Reviewer's report

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

Version: 2 Date: 8 December 2009

Reviewer: John Hughes

Reviewer's report:

Major Compulsory Revisions

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.

The talk by Martin Bland [3] at http://martinbland.co.uk/ shows how the effect of different "operators" can be accounted. This has to be done at the design stage so it is too late to adjust for this in this study.

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.

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.

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.

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.

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.

Minor Essential Revision

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

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 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.

3. Grouping in individually randomised trials Talk presented at the 4th Annual Conference on Randomised Controlled Trials in the Social Sciences, September 2009, York, UK. (September)
Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests.