Author's response to reviews

Title: Evaluation of pain and function after two home exercise programs in a clinical trial on women with chronic neck pain - with special emphasises on completers and responders

Authors:

Linn Karlsson (linn.karlsson@liu.se)
Esa-Pekka Takala (Esa-Pekka.Takala@ttl.fi)
Bjorn Gerdle (bjorn.gerdle@liu.se)
Britt Larsson (britt.larsson@liu.se)

Version: 3
Date: 13 November 2013

Author's response to reviews: see over
Author`s response to reviews

Title: Evaluation of pain and function after two home exercise programs: A randomized controlled trial on women with chronic neck pain

Authors: Linn Karlsson, Esa-Pekka Takala, Björn Gerdle, Britt Larsson

Version: 3 Date: 15 November 2013

Author`s response to reviews: Dear Editor-in-Chief,

We are very grateful for the positive and constructive comments regarding our manuscript id: 8735595831019231.

In our opinion we have been able to make the necessary revisions, and the text has been through a new professional linguistic revision. Therefore it is our hope that the revised manuscript now can be accepted for publication in BMC Musculoskeletal Disorders.

We have appended a point-to-point reply starting on next page. The revisions in manuscript and tables are marked with yellow highlight.

With best regards

Linn Karlsson

Rehabilitation Medicine,
Dept. of Medicine and Health Sciences (IMH)
Faculty of Health Sciences
University of Linköping
SE-581 85 Linköping
Sweden
Telephone: +46 101034933
E-mail: linn.karlsson@liu.se
<table>
<thead>
<tr>
<th><strong>Reviewer 1 Kenneth Jay Andersen’s comments</strong></th>
<th><strong>Our changes, corrections and comments</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>The subject is relevant and the article states clearly the purpose of the study. It is well written apart from very few grammatical mistakes.</td>
<td>Thank you for the positive comments.</td>
</tr>
<tr>
<td>1. Usually there is only one primary outcome. There can be several secondary outcomes.</td>
<td>1. Yes, we agree with reviewer KJA that usually there is only one primary outcome. However, in this study of subjects with chronic pain, we considered both pain and function, which are our primary outcomes, as equally important. A recent systematic review referred in the Introduction (Kay et al 2012) did not only consider pain but also function. Thus, we still prefer and would like to keep both pain as function as primary outcomes. (So far) no changes has been made in the text.</td>
</tr>
<tr>
<td>2. Did you consider a third group to serve as a control? Preferably providing them some sort of sham intervention.</td>
<td>2. Good point. We considered this, but due to ethical reasons we chose to not include a third group with a sham intervention. The trial should run for a whole year and we did not want to expose a group of subjects for a known ineffective treatment for such a long time period. Other alternatives – to follow a waiting list or “treatment as usual” - was not possible due to the recruitment of participants from a general population. As far as we understand, the reviewer (KJA) does not ask for the above arguments in the manuscript, thus no changes or additions has been made in the text.</td>
</tr>
<tr>
<td>3. It would be nice to see data presented as mean (95% confidence intervals)</td>
<td>3. Good point. If we had managed to include 50 subjects in each group we definitely had presented the data in mean and confidence intervals. But, unfortunately, we ended up with a smaller sample size than wanted. Thus, we chose to present the data with median and performed non-parametric analyses. Due to the relative small groups we could not ensure that data were normally distributed and not either, on a safe theoretically ground, rely on the central limit theorem. This together with ordinal data in primary outcomes in table 2 (a, b see below) made us prefer median and non-parametric analyses. To make the result presentation more transparent we have now separated descriptive data from the result-analyze in the former table 2 into two tables (2a and 2b). In table 2b we now also have presented the differences between baseline and the two follow-up time points. Data in both tables are now presented with 25 and 75 percentile instead of the previously min and max.</td>
</tr>
</tbody>
</table>
4. Please provide argumentation as to why the strength test of performing upright rows and shoulder abd. for as many reps as possible up to 50 is a valid approach to test for increases in strength. Usually a one rep max/MVC is the standard.

4. Point taken. We chose this way of testing the strength in the shoulder muscles based on following: We assumed that some of the participant would be unexperienced with handling dumbbells and together with the pain in the tested area it would be easier for them to lift a predefined weight as many times as possible than to manage to perform a maximal contraction. In addition, the pain would probably increase if the subject had to, repeated times, test which weight that were max. If the subject should find the right weight at once it would not be a concern, but if the subject had to try several weights, the result may not reflect the actual MVC (due to increased pain from the testing). In this way the chosen testing procedure was supposed to be more correct and consistent for the subjects in this trial. In addition, this way of testing strength has been done previously. Thus, in the text we now have added: “This way of measuring shoulder strength, rather than testing MVC, was chosen because the subjects’ chronic pain probably limited their ability of performing a maximal and immediate contraction in the affected area.”

Methods, Outcomes, Shoulder strength, page 11.

How did you control for varying technical execution of this?

Good point. The technique were showed to the subject by the skilled physiotherapist who also checked the full quality of performance carefully and counted the movements loudly so the subject could follow the counting. This was made in exactly the same way for all participants.

Was it up to the tester/investigator if a repetition counted or not?

Point taken. Yes, it was up to the skilled physiotherapist (test leader). The two sentences below in quotation marks now are added in text, Methods, Outcomes, Shoulder strength, page 11: “When the subject did not manage to perform the whole movement, the test was stopped. It was easy for the physiotherapist to decide when the whole movement was not performed.”

If so is that a bias and a limitation to the validity of the results?

Good point. We do however consider the test leaders counting as a probably insignificant bias. In the Discussion section, Limitations of the trial, has been completed with the following sentences about potential bias, page 21: “This trial was only partially blinded, which could imply a potential bias. We were aware of this and tried to compensate for the methodological weakness with strictly structured measurement procedures and where the test leader did not encourage the performance during the tests.”
<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Also, why did you not use the handheld dyno that you used in testing the neck?</td>
<td><strong>In addition, the two interventions were instructed and supported with same structure and frequency to ensure equal approach for all participants.</strong> Good point. We did not use the handheld dyno because we thought it could be a motivating factor to test the movements, and exercise with the same movements (the strength group), and then test again. Nothing is changed in the manus text. Furthermore, by this way of testing the strength it was possible for the subjects to follow their progression with this procedure which probably encouraged the subjects during the long study period.</td>
</tr>
<tr>
<td>5. Why weren’t the drop outs included in the statistical analysis? You could have analyzed the data following the Intention To Treat (ITT) principle, which accounts for missing values. In a RCT this is considered the appropriate procedure.</td>
<td>5. Good point. We agree with the referee that ITT analyses could have been done and we considered initially to do an ITT analyze. However, we find our result analysis to be an equivalent alternative while the effects of the interventions were the major foci and the drop-outs was often specifically due to lack of time (in the discussion section, page 21) and not to the non-appreciation of the interventions. Furthermore, we were especially interested in evaluating our particular interventions. One critique against ITT analyses is that the effects of the trial becomes underestimated due to the fact that some of the participant don’t follow the intervention as prescribed.</td>
</tr>
<tr>
<td>6. Why weren’t the investigators blinded to participant group allocation?</td>
<td>6. Point taken. The flaw in blinding was due to limited resources. <em>&quot;There were no other components of blinding in the trial due to limited resources&quot;</em> is now added in the text, Methods, Randomization and blinding, page 8.</td>
</tr>
</tbody>
</table>

**Reviewer 2 Petri Salo’s comments**

This is a well written manuscript. The language is good with only minor misspelling and the text flow is enjoyable.

**Our changes, corrections and comments**

Thank you for the positive comments.

1. However, the study seems to lack power and thus it should be considered more like a descriptive study.

1. Point taken. Due to this comment we have revised the title in order to put more emphases on the completers and responders and less on the design as a RCT. The revised title is now: *“Evaluation of pain and function after two home exercise programs in a clinical trial on women with chronic neck pain - with special emphases on completers and responders”*

We also due to this and other comments of the reviewers have pointed out the lack of power in the Abstract, Conclusion, page 3, in the Discussion, page 18 and in
Limitations of the trial, page 21.

Moreover, we have rephrased the conclusion and included the lack of power. The following sentences in italics is now in the Conclusion, page 22: “We found no differences in the two primary outcomes between the two interventions, a finding that could be due to lack of statistical power. Both interventions based on home exercise resulted in improvements in pain intensity and function, which may implicate economic advantages compared to supervised interventions. However, home exercise requires a substantial motivation to ensure consistent adherence. It seems important, based on the relatively low adherence in the present study, that future research should include analysis of how to improve adherence for unsupervised exercise in subjects with chronic pain. In order to design clinically relevant studies, it may be important to focus both on power calculations and on estimations of adherence in order to achieve conclusive results. It is also essential to continue to examine exercise physiology and the relationships with pain physiology in order to optimize exercise therapies. A trial - even though under-powered as the present - can give important clues concerning such mechanisms by analysing clinical responders both in completers and non-completers.”

2. Did the authors perform power calculation to assess the sample size? It seems that the number of subjects in this study is rather low to reach statistically significant differences between groups (as the authors admit). If not, this and reason for that should be stated in the study limitations.

2. Point taken.

We did actually consider power and made a sample size estimate when designing the study. This is now added in the text, Methods, Sample size, page 12.

“Sample size

When estimating the sample size for analysing changes in pain intensity (one of the primary outcomes) within the groups, we assumed that the mean difference should have a standard deviation of 3. Expectation of a mean improvement of two points on the NRS, which also represents a clinically relevant improvement [33, 34], required a sample size of 20 pairs of subjects to reject the null hypothesis with a power of 0.80 and a probability of <0.05 (two tailed). When estimating the sample size for analysing changes in pain intensity (one of the primary outcomes) between the groups, we assumed that the mean difference should have a standard deviation of 3. Expectation of a mean improvement of two points on the NRS, which also represents a clinically relevant improvement [33,34], required a sample size of 36 subjects in each group to reject the null hypothesis with a power of 0.80 and a probability of <0.05 (two tailed). Sample size calculations were made using the computer program Power and Sample Size Calculations (v. 3.0.43, http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSi
Based on the sample size estimations, also considering a probable presence of non-completers, and our available resources for running this trial, we aimed to include 50 subjects in each group. We did not fully succeed due to a lack of eligible subjects available during a reasonable time-period for the recruitment.”

<p>| 3. Lack of power affects the conclusions drawn from the data. In the discussion the authors state that their findings that both groups STRENGHT and STRETCH showed similar improvements, contradict the findings of Ylinen, Andersen and Zebis. |
| 3. Point taken. The Discussion now starts with the following sentences, page 18: “Despite our very detailed planning and adequate financial resources, we did not manage to recruit a sufficient number of subjects within the available time. Hence, due to this lack of power, especially with respect to between group differences, the results have to be confirmed by larger studies and no definite conclusions with respect to the main aim can be drawn. In future studies, it may be important to consider other ways of recruiting subjects. With this pointed out we still believe it is important to discuss our results, as these results could be useful in the design of future studies.” |
| 3a) However if the authors study was unable to find differences between STRENGHT and STRETCH due to lack of power (type II error), this conclusion is misleading. Strength training and stretching might be able to produce significant differences between groups, but this data was not able to show it. Thus assimilation of studies that show adequate power with studies that do not show adequate power misleads the reader. |
| 3a) Point taken. We agree and have added the following sentence in the former first now second paragraph of the Discussion, page 18: “As mentioned above with respect to the achieved power of the present study, it cannot be excluded that differences may exist between the two interventions in a study that provides better statistical power.” |
| 3b) This applies also to the sentence “Our results that no difference exists between high-intensity strength training and low-intensity exercise with respect pain intensity and function....” |
| 3b) Point taken. We have rephrased that sentence, word in italics are now added: “Our result, based on the completer and responder analyse, that no convincing difference exists between high-intensity strength training and low-intensity exercise with respect to pain intensity and function agrees with the findings in other studies”, is now in the text, Discussion, page 19. |
| 3c) And also to the sentence “This conclusion agrees with the results |
| 3c) Point taken. We have rephrased that sentence, word in italics are now |</p>
<table>
<thead>
<tr>
<th><strong>in our trial where significantly increase strength in the STRENGTH group was not linked to….”</strong></th>
<th>added in the Discussion, page 20: Similar connection was found in our trial <em>in the within groups analysis</em> where significantly increased strength in the STRENGTH group not was linked to a significant decrease in pain intensity but the in the STRETCH group significant improvements in both strength and pain intensity were reported.</th>
</tr>
</thead>
</table>
| 4. Were the physiotherapists that did the clinical examination and outcome measures blinded? | 4. Point taken. 
“The examiner conducting the standardized clinical examination of the neck and upper extremities during the inclusion process was blinded with respect to group affiliation.” is now added in the text, Methods, Randomization and blinding, page 8. In addition, this is now also in the description of the clinical examination, Methods, Participants, page 6. |
| 5. In Methods, Subjects is stated “The participants also had to agree to follow the treatment plan as prescribed”. What does this mean? Informed consent also states that one can stop to participate in the study at any time without any reason. | 5. Point taken. 
It is important that this not can be misunderstood. This sentence has been rephrased. “The participant also had to declare that they were motivated to follow the exercise protocol”, is now in Methods, Participants, page 6. |
| 6. In abstract the training frequency of the STRETCH group should be added. | 6. Point taken. 
“Both groups were instructed to exercise three times per week”, is now added in the Abstract, Methods, page 7. |
| 7. In abstract results: correct the spelling with ….outcomes between the groups… | 7. Point taken. 
According to referees (referee PS, # 3 and 13 about power) and ours intention to present the study in a more descriptive way - this sentence with the misspelling noticed by the referee has been omitted from the abstract. |
| 8. In abstract results: correct the spelling with ….reporting clinically important… | 8. Point taken. 
Spelling is corrected “…subjects reporting clinically…” is now in the Abstract, results, page 3. |
In Method, Participants (previous subjects), there is a reference to NRS added, page 7. |
| 10. In the intervention section description of the stretching exercises states that …comprised retraction of the neck by stretching the following | 10. Point taken. 
The sentence is corrected to “…which comprised retraction of the neck and stretching the following muscles…” Methods, Interventions, page 9. |
musc
tes: m. trapezius…. This gives an idea that retraction of the neck was done by stretching all those muscles. Perhaps the authors mean ...comprised retraction of the neck and stretching of the ..... 

11. In results concerning both 4 to 6-month follow-up and 12-month follow-up, difference between groups could be presented before within group differences. I assume that differences between groups were the authors’ main interest.  

11. We agree that it could be assumed that differences between the groups are our main interest. However, due to the power problem we prefer not to highlight the possible group differences and therefore we would like to present within group changes first.  

12. The authors chose to present the results in table 1 and 2 using median values instead of mean values. Could you explain why? In table 2 statistics make it look a bit confusing when median neck pain in STRENGT drop from 6 at baseline to 2.5 in 12 months without being statistically significant and in the same time in STRETCH from 5 to 3 with p-value .009.  

12. Point taken. If we had managed to include 50 subjects in each group we definitely had presented the data in mean and confidence intervals. But, unfortunately, we ended up with a smaller sample size than wanted. Thus, we chose to present the results with median in table 1 and 2 and make non-parametric analyses. Due to the relative small groups we could not ensure that data were normally distributed and not either, on a safe theoretically ground, rely on the central limit theorem. This together with ordinal data in primary outcomes in table 2 (a, b see below) made us prefer median and non-parametric analyses. The result in the example PS points out can indeed look confusing, but the explanation lies in the limitation of sample size and application of analyze method. In fact, when applying a parametric analyze, the p-value for STRENGTH at 12 months becomes 0.05 and with non-parametric the p-value actually was 0.06 (just missing the possibility to reject the null hypothesis). Corresponding values for STRETCH at 12 months was 0.006 when applying a parametric analyze and 0.009 when applying a non-parametric analyze.  

To make the result presentation more transparent we have now separated from in descriptive data and the result-analyze formal table 2 into two tables (2a and 2b), pages 26 - 27. In table 2b we have inserted the differences between base-line and the two follow-up time points. Data in both tables are now presented with 25 and 75 percentile instead of the previously min and max.  

13. Is it relevant to present p-values for comparisons between completers and responders when there are only 2 to 6 (or up to 19 at the most) subjects in each group? See  

13. The question is very relevant. However, in the revised manuscript we due to other comments have pointed out the lack of power very explicit both in the beginning of the Discussion and in the Conclusion as well as in the Abstract. In our opinion the power problem is the important issue and we have chosen to retain the p-values even though the
comments above concerning power. Thus the nature of this study should be more like descriptive and presenting % values among completers and responders would be appropriate means for that.

<table>
<thead>
<tr>
<th>number of subjects are low in several of the analyses as pointed out by the reviewer.</th>
</tr>
</thead>
</table>

14. The progression of lifting head up from the supine position is not described in the interventions section.

14. Point taken. It is interesting to know exactly how the progression was for lifting head up from supine. “The exercise lifting the head up from supine were performed without resistance other than the weight of their head. First, the subjects were instructed to do as many repetitions as possible with the goal of 3 sets of 20 repetitions. After that, they were instructed to increase the number of repetitions. “This is now in the text, Methods, Interventions, page 8.

**Reviewer 3 Per Kjaer’s comments**

<table>
<thead>
<tr>
<th>Our changes, corrections and comments</th>
</tr>
</thead>
</table>

1a. This is an interesting paper that investigates two interventions for women with chronic neck pain. The interesting part is the analysis of those adhering to the programme (completers) and comparison to those who do not in relation to clinical relevant outcomes (responders). However, in the current form the important messages from the study are unclear. After a thorough revision the study definitely adds important new knowledge to the management of people with chronic neck pain even though the study sample is very small and the power of the study seem not to be considered. I have many concerns about the paper and I recommend a thorough review of the content where the consort statements [1] for the conduction of randomized trials and for reporting are adhered to and if not the authors have to give arguments why. Furthermore, several parts of the method section are inadequately and inconsistently described and lack definitions in important areas.

1a Point taken. Several adaptations of the structure and content of the manus according to the CONSORT statements suggested by the referee has been possible to do. The adjustments according to CONSORT are:

- The method section has been rearranged and clarified.
- Two sections have been added (Sample size and Harms)
- Some headings and subheadings have been renamed.
- The Limitations section of the discussion now is partly rewritten.
1b Part of the methods are reported together with results, which I find very confusing.  
1b See specifically points 13 and 16 below.

| 2. Abstract | 2a Please consider to revise the aims to reflect what is actually done. | 2. Points taken.  
2a The aim is now worded as: “The aim of this trial, which focused on adherence, was to evaluate two home exercise interventions.” In Abstract, Background, page 3.  

| 2b Please revise conclusion (moderate). | 2b. The conclusion is now revised and has the following wording:  
“No differences in the two primary outcomes between the two interventions were found, a finding that may be due to the insufficient statistical power of the study. Both interventions based on home exercises improved the two primary outcomes, but the adherences were relatively low. Future studies should investigate ways to improve adherence to home exercise treatments.” In Abstract, conclusion, page 3. |

| 3. Background | 3a Page 4: The background needs to further elaborate on adherence its influence on the treatment outcome. Especially the last paragraph needs to be revised. | 3. Points taken.  
3a Point taken.  
Revision of the paragraph about adherence and elaboration on adherence is done. “Adherence to prescribed exercise should be considered in evaluations of exercise interventions. Supervised or individualized interventions and self-management techniques may enhance exercise adherence [20]. More research, however, is needed to explore different aspects of adherence and its relation to positive clinical outcome [20, 21]” now is added to the text, Background, page 5.  

| 3b The aims are very general and do not cover the main focus for the paper, which seem to be analyses of these adhering to exercise. | 3b. Point taken. The study had one main and an additional aim. We now have revised the aim in order to better reflect that also an additional aim existed. Background, page 6: “The aim of is trial was to evaluate two home exercise interventions, one year of strength exercises and one year of stretching exercises. Within this aim, we have focused on adherence to the prescribed interventions. The participants were women with chronic neck pain who were recruited from a general population. Primary outcomes were pain intensity and function.”  
We are of course aware now that the main aim was not possible to answer sufficiently due to lack of power, but that was not known before the data collection started. In the discussion we in the revised version of the manuscript |
<table>
<thead>
<tr>
<th>4. Methods: The description of the study hypothesis after the aims no not cover the second or maybe the primary hypothesis that those adhering to exercise will have a better result. Please include this aspect.</th>
<th>4. Point taken. We have added this (in italics here) to the last sentence of the Background, page 6: “We expected the results to be more pronounced in the strength-training group and that greater adherence to the exercise interventions would produce better results.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Methods: Somewhere, an explanation of the variability of the 4-6 month follow up is needed.</td>
<td>5. Point taken. The variability of the 4-6 month follow up now is explained in the text. Thus, the following text is now added: “The follow-up time point in the middle of the trial (4 to 6 months) varied due to practical reasons such as on-going recruitment phase, participants summer vacations, and the fact that the same investigators performed all outcome measurements.” Methods, Outcomes, Outcome measurement time points, page 10.</td>
</tr>
<tr>
<td>6. Subjects: Explain who (a) and where (b) the examination was performed</td>
<td>6. Points taken. a) “The examination was performed by a physiotherapist especially trained for this task and who was blinded with respect to group affiliation”. This is now added in the text. Methods, Participants, page 6-7. b) The examination performed at, Linkoping University Hospital, which is now added in the text. Methods, Trial Design, page 6.</td>
</tr>
<tr>
<td>7. Randomization. Explain what is meant by “skilled physiotherapist”</td>
<td>7. Point taken. The description of the physiotherapists has been replaced from the Randomization paragraph to Participants and Interventions paragraph respectively. A skilled physiotherapist in this study is a physiotherapist experienced and especially trained for the task. However, we now have replaced the word skilled referred to in the comment and replaced it by “especially trained for this task” page 7 or “experienced in this area” page 9.</td>
</tr>
<tr>
<td>8. Since this data collection is important for the definition of responders this data collection instrument has to be explained.</td>
<td>8. Point taken. We believe that the reviewer may refer to completers instead of responder. Responders in this study are defined as following in the text, Methods, Analyses of results, Responder and non-responder, page 13: “The responder definition for pain and function was based on criteria for clinically important changes in the two outcome areas. Thus, for neck pain and shoulder pain, a decrease of at least two</td>
</tr>
</tbody>
</table>
points on the NRS should be reported [33, 34] and for function, a decrease of the total NDI score of at least four points was required [35].”

For completers; the data were taken from the exercise diaries. “Data from the exercise diaries were used for the completer analyse.” is now added in the text, Methods, Analyses of results, Completers and non-completers, page 13.

9. Redundancy of the exercise diary, the support for the exercise section is very vague. This may be important information and we need to know if all participants had the same support and exactly how it was.

9. Point taken. More about the exercise diary and support is now added: “The exercise diary was based on long-term and short-term goals, the latter also functioning as a detailed exercise plan. Furthermore, the exercise diary contained a weekly evaluation of the implementation of the training. The diary had the same structure for both the STRENGTH and the STRETCH group. Furthermore, support for adherence to the home exercise programme was provided by phone or e-mail every four to eight weeks. The support was more frequent at the beginning of the one-year training period and it was conducted in the same way for both intervention groups.” Methods, Interventions, page 9.

10. Pain intensity: which week?

10. Point taken. Pain intensity measurement has been clarified in the text of the manuscript: “Both primary and secondary outcomes were measured at baseline (BL) one day about two weeks before the start of the intervention. All outcomes were followed up in the same manner after four to six months of training (4 to 6 months) and after 12 months of training (12 months). “ Methods, Outcomes, Outcome measurement time points, page 10.

In addition; “The participants assessed their pain intensity in the neck and shoulders during the previous week...” is corrected in Methods, Outcomes, Pain intensity, page 10.

11. Statistics, please provide means and medians with 95% confidence intervals also.

11. This is a good point. If we had managed to include 50 subjects in each group we definitely had presented the data in mean and confidence intervals. But, unfortunately, we ended up with a smaller sample size than wanted. Thus, we chose to present the data with median and performed non-parametric analyses. Due to the relative small groups we could not ensure that data were normally distributed and not either, on a safe theoretically ground, rely on the central limit theorem. This together with ordinal data in primary outcomes in table 2 (a, b see below) made us prefer median and non-parametric analyses.
<table>
<thead>
<tr>
<th>12. An explanation of the wide range for the first follow up is needed.</th>
<th>12. Point taken. The variability of the 4-6 month follow up now is explained in the Methods, Outcomes, Outcome time-point measurements, page 10. The following text is now added: “The follow-up time point in the middle of the trial (4 to 6 months) varied due to practical reasons such as on-going recruitment phase, participants summer vacations, and the fact that the same investigators”.</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. The method section should also include definitions of responders</td>
<td>13. Point taken. Definitions of responders and also completers are now in the method section, Analyzes of results, pages 12-13.</td>
</tr>
<tr>
<td>14. A section on sample size considerations. What differences are expected? What is clinically relevant?</td>
<td>14. Point taken. A section on sample size considerations including what is clinically relevant is now added in the Method, page 12: “Sample size When estimating the sample size for analysing changes in pain intensity (one of the primary outcomes) within the groups, we assumed that the mean difference should have a standard deviation of 3. Expectation of a mean improvement of two points on the NRS, which also represents a clinically relevant improvement [33, 34], required a sample size of 20 pairs of subjects to reject the null hypothesis with a power of 0.80 and a probability of &lt;0.05 (two tailed). When estimating the sample size for analysing changes in pain intensity (one of the primary outcomes) between the groups, we assumed that the mean difference should have a standard deviation of 3. Expectation of a mean improvement of two points on the NRS, which also represents a clinically relevant improvement [33, 34], required a sample size of 36 subjects in each group to reject the null hypothesis with a power of 0.80 and a probability of &lt;0.05 (two tailed). Sample size calculations were made using the computer program Power and Sample Size Calculations (v. 3.0.43, <a href="http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize">http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize</a>). Based on the sample size estimations, also considering a probable presence of non-completers, and our available resources for running this trial, we aimed to include 50 subjects in each group. We did not fully succeed due to a lack of eligible subjects available during a reasonable time-period for the recruitment.”</td>
</tr>
<tr>
<td>15. Results The Figure legend says month, not week?</td>
<td>15. Point taken. Figure legend is revised, page 24. “...showed in weekly mean frequency for each month for the strength training group (STRENGTH) and the stretching group (STRETCH)”, is now added in the figure legend.</td>
</tr>
</tbody>
</table>
| 16. Page 13, completers and responders: this reporting has not been introduced in the study aims or in the method section. Please revise method section to include the first two paragraphs from this section – and rephrase. | 16. Point taken. 

The previous about completers and responders (on previous page 13) are now partially rephrased and moved to the Method section. Analyses of results, page 13. Furthermore, relevant parts of the text about completers and responders has been rephrased in Results, Completers and responders, pages 16-17. |
| --- | --- |
| 17. Discussion The discussion lack discussion on the limitations of the study, especially sample size. Please elaborate further on this | 17. Point taken. 

In the discussion section, limitations of the study are clarified and further elaborated on. The following sentences in italics is now in the text, Discussion, Limitations of the study, page 21: “Lack of power in this trial decreases the possibility of capturing statistically discernible differences between the groups. We aimed at including 50 participants in each group, but did not succeed due to practical reasons. Despite the lack of power, this trial provides interesting knowledge about statistically significant changes within the groups and a description of proportions of subjects reporting clinically relevant improvements in pain intensity and function.” |
| 18. Conclusion The conclusion does not really cover what was found in this study. It did not matter which type of exercise. Only half of those adhering had a clinical relevant improvements. Please rephrase and stick to what was investigated. | 18. Point taken. 

The Conclusion, page 22, has been revised and now have the following wording (the two last sentences are not changed): “We found no differences in the two primary outcomes between the two interventions, a finding that could be due to lack of statistical power. Both interventions based on home exercise resulted in improvements in pain intensity and function, which may implicate economic advantages compared to supervised interventions. However, home exercise requires a substantial motivation to ensure consistent adherence. It seems important, based on the relatively low adherence in the present study, that future research should include analysis of how to improve adherence for unsupervised exercise in subjects with chronic pain. In order to design clinically relevant studies, it may be important to focus both on power calculations and on estimations of adherence in order to achieve conclusive results. It is also essential to continue to examine exercise physiology and the relationships with pain physiology in order to optimize exercise therapies. A trial, even though under-powered as is the present trial, can give important clues concerning such mechanisms by analysing clinical responders both in completers and non-completers.” |
<table>
<thead>
<tr>
<th>should be week)</th>
<th>Figure legend is revised, page 24. “...showed in weekly mean frequency for each month for the strength training group (STRENGTH) and the stretching group (STRETCH).” is now added in the figure legend.</th>
</tr>
</thead>
<tbody>
<tr>
<td>20. please keep headings with table – very difficult to read</td>
<td>20. Point taken. We assume the referee refers to the previous table 2, which now is altered in two tables: 2a and 2b, pages 26 – 27, (see our answer to point 3 of Reviewer 1 above). We hope that the tables are easier to read now.</td>
</tr>
</tbody>
</table>