Author’s response to reviews

Title: An update of stabilisation exercises for low back pain: a systematic review with meta-analysis

Authors:

Benjamin E Smith (benjamin.smith3@nhs.net)
Chris Littlewood (c.littlewood@sheffield.ac.uk)
Stephen May (S.May@shu.ac.uk)

Version: 3
Date: 22 October 2014

Author’s response to reviews: see over
22nd October 2014

Dear Editor

Thank you for your response to our manuscript. Below is a point-by-point response to the issues raised by the reviewers.

Reviewer: Luciana Macedo
Reviewer's report:

1. I think that the authors should not have pooled all of the studies together regardless of two things: the duration of pain (acute or chronic) and the treatment comparison. The later is very important and incorrect. Studies that compared stabilization exercises with a control group were pooled together with studies comparing stabilization exercises with other exercises. The literature is very clear that exercises have a small effect (thus stabilization exercises versus control or minimal intervention should give significant results). But the literature suggests that there is no difference between different types of exercises. Thus pooling everything together is not only heterogeneous due to fundamental differences between studies but also does not allow for a full interpretation of the results. For instance, stabilization versus control could have been clinically significant, but we don’t know this as this analysis alones was not conducted. I believe that only the subgroup analysis presented in the study should be included.

We chose to pool all interventions together as we felt there was a need to evaluate against a valid comparator. Currently, usual care would amount to some intervention, so including all studies in the main analysis might offer a more conservative estimate of effectiveness and could be considered a more accurate reflection of effectiveness in the real world. We feel this is akin to the movement towards intention to treat analysis in clinical trials where conservative estimates are preferable to reflect real world practice

Minor Essential Revisions:

1. The authors should perform a systematic review of systematic reviews and make sure that they have not missed any given that many systematic reviews in the topic have been published. Including Macedo et al; Phys Ther 89:1; pg 9-25

Our review found 23 publications not included within Macedo et al review. Therefore we feel an update is more appropriate, rather than a review of systematic reviews. However, we have now referred to the previous systematic reviews in the introduction and discussion.

2. The agreement in study selection was very low. This is surprising as normally the agreement tend to be very high, especially when there is a defined inclusion criteria. I think it would be good if the authors provided information where the disagreements occurred: type of exercise, population, etc?
The following information has been added: “Initial disagreements were due to; intervention criteria [24–28], study population [29,30], study design [31–34] and duplication of results from another publication being missed [35,36].”

3. Given the controversy in using quality scores to pooled data, why did you do this? It is still not clear to me why you would choose this approach.

Pooling data by quality score is a recognised approach to minimise the impact of bias on the results.

4. You mentioned on the results “That left a total 18 publications for inclusion [25–28,57,62–75]”. (this is actually 19 references).

This has been correct to: “That left a total 18 studies for inclusion [28,38–40,61,64–77].” There were 18 studies, over 19 publications.

5. Conclusion: “there is a trend of worse fear avoidance scores” This is confusing, I did not understand what you meant with this statement.

The results show that fear avoidance score were favouring the control when compared with stabilisation exercises, but that there results were non significant. There was a non significant medium term mean difference of –2.0 (95% CI –5.1 to 1.0) and –2.7 (95% CI –7.6 to 2.1) respectively in favour of the stationary bike. Short term mean difference for FABQ (physical activity) was non significant at -1.58 (-4.00 to 0.84) and -0.18 (-2.42 to 2.07) in favour of sling and general exercises, respectively. Mean difference for FABQ (work) was non significant at -0.40 (95% CI -3.81 to 3.01) in favour of slings and 0.25 (95% CI -2.74 to 3.24) in favour of stabilisation exercises, when compared to general exercises.

6. I think you might need to better define the interventions that were included as part of stabilization exercises. For instance the studies by Marshall et al and Xueqiang don’t explicitly say that TrA or MF were of interest and thus, it is left to the readers imagination to whether these exercises actually contained the isolation of the deep trunk muscles. It is very possible that it did not and that stabilization exercises that deals only with the global muscles (as suggested by Stu McGill) could have been used.

This review uses definitions previously accepted within the literature, and one that matches the publications with which it is updating.

7. Why would a pilot study not be eligible for inclusion in the review? Why did you exclude this study?

By definition a pilot study is not powered to detect a difference between interventions and hence any attempt to conduct such an analysis is flawed and inappropriate. Hence, pilot studies were excluded from this review.

8. You mentioned that three studies included a package of treatment that
included stabilization. How can you assess the effect of the intervention within this package? Why did you decide to keep these studies in the review?

Akbari and Javadian were general exercises and core versus general exercise. Inani was core and electrotherapy versus general exercise and electrotherapy hence the effect of core exercise was ascertainable within these study designs.

9. I found that there is a lot of descriptive information about the studies on the text of the results. I think that this would be better represented on a table with duration of symptoms, type of comparisons, quality of the trials, etc.

We have slimmed down the descriptive information about the studies, as requested. All data is represented within the characteristics table.

10. Given that you had studies including populations with different duration of pain: acute, subacute and chronic, I think it is probably better to separate these studies. This is what is recommended given the better prognosis of acute and subacute patients and the fact that exercise is not often recommended for this group.

This traditional concept of acute versus chronic LBP is not supported by modern epidemiological studies. These show that the ‘good prognosis’ for acute LBP is not always apparent with 40-50% still haing symptoms at 3 and 12 months, and of those who recover 40-50% having another episode in the following year.

11. How was clinical significance assessed? You do not mention this on your methods or results and it is left to the reader’s imagination to guess this.

Clinical significance was assessed through the Minimal Clinical Important Difference (MCID), which was discussed in paragraph 1 of the discussion. We have added more information within the results section.

Pain: “However, the difference between groups was clinically insignificant with Minimal Clinical Important Difference (MCID) for pain being suggested as 24 to 40 [97]”

Disability: “The difference between groups was clinically insignificant, with MCID for RMDQ 17 to 21 and 8 to 17 for ODI (converting all to 0 – 100 scale) [97].”

12. It is odd to me that statistical heterogeneity was presented in the same sentence as clinical significance. This may cause one to think that they are related. Maybe this needs to be made clearer

We have been unable to identify where this has occurred, even after using the search feature of word.

13. At long term there was no significant statistical or clinical difference; should read: At long term there was no clinical or statistical significance.
To be clearer the sentence now reads “At long term there was no statistical or clinically significant difference”.

14. I2 does provide some information on the strength of the results but I don’t think it should be the sole measure to evaluate the robustness of the analysis. This could even be a chance finding from small trials or something like that. Other things need to be taken in consideration. I think that the meaning of I2 has been a little overused in the study. You should consider using GRADE or any other method to evaluate the strength of the results. At this point you used only I2 to evaluate the strength of the evidence, whether other methods are more comprehensive.

We haven’t used $I^2$ as the sole measure of robustness of the analysis, but simply as an indication of heterogeneity.

15. Why did you include the results of the FABQ in your manuscript since this was not in your methods?

We feel that the FABQ offers an alternative outcome measure for functional disability for patients with low back pain. It has been validated as a disability outcome measure by George et al Clin J Pain. 2006 Feb;22(2):197-203, and fits within our clearly stated aims and methods.

16. I think that supplement 1 could be made smaller and simpler and presented as part of the manuscript, especially if the manuscript text can be reduced.

As per comment 9, we have slimmed down the descriptive information within the text of the document. Therefore the information within the characteristics table is solely placed there, and essential.

We specifically had a characteristics table that had full details of the intervention and outcome measure, based upon feedback from clinicians. It was felt it allowed better understanding and comparison of the specific exercises used.

17. Also, I don’t see the need for supplement 2 given that the PEDro scores are readily available on the PEDro database.

This has been removed, as recommended. “The PEDro scores ranged from 4 to 9 [42], with mean score of 6.6 (please refer to the PEDro website for score breakdowns).”

Reviewer: Neil O’Connell
Reviewer’s report:

Abstract
Intro - I would reconsider the sentence regarding the lack of evidence to support its use as it is ambiguous. A lack of evidence or a lack of positive evidence? There is a fair amount of evidence about - hence your review. If a lack of supportive evidence would that not preclude the question of your review?
Modified as requested. Sentence now reads: “Despite it being the most commonly used form of physiotherapy treatment within the UK there is a lack of positive evidence to support its use.”

Results – When presenting percentage effect sizes specify what it is a percentage of. When mentioning clinical importance please specify the threshold.

We have referred to other published examples in BMC MSK disorders and it appears that the level of detail we have presented is commensurate with convention. Within the confines of the specified word limit it is not possible to add further detail.

However, further detail on the outcome scale is now in the abstract.

When mentioning significance state that it is statistical significance.

Sentence now reads: “When compared with alternative forms of exercise, there was no statistical or clinically significant difference.”

Conclusion – What is meant by “person specific”. Do you mean individually tailored? If so this is not explicitly specified in the inclusion criteria later and is unlikely to be true of group exercise studies.

Removed, as requested. Sentence now reads: “There is strong evidence stabilisation exercises are not more effective than any other form of active exercise in the long term for non-specific low back pain”.

Background
NSLBP - is not really defined as without cause, but without a cause that we can reliably identify.

This review uses definitions previously accepted within the literature.

The 62% statistic from Hestbaek for long term problems would benefit from some qualification. A much smaller % might be expected to have serious disabling chronic problems.

As requested, sentence now reads: However, 62% of people experiencing their first episode of LBP will develop chronic symptoms lasting longer than one year, with 16% of people still sick listed from work at 6 months [4].

What is meant by the median duration of survival time?


Do you mean improvement or growth of the evidence base?

Sentence corrected to: “Since 2006 there has been considerable growth in the evidence base, with a large number of new trials being published”
In justifying the review it would be worth also referring to the review by Bystrom et al (2013 Spine 38:6) and its critical summary by the CRD http://www.crd.york.ac.uk/crdweb/ShowRecord.asp?LinkFrom=OAI&ID=12013023228

We have now referred to the previous systematic reviews in the introduction and discussion.

Why would meta-analysis be a secondary aim - it is a critical tool in achieving the primary aim. Were you not interested in comparisons with no treatment or usual care? If not it might be useful to explain why.

A meta-analysis is useful if studies are considered suitably heterogeneous. The original review this one is updating was unable to do a meta-analysis and performed a qualitative analysis instead. Therefore our plan was to conduct a meta-analysis only if it was possible.

Our comparison has been corrected to: “Comparison group of any other intervention, placebo or control were considered appropriate.”

Methods

The search strategy seems quite limited. It is possible that a broad selection of terms might have been more sensitive. EMBASE or SCOPUS were not searched - could the authors comment on whether they may have missed any trials that way?

An inclusive search was conducted which meets relevant criteria. Our discussion covers this topic under the heading “limitations of this review”.

Was any search made for unpublished studies or of the grey literature?

Our method clearly states we only included studies published in peer review journals, and our discussion covers our justification. “It is thought that identifying unpublished trials minimises publication bias [102]. However, this approach has been questioned by others, who suggest that truly unpublished trials frequently have poor methodology, and ones with better methodology often eventually become published [103]. It is not possible to know if the inclusion, if available, of any unpublished trials would considerably alter our conclusions, or if this truly is a weakness of this review.”

Study selection - were patients with radiculopathy included?

Yes, our definition of back pain was: “Low back pain defined as, but not restrictive to, pain and/or stiffness between the lower rib and buttock crease with or without leg pain”
Study selection - why have a single author screen the titles and abstracts?

This was undertaken for pragmatic reasons relating to time. It is recognised that use of a single reviewer might increase the potential for bias and error during this stages. However, it is interesting to note that there is movement in the field of systematic review methodology towards an appreciation of rapid reviews. Frequently such reviews use one reviewer at the various stages for pragmatic reasons and although it is recognised that the potential for error is higher, it is generally suggested that most errors or omissions do not lead to substantial changes in any conclusion, e.g. Jones et al (2005). High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol. 2005 Jul;58(7):741-2.

Can the authors comment on possible reasons for the low level of agreement relating to full text screening?

The following information has been added: “Initial disagreements were due to; intervention criteria [24–28], study population [29,30], study design [31–34] and duplication of results from another publication being missed [35,36].”

Page 8 line 156 - correction - authors were contacted.

Corrected to: “One study published median outcome scores, and the authors were contacted and provided mean outcome data [40].”

Were authors contacted where data required for meta-analysis were not available?

All data were available.

Quality Assessment - correction - in a number of places “were” should be replaced by “was”.

Corrected

Can the authors comment on the decision to use PEDRO scores from the database rather than quality assessing the papers themselves with 2 independent reviewers. This feels like a shortcut.

We feel this was a very sensible decision removed any chance of errors. The PEDro database is already double scored by experienced reviewers, with systems in place to check and monitor scoring.

Were studies scoring <6/10 considered to be of low quality? The use of these cut-offs is common but questionable - for example you have a “high quality” paper that reported insufficient data to analyse. So it is at high risk of bias for selective outcome reporting.
This approach has been validated, and is consistent with previous published reviews, for example the review that this review is updating. It is also consistent with the previous review on this topic previously mentioned Bystrom et al (2013 Spine 38:6)

Statistical Analysis – why were results converted from their original scales? How was this done for disability scales such as the RMDQ? Why not use the standardised mean difference? This section requires substantially more detail. What was the established a priori plan for the evidence synthesis? Which comparisons were planned? With regards heterogeneity was the Chi squared test also used to assess the statistical significance of the heterogeneity? How was it decided whether trials were “sufficiently homogenous”. Was any subgroup analysis planned? Was any investigation of small study effects made?

Data results were standardised, so there was comparability between different trials using different outcome measures. This had been done in the previous review, which this review is an update of, where it was supported by this reference: Herbert RD. How to estimate treatment effects from reports of clinical trials. Aust J Physio 2000;46:229-235.

The priori plan for the evidence synthesis is stated within the aims of the review: “to determine whether stabilisation (or ‘core stability’) exercises are an effective therapeutic treatment compared to an alternative treatment for people with non-specific low back pain. The secondary aim is to determine if stabilisation exercises are as effective as other forms of exercise”.

The Chi square test was not used to assess the statistical significance of the heterogeneity. The Chi squared test has been shown poor at detecting true heterogeneity, especially when total number of included studies are low, as is often the case with meta-analysis (Higgins et al., 2003). Higgins et al. (2003) devised the $I^2$ test, which they showed to be more robust and a better method for detecting true heterogeneity between studies. Studies were considered ‘sufficiently homogenous’ from the studies characteristic comparison, couple with the $I^2$ test. (Higgins J, Thompson S, Deeks J, Altman D. Measuring inconsistency in meta-analyses. BMJ. 2003 Sep 6;327(7414):557–60. )

All planned subgroup analyses were included within the written report.

Results

It is odd to reject papers in an update of a systematic review that were included in the previous versions. More detail on how the participants in those studies no longer met the criteria are needed.

More detail added: “From the 18 included studies from the 2008 review, seven were rejected for this review, five due to this review only including patients with non specific back pain [20,78–81], one because it was a pilot study [82] and one due to inappropriate outcomes [83].”
Study quality and bias - correction “….failed to assess baseline comparability”

Corrected to: “and one study assess baseline comparability [72]"

Be clear when using the word significant whether you are referring to statistical or clinical significance.

This is been re-written to be clearer.

There needs to be a clarification of the scales - do they represent the scales all converted to 0-100 or some form of percentage as implied in the abstract. If a percentage what is it a percentage of? Baseline score? Post treatment control group score? What threshold are you using for clinical significance? Without these details it is not possible to meaningfully interpret the results. They appear in the discussion but need to be more clear and precise, presented in the methods and stated whether they were established a priori.

A scale of 0-100 is a percentage. For clarification the text has been modified for consistency, with ‘scale of 0-100’ used. More detail on clinical significance has been added.

Forest plots - it would be useful to see the plots for the primary analyses rather than just those for subgroup analyses. It would also be useful to see the number of participants clearly displayed in the forest plots.

Modified, as requested. Unfortunately our stats package cannot add participant numbers to the forest plots.

Were the subgroup analyses pre-planned or post hoc? It is vital that this is made clear.

The priori plan for the evidence synthesis is stated within the aims of the review: “to determine whether stabilisation (or ‘core stability’) exercises are an effective therapeutic treatment compared to an alternative treatment for people with non-specific low back pain. The secondary aim is to determine if stabilisation exercises are as effective as other forms of exercise”.

Discussion:

The most likely explanation for the high heterogeneity is arguably the different comparisons being made between trials - this should be acknowledged.

As per your comments the following line has been added: “The high heterogeneity is possibly due to the different comparisons being made between trials, and this reduces the robustness of our short to medium term results”

You state here that 2 authors independently extracted data but in the methods one author did this with another author checking. This needs to be consistent
across the paper.

Corrected to: “with two reviewers screening full texts independently for inclusion and the data extracting independently checked by”

Discretionary revisions
Title
This could be clearer - what is the specific question - effectiveness? Compared with?

Our title is consistent with the title of the review which is being updated.

The discussion is quite strident in suggesting the increase in fear avoidance scores is the fault of the core stability paradigm. While I have sympathy with the theoretical suggestion this should be substantially toned down as there was no significant difference and the difference may also be clinically meaningless if it is there at all.

As suggested, the theoretical suggestion has been toned down significantly.

We trust this response is satisfactory and look forward to hearing from you in due course.

Yours faithfully

Benjamin Smith, Chris Littlewood & Stephen May