we have banned arthroscopic interventions in knee OA patients based on the fact that the effect of the real arthroscopic intervention did not surpass the effect of the placebo procedure. The placebo effect of such surgical interventions seems to be huge; at least higher than that of exercise. The same huge placebo effect most probably is true in the present arthroscopic intervention and could explain the difference in self-reported outcomes found between physical therapy and arthroscopic intervention. Like comment 2 this aspect needs more discussion.**

Response 11: We have now expanded our discussion of the placebo effect, which is likely to be present in this study (also see Response 6).

4) **For the same reason I also would like the authors to report on the patients' expectation of the intervention was measured beforehand (if measured), and what the impact of this could be on the outcomes, as well as what proportion of the patients already had received physical one year or more before the inclusion in the trial.**

Response 12: At baseline, participants were asked to complete an 'expectation' HOS ADL, which is available for 184 of 188 included in the primary analysis. When included as a continuous covariate in the primary analysis model, it was not statistically significant and changed the treatment effect very slightly from 10.0 to 9.9. We therefore concluded that adjusting for expectations does not affect the trial results.

No patient completed a programme of physiotherapy targeting femoroacetabular impingement syndrome in the year prior to recruitment into the study. A significant proportion of patients reported receiving some form of physiotherapy within the year preceding referral to a specialist clinic (59%). There was no difference in proportion between the arthroscopic surgery (58%) and physiotherapy groups (60%). This observation reflects current practice, where a significant proportion of patients receive physiotherapy in the community prior to referral to secondary or tertiary care.

5) <u>The authors talk about "to seed the minimisation system" (page 9, line 7); I recommend to use other words so that clinical readers can understand what is meant with this. The same is true for minimization factors (page 12, line7).</u>

Response 13: Although we believe our previous text: "Randomisation for the first 12 participants (10% of original sample size) was based on a simple random list to seed the minimisation system" was a more accurate description of our methodology, we have modified our text to:

Randomisation was performed by a research nurse at each site using an automated computer-generated telephone randomisation system provided by OCTRU (Oxford Clinical Trials Research Unit). Randomisation for the first 12 participants (10% of original sample size) was based on a simple random list and subsequent participants were randomised using a minimisation algorithm. This algorithm included a random element (80%) and aimed to generate balanced treatment allocations by age (< 40 or ≥ 40 years), gender, baseline Activities of Daily Living subscale of the Hip Outcome Score (HOS ADL) (<65% or ≥65%), and study site[12].

6) <u>Apparently, the physical therapy in the UK does not exceed 8 sessions. To what extent were the goals achieved in the physical therapy group at the end of the sessions? Could it be that the physical therapy should be more extensive?</u>

Response 14: Patient goals prior to commencing physiotherapy were extremely varied and ranged from 'sleeping through the night' to 'rowing at national level' or performing martial arts 'roundhouse kicks'. The most common theme was return to sport. When patients were asked whether they had achieved their goals on completing physiotherapy, 31/88 (35%) responded 'yes', 35/88 (40%) responded 'no', and 22 participants did not answer this question (mITT population). The 35% completing their goals is comparable to the 32% with an improvement in HOS ADL that exceeded the minimal clinically important difference. At eight-month follow-up, two patients crossed over to receive arthroscopic surgery. It is possible that additional physiotherapy may have provided improved outcomes, however, this is not currently funded on the NHS (please also see Response 8). We have expanded our discussion accordingly.

<u>Additional Questions:</u>
<u>Please enter your name: Sita Bierma-Zeinstra</u>

<u>Job Title: Professor</u>

<u>Institution: Erasmus MC - University Medical Center Rotterdam</u>

Reviewer: 2

Comments:

I appreciate the opportunity to review this manuscript.

This is a pragmatic multicenter two-armed randomized control trial comparing physiotherapy to hip arthroscopic surgery for the treatment of femoroacetabular impingement (FAI). There is a notable paucity of high-quality clinical trials comparing the outcomes between these two approaches. Several similar RCTs (e.g., the UK and Austrialian FASHIoN trials) are ongoing, but this study appears to be the first complete report. The analysis showed strong evidence of better outcomes (primary outcome: HOS ADL) achieved by hip arthroscopy, compared to physiotherapy, at eight months post-randomization. The study appears to be well designed and implemented, with a good adherence to CONSORT guidelines. The results provide very important knowledge to enhance the current understanding of the risks and benefits of these two mainstay approaches. Below are a few comments:

1. Discrepancy between the current study and the published protocol.

The primary outcome of this study was the HOS ADL measured eight months post-randomization. In the published protocol (Palmer et. al, 2014, Bone and Joint Research, 3-11), the primary outcome was the "change in the HOS ADL".

Response 15: Our protocol states that our primary outcome measure is 'change in HOS ADL' eight months post randomisation, which we assessed using outcome HOS ADL adjusted for baseline values in an ANCOVA model. This methodology was agreed and signed-off in our Statistical Analysis Plan prior to unblinding the study data for analysis. Calculating 'Outcome HOS ADL' gives the same statistical significance as 'Change in HOS ADL' (Senn S. Change from baseline and analysis of covariance revisited. Stat Med 2006).

We report change in HOS ADL as the proportion of patients with an increase in HOS ADL:

	Physio	*Arthroscopy*
At Least One Point Improvement	*50% (39-61%)*	*70% (61-79%)*
At Least Nine Point Improvement (MCID)	*32% (22-42%)*	*51% (41-61%)*

2. Radiographic inclusion criteria not clear

Inclusion was defined as "… have clinical and radiographic evidence of FAI" and "surgeons made a qualitative assessment of hip morphology to diagnose FAI". The qualitative assessment in the current study appears subjective. The consistency between centres in patient recruiting needs to be established. Although the authors argued that there is no consensus on a quantitative imaging measurement as inclusion criterion, more details should be described on how the qualitative assessment was performed. Currently it is unclear whether there is an effort to standardize this assessment among participating surgeons (If this is also part of the pre-trial meeting, please mention).

Response 16: Inclusion criteria were discussed extensively amongst participating surgeons, who all have a specialist interest in young adult hip surgery, prior to commencing the study. None of the participating surgeons routinely perform quantitative measurements to diagnose femoroacetabular impingement morphology. Given the pragmatic nature of the study, we did not wish to enforce diagnostic thresholds to ensure that the study reflected current clinical practice and to improve generalisability of study findings. The diagnosis of FAI syndrome was based on patient history, clinical examination, and imaging studies. We

accept that there is subjectivity in qualitative assessment and we therefore compared quantitative measures of cam and pincer morphology between centres and found these to be comparable and consistent between sites (see table below). Supporting Analysis B included imaging measurements as co-variables.

	Cam morphology: MRI radial slices				Pincer morphology: Anteroposterior Radiograph
	Baseline average alpha angle (bone)	Baseline max alpha angle (bone)	Baseline average alpha angle (cartilage)	Baseline max alpha angle (cartilage)	Lateral centre edge angle
Centre 1					
Participants	148	148	148	148	148
Mean	67.2	86.1	67.5	86.2	28.7
Standard Deviation	12.7	17.8	11.6	15.9	6.8
Centre 2					
Participants	30	30	30	30	32
Mean	67.1	86.1	67.7	86.5	28.7
Standard Deviation	10.2	12.0	8.7	12.8	6.6
Centre 3					
Participants	6	6	6	6	6
Mean	67.9	88.3	68.3	87.4	32.4
Standard Deviation	11.9	19.6	11.9	19.4	8.4
Centre 4					
Participants	Not possible to analyse MRIs				18
Mean					28.4
Standard Deviation					6.9
Centre 5					
Participants	5	5	5	5	7
Mean	62.5	86.0	56.8	74.0	29.0
Standard Deviation	8.4	19.8	2.6	8.1	5.1

3. "Aim of this study" vs. "aim of this RCT"

In Introduction, the authors stated: "The aim of this study is to … and preventing the development or progression of osteoarthritis in patients …". However, there is no data on the prevention of OA development at this stage. To avoid confusion, please say: "the aim of FAIT is to …" or "the aim of this trial is to…", or something alike.

Response 17: We have modified this paragraph to:

The aim of the Femoroacetabular Impingement Trial (FAIT) is to compare arthroscopic hip surgery with physiotherapy and activity modification in patients referred to secondary or tertiary care with symptomatic FAI. In this manuscript, we report the primary end-point of patient-reported outcomes at eight months post-randomisation. Osteoarthritis progression will be evaluated at three-year follow-up. Study design was based on a prior feasibility study that demonstrated surgeon and patient equipoise.

4. PROMs of the late treatment group

I am not sure if understand this correctly - It appears that "six months after intervention" was used interchangeably with "eight months post randomisation", e.g., in "what this study adds". But the primary outcome was set at eight months post randomisation. For patients who received treatment less than 12 weeks, these two time points were considered equal. But for patients who commenced their treatment more than 12 weeks post randomisation (page 12), "PROMs were collected at eight months post randomisation (primary outcome measure) and six months post intervention in this group." - isn't it better to use PROMs at six month post intervention as the primary outcome for analysis? The delay in treatment after randomisation will affect the time from the start of intervention to PROM collection and therefore make this time from intervention to PROM more variable.

Response 18: We have modified 'What this study adds' to reflect the primary outcome measure, which is HOS ADL measured eight months post randomisation. During study design there was extensive discussion as to whether the primary outcome measure should be post-randomisation or post-intervention. The Surgical Intervention Trials Unit (SITU) frequently debate this issue.

Clinically, assessment post-intervention may seem preferable, particularly when one intervention has longer waiting times. However, this approach limits the ability to infer a causal relationship between intervention and outcome due to systematically different times to intervention after making the decision to treat. Whereas groups are balanced at the time of randomisation, this may not be true at any other timepoint. We therefore did not feel that post-intervention analysis should be used as the primary analysis.

We also performed a post-invention analysis (Supporting Analysis D in Table 3), and the results are consistent with the post-randomisation analysis. The treatment effect is a little higher when measured post-intervention, potentially due to longer recovery times from surgery for patients receiving arthroscopic surgery compared with post-randomisation assessment.

Additional Questions:

Please enter your name: Yun Peng

Job Title: Research Staff

Institution: Massachusetts General Hospital

Reviewer: 3

Recommendation:

Comments:

Originality

This study adds RCT data to support the decision between conservative physiotherapy and activity modification or arthroscopic treatment of femoroacetabular impingement. With physiotherapy revealing subjectively inferior outcomes in comparison to arthroscopy, currently holding a low priority in the NHS, these results adds relevant knowledge for general practitioners and policymakers.

Importance

Although arthroscopy presented beneficial and clinically significant outcomes in comparison to physiotherapy, the pragmatic approach, in this case, leaves some questions unanswered in order to fully answer the stated research questions. With 51% of included patients in arthroscopy group and 32% in the physiotherapy group achieving important symptomatic improvements, success-rates for both treatment groups are relatively low. Consequently, inclusion of planned long-term data regarding osteoarthritis and hip replacement surgery risk could potentially have contributed considerably to the decision towards arthroscopy if able to serve as tertiary prevention as hypothesized. Although adding data to the considerations of the general practitioner when referring patients and policymakers when prioritizing health care services, I am uncertain whether the condition and clinical implications are frequent and significant enough for a general journal.

Response 19: In our feasibility study, patients placed equal importance on improving symptoms and preventing future osteoarthritis, hence the co-primary outcomes of this study.

Long-term follow-up is ongoing to determine whether arthroscopic hip surgery has disease modifying potential, and FAI is thought to be responsible for up to 50% of all hip osteoarthritis.

At present, there is little evidence to inform current practice for symptom management. Femoroacetabular impingement syndrome is an increasingly recognised condition seen frequently by primary care and musculoskeletal practitioners who do not have a specialist interest in this field, hence the relevance to a general journal. In a study of 19,185 patients aged 15-60 years in the Netherlands, femoroacetabular impingement syndrome was confirmed in 17% and suspected in 30% of patients presenting with groin pain (Rolling et al. 2016). The number arthroscopic hip procedures performed continues to rise worldwide and the number performed annually in England between 2002 and 2013 increased by 727% (Palmer et al. 2016).

Scientific reliability

- The introduction is well-written, well composed, and the research questions concisely stated. The feasibility study and thoroughly considered protocol are strengths of this study. It is unclear what is meant by "equipoise between surgeon and patient" in page 8, line 9.

Response 20: We have rephrased this statement to "surgeons and patients both have equipoise with respect to physiotherapy and activity modification versus arthroscopic surgery".

- The methods section describes appropriate and convincing efforts to minimize bias with regards to power calculations, intention-to-treat, and robust questionnaire information. However, the pragmatic and feasible approach to the physiotherapy intervention (non-specialized physiotherapist and eight sessions of physiotherapy over a period of five months) pose a problem when compared to a close to optimal arthroscopy intervention (performed by specialists and combined with exercises). More elaborate specifications of interventions could potentially alleviate this concern but is not reported. An intervention group receiving both arthroscopy and subsequent physiotherapy could possibly be interesting for future studies in order to further improve patient outcome from current treatments. Based on baseline characteristics of included participants, validity for mixed- and pincer-type hip impingement is questionable as the isolated Cam-type make up for 94% of the distribution of types. Exclusion criteria and baseline characteristics are clearly and adequately described and no checklist, ethical, or reporting issues were identified. The method section is very elaborate but would benefit from being more focused.

Response 21: We studied the methods section to seek ways of making it more focused. Whilst we have not identified information we wish to exclude, we would be agreeable to moving sections to supplementary data.

- Results were credibly reported. The extensive tables would benefit from containing lesser amounts of information, perhaps creating an online supplemental file for full table information and exclusively presenting essential and particularly interesting results in the main manuscript. The elaborate questionnaire data, radiography-measures and supporting analyses structure are strengths of this study but is not particularly clearly presented.

Response 22: All tables have been modified in an attempt to aid clarity.

- Discussion and drawn conclusions are appropriate, but should be more explanatory of field-specific measures when targeting a general journal. As physiotherapy is still recommended as first-line treatment as a consequence of potential complications of surgery, a description of suggested place and/or conditions of arthroscopy is warranted to state the precise clinical implications of this study. A bullet point table or written summary, possibly in a Conclusion section, would enable readers to get a quick resume and authors to emphasize key messages. Abstract is well-written but needs to include the recommendation regarding preservation of physiotherapy as first-line treatment (on the basis of potential arthroscopy complications) to be representative of the discussion.

Response 23: We hope the 'What This Paper Adds' section provides a valuable bullet point synopsis of the key findings. We have provided more detail on the suggested provision of arthroscopic hip surgery in the discussion.

We are reluctant to state in the abstract that we recommend physiotherapy should remain first-line treatment, as the statement is not directly supported by study data, but is our suggestion given the proportion of patients with a clinically important symptom improvement in each treatment group.

- No problems with references identified.

Additional Questions:
Please enter your name: Jeppe Bo Lauersen

Job Title: Medical doctor

Institution: Department of Orthopedic Surgery, Amager-Hvidovre Hospital, Hvidovre, Denmark

Reviewer: 4

Recommendation:

Comments:

Thanks for inviting us to review this interesting study on the effect of hip arthroscopy versus physiotherapy. This review is performed by Gerjon Hannink, medical statistician and myself, Wim Schreurs. To perform a thorough review in a study with a set up like this I can not do this without experienced statistical advice.

Overall comments

This an impressive study by an extensive team of experts trying to study the effect of hip arthroscopy versus physical therapy in patients with femoro-acetabular impingement. There is a high demand for studies like these, given the enormous increase in hip arthroscopies worldwide for this indication, without a sound evidence of the effect of these arthroscopies.

They set up an RCT in 7 high volume arthroscopic centres, and surgeries are done by experts on hip arthroscopy. The control patients were referred for physical therapy, they had maximal 8 treatment of therapy, most had less, and were treated by physical therapist who had instructions on the training program.

There are many outcome parameters used and I was impressed by the number of tests used. This is the first study based on this extensive study set up.

The conclusion at 8 months after randomisation, in practise around 6 months after the surgery, must be somewhat disappointing. They claim that after this period the mean HOS ADL was 10 points higher than in the physiotherapy group, however this difference is clinically not very impressive. It is not expected that this benefit will increase after a longer follow up, the authors state in the paper that most gain will be obtained in the first 6 months after the surgery. It is worrisome that only half of the patients experiences a clinical benefit after this procedure, so that also means that 49 % of the patients will have a quite costly surgery without benefit.

Response 24: The 10 point difference between groups exceeds the MCID, suggesting a clinically important difference between treatments. However, only 51% patients receiving arthroscopic surgery exceeded this MCID (assuming the MCID of 9 points is valid for our cohort). Our interpretation is that it is future work should focus on identifying individuals most likely to benefit from surgery (and physiotherapy). Cost analysis will be performed after three-year follow-up, as per protocol, which will take into account long-term outcomes and cross-over between treatment groups.

In their conclusions the authors state that arthroscopy outcomes for this indication is superior to non-operative measures, and hence this study supports the provision of arthroscopy. We are not sure that the authors can claim this. First, this is arthroscopy performed by excellent surgeons versus physical therapy. The level of surgery will be high and even in the excellent surgical hands the clinical difference is not impressive 6 months after surgery. And only 50% of the patients has a minimal clinical benefit. Of course, future better inclusion criteria can probably improve the outcome, however this is the state of the arthroscopy in the hands of experienced surgeons at the moment.

Response 25: We believe the surgery delivered in this study represents what patients can expect to receive within the UK NHS. Centres in this study include five District General Hospitals and two University Teaching

Hospitals. In these centres, all surgeons who perform hip arthroscopy participated in the study. Our selection criteria of high volume surgeons ensured that individuals performing the procedure were beyond their learning curve, which is steep for this procedure. In the NHS, hip arthroscopy is only performed by surgeons with a specialist interest in young adult hip as per NICE guidance. Outcomes from arthroplasty surgery have been shown to be inferior for low volume surgeons (Ravi et al. 2014 BMJ) and a suggested minimum annual procedure number is emphasised in the Getting It Right First Time (GIRFT) initiative. The same principle is likely to apply for hip arthroscopy and the manuscript discussion now addresses these issues.

There is also a severe potential bias, there is very little information on the after treatment after the arthroscopy. They state that patients after arthroscopy got also physical therapy, but what was the protocol, how many treatments did they get and who trained them. This is compared to physical therapy in the conservative treated group who had a very low number of physical treatment session during their treatment, with a maximum of 8 session in 8 months. So, what is the effect of physiotherapy after the scopy on the outcome?

Response 26: The focus of physiotherapy post arthroscopic surgery was to maintain range of movement and guide return to activity, whereas the focus of physiotherapy treatment for femoroacetabular impingement (intervention in physiotherapy group) was to improve pain and function, and address patient-specific goals.

The post-operative physiotherapy protocol for this study represented routine care in this pragmatic study. Study participants were assessed by the physiotherapy team prior to discharge from hospital to ensure patients could mobilise safely and a referral was made to physiotherapists local to the patient (not members of the study team). Patients were asked to commence active range of movement and isometric exercises on the day following surgery (demonstrated by a physiotherapist prior to discharge and illustrated in patient information sheets). It was recommended that stretching exercises and static bicycle work without resistance commenced within the first postoperative week. Strengthening and functional exercise was advised three weeks after surgery, usually under outpatient physiotherapy guidance. At this point low impact exercise was permitted, such as swimming or the use of a cross-trainer machine, whilst running and deep squats were to be avoided for six weeks. During weeks 6-12 postoperatively, the emphasis moved to improving muscle endurance, with sport specific rehabilitation beyond 12 weeks when appropriate.

With respect to the physiotherapy treatment of femoroacetabular impingement syndrome, NHS commissioning funds a programme of up to eight physiotherapy sessions (please also see Response 8), which were delivered over a five month period reflecting current practice. The initial physiotherapy appointment consisted of a 60 minute assessment with the development of a tailored treatment plan based on symptoms, clinical examination, and patient goals. Treatment was delivered via a home exercise programme with up to seven 30 minute follow-up appointments. There were broadly six elements to the physiotherapy programme:

1) *Activity Modification: Cessation or adaptation of activities and postures that cause pain, and avoidance of painful terminal range of movement (particularly flexion and internal rotation).*
2) *Movement Modification: Establish limits of pain-free active movement and ensure exercises do not exceed these limits. Passive stretch of iliopsoas, external rotators, gluteal muscles and lumbar extensors.*
3) *Muscle Strength: Targeted strengthening exercises with emphasis on the external rotators and abductors for movement control and avoidance of the impingement position.*
4) *Segmental Stabilisation: Progressive strength training for core muscles and lumbo-pelvic stabilisation with further improvement in movement control.*
5) *Functional Activity: As muscle strength and global movement control improves, increased intensity and complexity of dynamic exercises based on individual activity goals. Includes sensory motor training.*

6) *Return to Activity: Phased return to sporting activities as per patient goals with modified movement and posture where possible. Strengthen hip stabilisers during loading and address any residual imbalances.*

Patients progressed through their physiotherapy programme at different rates and were discharged once goals were achieved or on completion of the physiotherapy programme, with advice to continue their physiotherapy exercises. If at the eight-month post randomisation assessment patients remained symptomatic and wished to consider arthroscopic surgery, they were reviewed by the treating consultant. After discussion the potential risks and benefits of arthroscopic surgery, if patients wished to proceed they were listed for surgery in the absence of contraindications.

The median number of physiotherapy sessions in the post arthroscopy group was 4 (IQR 2.5 to 6) and for the physiotherapy treatment group was 6 (IQR 4 to 8).

Although not part of this session, there is also a considerable cost aspect. As suggested by the authors in their paper, later on they will focus on costs. However, one cannot start promoting arthroscopy for an indication for which arthroscopy is widely used (FAI) and later come back on the costs.

Response 27: We have now included the limitation that the results of the cost analysis must be awaited to fully justify recommending arthroscopic surgery. We will perform cost analysis on trial completion at three-year follow-up (as per protocol) to account for long-term outcomes and further treatment.

It has to be acknowledged that in many patients the effect of this surgical treatment is limited, it is completely unclear if this surgery will really be effective in the longer follow up given the early results and it is at this stage also unclear if an arthroscopic early treatment will prevent the big problem on the long term, hip osteoarthrosis.

Response 28: We have now ensured that the conclusions of this study sufficiently reflect the fact that short-term benefits of arthroscopic surgery may not be maintained in the long-term, and that there is currently no convincing evidence to suggest disease-modifying potential.

FAI is a pathology seen in many individuals in the population, about 20 %. Of these 20 % also around 20 % is symptomatic at a relative young age. So the number of patients with symptomatic FAI who can be considered as a candidate for arthroscopy is high. For communities, they cannot afford an enormous increase in arthroscopies without a sound scientific and clinical back and do we really prevent later problems like osteoarthritis.

Response 29: In our discussion, we suggest that based on current evidence, physiotherapy remains first line treatment. However, it is possible that arthroscopic surgery is more cost effective than physiotherapy or no treatment, as suggested by a recent health economic study from USA (Mather et al. 2018). Femoroacetabular impingement syndrome produces a significant socioeconomic burden and the authors of this study concluded that arthroscopic surgery conferred a mean cumulative 10-year societal savings of $67,418 per patient versus non-operative treatment. Our discussion emphasises the important of our future health economics analysis.

We certainly would recommend this impressive study for publication in BMJ, this is the first study based on this trial and hopefully many new insights will be presented in future studies. However, for now, the interpretation of the data and the conclusion should be more balanced.

Response 30: We have ensured that our discussion offers a more balanced interpretation of the results, with less certainty over the clinical benefit of arthroscopic surgery over physiotherapy.

Comments Gerjon Hannink

Well conducted and reported study given the challenges associated with surgical trials.

Throughout the manuscript, the terminology '8 months post-randomization' and '6 months post-operative' is somewhat confusing.

Response 31: We have modified the manuscript to ensure we refer to eight months post randomisation (primary outcome measure) except when we are specifically addressing Supporting Analysis D (post intervention analysis).

How generalisable are the results since the surgery was performed in/by experienced high volume surgeons/centers? This could have overestimated the effect of surgery.

Response 32: Please see Response 25.

What are the authors comparing? Please be more precise / specific.

In the abstract it is stated that hip arthroscopy (followed by routine post-operative care) was compared with goal based physiotherapy. In the methods section it reads that arthroscopy patients received post-operative physiotherapy (routine NHS care). In what was this post-operative physiotherapy different from the goal based physiotherapy in the control group? This could be clarified.

Response 33: Please see Response 8 and 26.

Analysis was performed as per modified intention (mITT) to treat including those patients with available outcomes data based on their randomized treatment allocation (regardless of compliance). ITT analysis reflects the practical clinical scenario as it admits noncompliance and protocol deviations. The mITT analysis allows a subjective approach in entry criteria, which may lead to confusion, inaccurate results and bias (Gupta, Perspect Clin Res. 2011). I am slightly worried that the mITT might have influenced the results. If noncompliant subjects and dropouts (no outcome available) are excluded from the final analysis, it might create important prognostic differences among treatment groups. Moreover, subjects may be noncompliant or may drop out from the study due to their response of treatment. This should be discussed.

Response 34: We acknowledge that our primary analysis based on the mITT population may open our trial results to bias due to missing data, and we have now emphasised this in the discussion. However, our primary analysis adjusts for the most important prognostic factors, and the treatment effect is preserved for all supporting analyses including multiple imputation (Table 3). Multiple imputation is often considered to be the gold standard for handling missing data as it allows the inclusion of all randomised participants in the analysis and makes assumptions about the underlying missing data mechanism (assuming data are missing at random) to accurately estimate real values. However, a recent study did not find any advantages of multiple imputation over mITT analysis with appropriate covariate adjustment in a range of comparable scenarios (Sullivan et al. Should multiple imputation be the method of choice for handling missing data in randomised trials? Stat Methods Med Res 2016). Since the assumptions about data being missing at random are untestable, we also performed missing not at random sensitivity analyses (presented in the

supplementary material). The trial conclusions based on the mITT population, the ITT population based on multiple imputation, and various missing not at random analyses are all consistent in the trial conclusions reached. We therefore believe our conclusions are robust to missing data in this trial.

The conclusion (and discussion) should be more balanced. I find it misleading shorthand to conclude that hip arthroscopy is superior to physiotherapy based on an MCID. Although the MCID is a useful tool to define general guidelines to determine whether a treatment produces clinically meaningful effects, the MCID value is impacted by the method used to calculate it (anchor, distribution), by the type of anchor chosen and by the definition (threshold) of improvement. The MCID is also dependent on the population characteristics such as disease type and severity, sex, age, etc.

The authors showed a statistically significant difference of 10 points/percent in HOSADL score between surgery and physiotherapy. This difference exceeds the MCID of 9 points reported by Martin & Phillipon (Arthroscopy 2008) and was therefore concluded to be clinically relevant. However, comparing the finding with the reported 6 months MCID in HOSADL of 15 points by Chahal et al. (Orthop J Sports Med 2014) would make the difference less relevant...

This said, most importantly, for appropriate use, the MCID should be applied to changes in individual subjects, not to group changes (Katz, J Orthop Surg Res. 2015).

Response 35: We ensured our conclusions are more balanced and we have included the limitations of MCID in our discussion. We also revisited the literature. As the reviewer highlights, Chahal et al. published an abstract in 2014 describing outcomes from 130 patients who underwent arthroscopic surgery for femoroacetabular impingement. They reported a HOS ADL MCID of 15.0 alongside scores for Patient Acceptable Symptomatic State (PASS). The authors later publish the outcomes from this cohort as a full article (Chahal et al 2015. American Journal of Sports Medicine). In this manuscript they report the PASS data, but not their MCID calculations, and instead cite HOS ADL MCIDs of 9 (Martin et al. 2008) and 5 (Kemp et al. 2013). In a subsequent publication (Levy et al. 2016), the authors again cite these MCIDs rather than the 15 points stated in their previously published abstract. Nevertheless, we accept that the MCID may be greater than 9 in our specific cohort and have included this limitation in the manuscript, although it could equally be argued that the MCID may be less than 9. We also acknowledge that MCID should be applied to individuals, rather than groups, and our updated manuscript includes the proportion of patients who achieved the PASS in addition to the already reported MCID data (please also see Response 7):

	Physio	*Arthroscopy*
Exceed MCID (HOS ADL +9)	*32% (22-42%)*	*51% (41-61%)*
Exceed PASS (HOS ADL >87)	*19% (11%-28%)*	*48% (38%-58%)*

Additional Questions:

Please enter your name: Wim Schreurs and Gerjon Hannink

Job Title: Orthopedic surgeon and Medical statistician

Institution: Radboudumc Nijmegen The Netherlands

President of the European Hip Society 2016-2018
(Schreurs)

Reviewer: 5

Recommendation:

Comments:

Thank you for the opportunity to review this interesting trial. I have reviewed from a statistics viewpoint, and many analyses are well done (e.g. adjusting for baseline values and site, in a mixed effects model). However, I have some comments to be addressed before the BMJ can make a proper evaluation of the paper.

1) **The drop out in the groups is a concern, but the authors do investigate this through imputation and missing not at random analyses. Also, drop out is perhaps expected in these field. Therefore, I do not think this is a fatal flaw, but it should be discussed in more detail as a limitation in the Discussion.**

Response 36: We have now included patient drop out in our discussion of study limitations, including the small difference in baseline scores between patients included and excluded from the mITT (Table S2) (please also see Response 37). Despite drop outs, our sensitivity analysis supports the trial results and the study remains adequately powered.

2) **Baseline characteristics for each group should also be given for the subset of patients who actually gave follow up results toward the final analysis. Indeed, this is more important, as the large drop out of patients may well have caused baseline imbalances, and we can't see this at the moment.**

Response 37: We have now included this table from our Statistical Report as Supplementary Data in the revised manuscript (Table S2). It does show a slight baseline imbalance within the physiotherapy group, with lower baseline scores in patients excluded from the mITT population (primary outcome). However, our results appear robust to different assumptions regarding missing data (missing at random and missing not at random) in our sensitivity analysis (Figure S3).

	Physiotherapy Participants in mITT population	**Physiotherapy** Participants excluded from mITT population	**Arthroscopic Surgery** Participants in mITT population	**Arthroscopic Surgery** Participants excluded from mITT population	**Total Participants** mITT population	**Total Participants** excluded from mITT population
Participants	88	22	100	12	188	34
Male	30 (34%)	7 (32%)	32 (32%)	6 (50%)	62 (33%)	13 (38%)
Age (years)	36.4 (10.3) [18, 60], n=88	34.2 (7.7) [21, 48], n=22	36.2 (9.8) [18, 59], n=100	38.2 (8.6) [24, 49], n=12	36.3 (10.0) [18, 60], n=188	35.6 (8.1) [21, 49], n=34
BMI	26.8 (5.0) [18, 41], n=87	25.5 (3.9) [18, 31], n=19	25.8 (4.7) [17, 42], n=97	26.7 (6.0) [20, 39], n=12	26.3 (4.9) [17, 42], n=184	26.0 (4.8) [18, 39], n=31
HOS ADL	67.5 (18.9) [12, 99], n=88	58.7 (17.5) [28, 93], n=22	66.2 (18.7) [28, 99], n=100	65.5 (18.2) [37, 97], n=12	66.8 (18.8) [12, 99], n=188	61.1 (17.8) [28, 97], n=34
HOS Sports	49.7 (22.7) [0, 94], n=88	39.0 (21.4) [6, 75], n=22	48.5 (22.6) [0, 100], n=99	52.1 (29.3) [0, 100], n=12	49.1 (22.6) [0, 100], n=187	43.6 (24.9) [0, 100], n=34
OHS	29.7 (9.6) [5, 47], n=87	23.1 (11.1) [8, 40], n=22	28.6 (9.5) [8, 45], n=100	27.0 (6.0) [16, 38], n=12	29.1 (9.6) [5, 47], n=187	24.5 (9.7) [8, 40], n=34
iHOT	3.8 (2.2) [0, 9], n=88	2.6 (1.8) [0, 6], n=22	3.5 (2.2) [0, 8], n=100	3.7 (2.1) [1, 8], n=12	3.6 (2.2) [0, 9], n=188	3.0 (1.9) [0, 8], n=34

3) **This is a major point:**

The subgroup analyses are rather sub-standard. The authors dichotomise variables at arbitrary values, such as age at 40. Also, there is no estimate of the interaction (difference in effect between subgroups). Therefore, the authors state that there is a subgroup effect for age, but do not actually quantify this properly. The authors therefore need to address these things, by analysing covariates on their continuous scale (or at least use 4-5 categories), and the potential for non-linear trends / interactions, and by quantifying interactions (differences between subgroups).

Response 38: We acknowledge the limitations of our sub-group analysis, particularly in relation to the inadequate power with our sample size. However, our aim was to explore consistency of treatment effects across different patient groups, and for hypothesis generation rather than definitive conclusions. We dichotomised continuous variables to reflect our minimisation algorithm and enable us to present the subgroups as a forest plot (pre-specified in statistical analysis plan). The primary outcome measure for the study was adjusted for age and baseline HOS ADL as continuous variables.

We repeated our sub-group exploration using linear regression with baseline HOS ADL and age as continuous variables. There was a weak but statistically significant association between age (linear continuous variable) and eight months post randomisation HOS ADL in our primary analysis (adjusting for continuous baseline HOS ADL, gender, and allowing for clustering by site). When adding an interaction term between age (continuous) and treatment, this interaction was statistically significant (p=0.001), suggesting the observed treatment effect for arthroscopic surgery over physiotherapy may decrease with increasing age. 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. However, we would be agreeable to including this data at your request.

There was no statistically significant relationship when adding an interaction term between HOS ADL (continuous) and treatment (p=0.084).

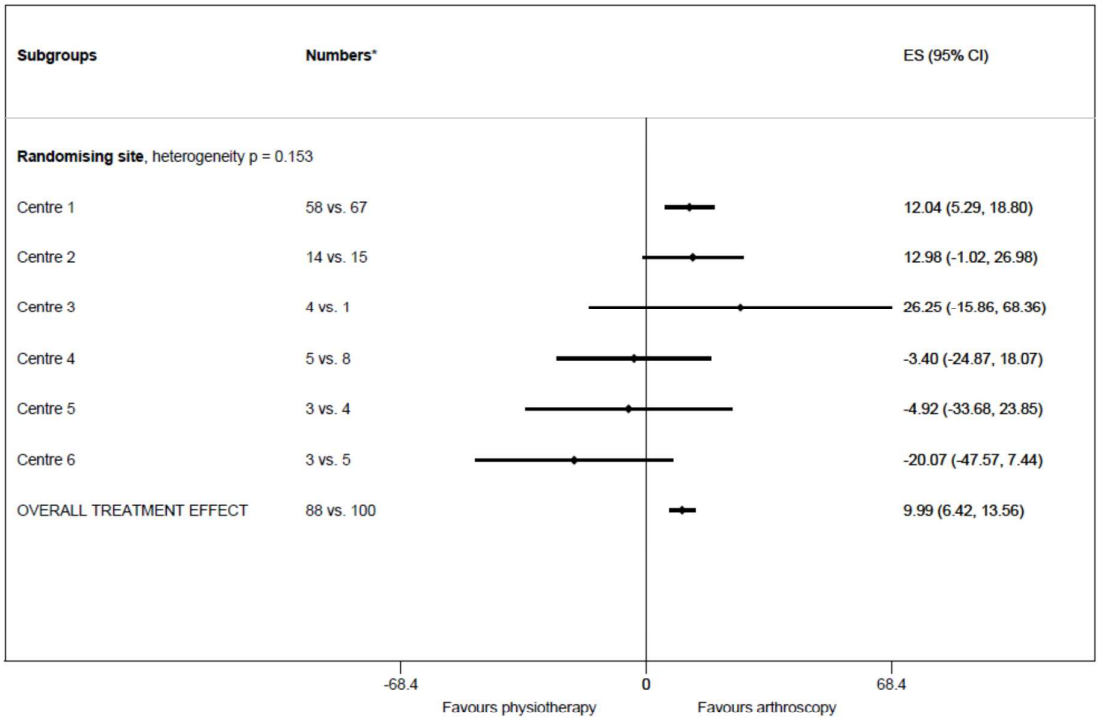
4) **Also, there are a number of paragraphs hyping the subgroup effects (despite the note from the authors that the study was not powered for this), and this should be removed or toned down.**

Here is an example of discussion that adds nothing new and is based on data dredging and speculation: "Patients with more severe symptoms pre-operatively have been shown to have a greater likelihood of achieving a clinically important improvement in HOS ADL after hip arthroscopy33,35. Using a threshold score of 65 on the HOS ADL, we did not detect a difference in treatment effect, although this may reflect the threshold selected or inadequate power."

Response 39: We have removed this sentence and rewritten the discussion of our exploration of subgroups.

5) **There are multiple centres. Of interest, therefore, is whether there is heterogeneity in the treatment effect across centres? Was this accounted for in the analysis? Can the authors provide a forest plot of the centre-specific results? It was add credence to the overall results, if the direction of effect is similar in all the centres, with little heterogeneity.**

Response 40: Our statistical analysis accounted for clustering within centres using the 'cluster' option in Stata's regress command to ensure that cluster-robust standard errors were used. A forest plot of the centre-specific is included below. On this plot, the treatment effect for sites recruiting more than 20 patients are very similar (Centre 1 and 2), and in favour of arthroscopy. There is more variation in the treatment effect at sites with small numbers of participants, which also had higher dropout rates. We consider these estimates to be less robust, hence we did not conclude that treatment effects were different between centres.



6) <u>Interesting reviewer point that "The assessor blinded outcomes deliver no difference between the groups, while the self-reported (non-blinded) outcome (including the primary outcome) clearly are in favour of the arthroscopic intervention. I think this aspect absolutely needs more discussion"</u>

<u>However, actually I do not find that the clinically defined outcomes are presented clearly, with effect estimates and CIs missing from Table 6. Is there actually evidence of a treatment effect or not for these outcomes?</u>

Response 41: Table 6 has been updated and discussed in detail in Response 5.

The primary aim of our study was to evaluate patient reported outcomes. Secondary outcome measures included clinical assessment. We found a treatment effect for hip flexion in favour of arthroscopic surgery, but are uncertain whether this is clinically important. Differences between other variables may reflect multiple statistical testing, but there is a suggestion of reduced pain on movement in the arthroscopic surgery group compared with the physiotherapy group.

7) **The discussion says: "The greatest improvement in symptoms was seen within the first five months post randomisation in the hip arthroscopy group, with continued improvement between five and eight months." – where are the results at 5 months? Perhaps I have missed them, but I cannot see them in the paper. Is there actually a decreasing trend over time, as implied? This comes out of the blue. Indeed, the 5 months is confusing, as in the results the authors actually say: "outcomes were measured eight months post randomisation and six months post intervention" – no mention of 5 months.**

Response 42: We acknowledge that this data requires clarification. As per the published protocol, we also recorded HOS ADL five months post randomisation, although it is not a primary outcome measure. Five-month post randomisation scores were available for 184 of 188 participants (100 in arthroscopic surgery group and 84 in physiotherapy group) of the mITT population, however, these scores include a heterogeneous population of patients pre and post intervention. We therefore had concerns over the validity of this results, but included it in the supplementary data as it suggests the majority of symptom improvement was achieved within five months post randomisation, consistent with other studies. Applying the same model as for the primary analysis, the treatment effect is 10.6 (95% CI 3.9-19.8) p=0.008. We have removed this data from the manuscript to avoid confusion.

8) **The CI for the treatment effect for the main outcome is wide and lower values within it may not be clinically important (based on the authors stating that 9 points is a clinically meaningful change). This needs discussion.**

Response 43: We clarified that there is a statistically significant difference in treatment effect between groups, of a magnitude that suggests clinical benefit, although the true difference may fall below the MCID. We also offer a more balanced interpretation of the study results.

9) **At the manuscript meeting, it was also noted that the control group is perhaps sub-optimal, with: physio therapy not very intensive (8 sessions over 5 months). It would be interesting to understand the authors' view on this.**

Response 44: Please see Response 8

10) **Table 2, the groups should be in separate columns to allow an easier comparison**

Response 45: Table 2 has been reformatted with the groups in separate columns.

11) **"Eight month post randomisation HOS ADL scores were higher than baseline in 70% patients who received arthroscopy compared with 50% patients who received physiotherapy. Clinically significant improvement, defined as a HOS ADL score greater than 9 points, was reported in 51% of patients who received arthroscopy and 32% of patients who received physiotherapy."**
- can we have some CIs please for the differences?

Response 46: We have inserted 95% confidence intervals for these results, as presented below:

	Physio	*Arthroscopy*
At Least One Point Improvement	*50% (39-61%)*	*70% (61-79%)*
At Least Nine Point Improvement (MCID)	*32% (22-42%)*	*51% (41-61%)*

12) <u>Why is the CI for the treatment effect WIDER after multiple imputation? Surely, the additional patients will add more information to the analysis, and therefore narrow the CI. Can the authors explain more about how their imputation was done in the methods and why this finding occurs.</u>

Response 47: The multiple imputation model presented in the manuscript included all variables in the primary analysis, as well as variables identified as important prognostic factors in Supporting Analysis B.

The model was implemented using multiple imputation by chained equations (MICE) to also obtain imputation for the low number of missing baseline variables, including missing radiological variables (lateral centre edge, maximum alpha angle, and KL grade). In line with best recommended practice, imputations were run separately by treatment arm. 30 imputations were generated, and imputations obtained using a predicted mean matching approach (pmm), choosing values to be imputed from the five observed data points with the most similar predicted scores (Stata command mi impute chained (pmm, knn(5))). Overall, 15% of observations were imputed, 20% in the physiotherapy arm.

Although we used the variables that were most predictive of outcome in our multiple imputation model, there is still considerable uncertainly around these estimates with some participants displaying significant variation in their imputed outcomes, giving larger CIs. Thus, rather than gaining power and precision by including additional participants, the imputation model produces results that are less precise than those from the complete cases mITT analysis. Instead, they minimise the bias, assuming incomplete data are missing at random. As such, we believe that our imputation results are an adequate reflection of the data, given the uncertainty around missing values.

We acknowledge that the use of simple imputation would have narrowed the confidence intervals, but by incorrectly assuming no uncertainty around these imputed values.