Author's response to reviews

Title: Effects of an exercise and manual therapy program on physical impairments, function and quality-of-life in people with osteoporotic vertebral fracture: a randomised, single-blind controlled pilot trial

Authors:

Kim Bennell (k.bennell@unimelb.edu.au)
Bernadette Matthews (matthews@unimelb.edu.au)
Alison Greig (alison.greig@ubc.ca)
Andrew Briggs (A.Briggs@curtin.edu.au)
Anne Kelly (amkelly@unimelb.edu.au)
Margaret Sherburn (m.sherburn@unimelb.edu.au)
Judy Larsen (hctjudy@hotmail.com)
John Wark (jdwark@unimelb.edu.au)

Version: 3 Date: 19 January 2010

Author's response to reviews: see over
Dear Sir/Madam

We thank you for the opportunity to revise our manuscript based on the reviewer’s helpful comments. We outline below our responses to the reviewers comments. We have also highlighted the parts of the manuscript in yellow where we have made changes to the text and table 2. The main change is that we have re-analysed the data according to the suggestion of the statistical reviewer (Reviewer #3).

We hope that we have addressed the comments to your satisfaction and that our revised manuscript is now suitable for publication. However, we are happy to make further revisions if deemed necessary.

Yours sincerely,

Kim Bennell on behalf of the authors

**Reviewer 1**

*The reviewer outlined the limitations of the study including small sample size, lack of long term followup and lack of control intervention*

As the reviewer highlights, these have all been discussed in the limitations paragraph of the manuscript. No changes were requested by the reviewer.

*Page 6: were the thoracic spine mobilizations central of unilateral?*

The mobilizations were central and this has been added to the sentence on page 6 as well as in Table 1 which outlines the treatment protocol.

*Page 7: end of the first paragraph: do you have any references to support the use of the 10 week intervention in order to achieve strength gains in postural muscles and to increase spinal mobility?*

Two references to support this statement have been added to the manuscript.
Reviewer 2

I would suggest using a non-parametric method for assessment of between group differences.

Given the comments by the statistician reviewer (Reviewer 3) and their direction to use ANCOVA (general linear models adjusting for the baseline value of the measurement), we will not perform non-parametric methods of assessment.

There are discrepant data in relation to the number of men. Three men are mentioned in the results section (page 10) while four are implied from Table 2.

Thank you for spotting this discrepancy. The number in Table 2 was incorrect and the number should have been 3 males. This has been corrected in the Table.

The SDs of change (Wk 11-0) for most variables is higher for the intervention than the control arm (table 3). The independent t-test assumes similar variance which does not appear to be the case here. I would suggest that the authors use also non-parametric statistics to look at between group differences.

Given the comments by the statistician (Reviewer 3) and their direction to use ANCOVA, we will not perform non-parametric methods for assessment

Reviewer 3 (Statistician)

1. The outcomes in this study measure change from baseline between intervention and control. The appropriate method of analysis is an analysis of covariance that compares treatment groups with the measurement at baseline as the covariate, not a two sample t-test [1][2]. Analysis of covariance will be more powerful and allow the inclusion of additional covariates, although the number of these will be limited by the sample size. Therefore this method of analysis should be used.

As directed by the statistical reviewer, we have re-analysed the data using ANCOVA (general linear model) adjusting for the measurement at baseline as the covariate. Given the small sample size and the fact that the groups were similar at baseline, no other covariates apart from the baseline value of the measurement were added. Table 3 has been changed to reflect the new numbers. This new analysis changed the results slightly. Qualeffo total and Qualeffo pain which were significant in our prior analysis just failed to reach significance when re-analysed using ANCOVA. Thus, we have changed the abstract, results and discussion sections where appropriate to reflect this. However, overall the conclusion of the study has not been altered as the
primary outcome of pain is still significant as it was in the prior analysis.

Re-analysing the data by excluding the three males, all from the intervention group, achieves nothing except further reducing power in an already under-powered study.

We agree and this sentence has been removed from the results section.

2. Differences between groups should be presented with their 95% confidence intervals as the authors do in Table 3 not as p-values which are presented in the text. P-values by themselves only show that the mean difference was significantly different from zero. It is impossible to evaluate the importance of a mean difference without stating the clinically important difference. This was only given for the pain score.

The p-values have been removed from the text as suggested by the reviewer. Minimal clinically important differences (MCID) were only provided for the pain score as it was the primary outcome and also the MCID has not been established for the other outcomes particularly in this patient population.

Statistical significance inevitably arises by chance, as the authors acknowledge, and this is one of the reasons that the p-value is conventionally set at 0.05, not because the study is a pilot. There is no adjustment of the p value to account for multiple tests (approaching twenty). If an adjustment is made it is likely that most significance, if not all, would disappear.

The sentence stating that the p value was set at 0.05 because this was a pilot study has been removed and we have just stated that the p value was set at 0.05. We set the p value at p<0.05 rather than a more conservative p<0.01 given our multiple tests because we recognize that the power of our study is limited by our small sample size. The issue of whether or not to adjust for multiple statistical tests is contentious and the use of Bonferroni tests to correct for multiple testing is often criticized (eg. Nakagawa 2004). In light of the controversy and the fact that the reviewer did not advise this to be done, we have not corrected for the multiple tests and this point was raised in the discussion.

3. The statement that "The intervention group showed clinically meaningful reductions in pain on movement and at rest following the 10 week treatment" contradicts the earlier statement that "a change of at least 2 points is thought to represent a clinically meaningful improvement" as both changes in pain within this group were smaller than 2 points (see Table 3). But as the aim of the study is the comparison between intervention and control groups not change from baseline within groups this comparison is unnecessary.

We agree that the way the sentence was written was misleading suggesting that the change in pain within the treatment group was ≥ 2 units which is the clinically meaningful amount. The reviewer is correct in that change in pain within the treatment group did not reach this. However, the difference in
change between the two groups did. Therefore, we have rewritten this sentence to make it clear as to what we were alluding to.

4. Subjects rated their perceived change in back pain on an ordinal scale (1-much worse, 2-slightly worse, 3-no change, 4-slightly better, 5-much better). Fisher's Exact test depends on the hypergeometric distribution and is distinct from the chi-squared test. It is used when the assumptions needed for the chi-squared test are not met. These include sufficient counts in each cell, this could not possibly be met with 20 observations spread over 10 cells, as a result a five point ordinal scale has been illogically split between "5-much better" and everything else, this makes "4-slightly better" the same as "1-much worse". This cannot be right. The calculation of the relative risk is just as wrong and is an unnecessary second analysis on the same data. It is questionable whether these analyses based on "much better" versus "everything else" demonstrate anything. It may be preferable to wait for a subsequent study and perform an appropriate analysis on a sufficiently populated 2 x 5 table. The responses from the pain questionnaire should be presented as collected which avoids loss of much relevant information.

We agree that given the fact that there were insufficient counts in each cell to do a Chi square test, it is best to present the data descriptively and so have we have described the number of responses for each of the categories in the text. As suggested, we have not dichotomized the data and have removed the calculation of relative risk. The abstract and results have been changed to reflect this.

5. There is no reference to the statistical software used to analyse the data.

SPSS was used – this has been added to the statistics section in the methods

From the data presented in the manuscript it is only possible to perform a sample size calculation for back pain as this was the only outcome where a clinically meaningful effect size was stated. This was stated as "at least 2 points". The pooled estimate of the standard deviation is approximately 2. This gives a standardised effect size of 1. Using Machin and Campbell et al [4]. Using the authors’ specified p-value of 0.05 Table 3.5 shows that to achieve a power of 0.90, the usual power in clinical trials, 27 (twenty seven) subjects per group would be needed. The power actually achieved with nine in one group is about 0.5. Equal sized groups is the most powerful design, imbalance reduces power. Equal groups could have been achieved by using block randomisation and this should be considered for any subsequent study. The wide confidence intervals in Table 3 are a consequence of low power.
We thank the reviewer for this comment. We agree that the study had lower power because of the small sample size and that the wide confidence intervals in Table 3 are a consequence of this. This was stated in the limitations paragraph in the discussion. However, despite the lower power, we did detect significant differences in several outcomes between groups.

Reference: