PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Pharmacist-led management of chronic pain in primary care: results from a randomised controlled exploratory trial
AUTHORS	Bond, Christine; Bruhn, Hanne; Elliott, Alison; Hannaford, Phil; Lee, Amanda; McNamee, Paul; Smith, Blair; Watson, Margaret; Blyth, Annie; Wright, David; Holland, Richard

VERSION 1 - REVIEW

REVIEWER	Peter Watson Statistician MRC Cognition and Brain Sciences Unit 15 Chaucer Road Cambridge CB2 7EF UK I have no competing interests connected to the research in this paper.
REVIEW RETURNED	06-Dec-2012

THE STUDY	I was confused by the rationale of the paper. The authors state that since it is an exploratory study with no previous work it is not powered for the purpose of statistical testing (page 9). In this scenario descriptive statistics and effect sizes would be appropriate. The study however features (Tables 2-5 on pages 22-25) p-values which are carried out to test apriori statistical hypotheses but giving p-values assumes (a) there are apriori hypotheses and (b) the study is adequately powered to allow differences to be found using statistical tests and preclude the 'file-drawer' problem.
RESULTS & CONCLUSIONS	I was confused by what the p-values in Tables 2 to 5 (pages 22-25) were testing. Could these also be explained clearly in the table to explicitly say what is being tested. For example the between group p-values in Tables 2 and 3 appear to test for a time by group interaction to see if there are different rates of improvement between baseline and follow-up in the various groups. I am also not clear of the reason for the 'Not Valid' p-values in Table 4 when exact tests are available and find the p-values for (I assume) the marginal homogeneity tests of the signs of change counterintuitive but this may be because it is not clear what is being tested.
GENERAL COMMENTS	The aim of this randomised control trial (page 5) is to compare responses in patients who are randomly allocated to one of three treatment groups (page 7): treated by a personalised prescription issued and followed-up by the pharmacist, a review of existing medication followed up by the GP or an ‘as usual’ (control group). Questionnaires are posted (page 8) at three and six months after follow-up to compare responses with baseline. The responses (page 8) compared are the Chronic Pain grade (CPG in Table 2, page 22)

	disability and intensity scores, a general health index the SF-12v2 (in Table 3, page 23) with physical and mental components and the Hospital Anxiety and Depression Scale (HADS) (in Table 4 on page 24). In addition to changes in average responses over time (comparing baseline and six months) the sign of these differences is also compared across the three groups using numbers of individuals with negative, zero or positive differences. I think the statistical tests are appropriate but find the description of the results (especially the p-values) in the tables somewhat confusing and also would like a little more focus on what the effect sizes are and at least a mention of a caveat regarding the interpretation and use of p-values in an exploratory study that has, consequently, not been powered (page 9, third paragraph). One could argue that descriptive statistics only are more meaningful in a truly exploratory analysis as there will be no apriori hypotheses to either test or power so no need for any p-values or statistical inference. Page 9. An aim quite correctly in any exploratory study such as this is to estimate an effect size (third paragraph). I, therefore, wondered what form(s) this effect size takes. Is this a standardised difference in means between groups or the unstandardised difference in means between groups or an eta-squared (for the one-way anova)? Would a phi coefficient be appropriate for the chi-square tests? Field (2005) also gives an easily computed effect size for nonparametric tests. On a related note do we have any idea how large a difference would need to be, irrespective of its statistical significance, to be regarded as clinically interesting? The effect size is key here as the study has not been powered which would mean any statistically non-significant results may be due to having too small a sample size to detect a clinically meaningful difference. I, therefore, wonder about the usefulness of quoting p-values in Tables 2 to 5. The effect size, however, should not be influenced by sample size so should give us a good idea of any clinically large differences and should be expressed for all comparisons. If any large effect sizes are found one can then (as suggested on page 9, third paragraph) see if these effects would be replicated in a larger study which would be adequately powered to have enough people in it to detect such a difference and p-values could then be used to test for any associated hypotheses in the larger study that were suggested by this pilot study. Page 9. Did you consider testing for between factor (group) by within factor (time) interactions? This is not explicitly mentioned in the final (analysis) paragraph on page 9 but I think it is tested for in Table 2 by the between groups p-value (far right) testing the difference over time within group (ie a group by time interaction). Pages 22 to 25. I think the p-values could be more clearly expressed in the captions and body of Tables 2 to 5. For example you could put the p-value next to the statistic it relates to e.g. difference CPG intensity=-8.0 (16.34),p=0.002 in Table 2 and the p-value for testing between groups (presumably the three differences in average CPG intensity between baseline and 6 months follow-up in Table 2) on the same row as the actual numbers it is testing (see Table 2 the first two P(between groups) p-values). I assume the between groups p-values for the counts at the bottom of Tables 2 (page 22) and 4 (page 24) are testing the group profiles of the 3x3 tables of counts for CPG grade differences (Table 2) and HADS-A so testing to see if

	groups improve at different rates (ie the group by time interaction). I am also not clear why nonparametric tests (with medians) are used for the CPG disability and SF12 mental scales but parametric tests (with means) used for the CPG intensity and SF12 physical scales. Is one subscale in each test skewed or having a limited number of responses to motivate the use of a nonparametric test? Page 24. It is not clear to me what the within group p-values are comparing in Table 4. For example, I notice a p-value of 1.0 in Table 4 is given for Prescribing Difference HADS-D which I assume is testing whether the counts for the signed HADS-D difference of 5,34,5 are equal (where there seem a lot more zero differences than positive or negative ones). I therefore wondered if you were testing if the number of negative HADS-D changes was equal to the number of positive HADS-D changes which could give a p=1.00 for the prescribed group but then in the next column for the review group we have three counts of 4, 37 and 4 respectively for <=-1,0 and >=1 which has a p-value of 0.71 rather than 1.0, despite the number of positive and negative changes in the review group also being the same. The “Not Valid” in Table 4 for P(between groups) suggests it is not possible to do a statistical test of the two-way frequency tabulation owing to small numbers in the cells but a Fisher exact test can still be used here (recommended by Howell (1997) when the expected count in any cell is less than 5). It might also aid interpretability if the differences in CPG (Table 2, p22) and HADS grades (Table 4, p24) was expressed as improvement (presumably <=-1) or deterioration (>=1) or no change in these tables. References Field, A (2005) Discovering Statistics sing SPSS. 2nd Edition. Sage:London. Howell, DC (1997) Statistical methods for psychology. Fourth Edition. Wadsworth. Belmont,CA.

REVIEWER	D.K. Raynor, Professor of Pharmacy Practice, University of Leeds, UK and Academic Advisor, Luto Research
REVIEW RETURNED	19-Dec-2012

THE STUDY	Slight rewording needed in the Abstract and Key Messages
RESULTS & CONCLUSIONS	The interpretation and conclusions are in general appropriately described, but I suggest below a slight softening of the wording in the Abstract, Key Messages and Conclusion.
GENERAL COMMENTS	GENERAL This is an important paper in a field where controlled studies are few and far between. My only significant concern is that in the Abstract and the Conclusion the choice of words could be argued to 'over-claim'. At the end of the Abstract I would suggest replicating the wording used in the Discussion later to make the wording here '...and suggests THERE MAY BE a benefit for patients....'. This also applies to the first Key Message. Then in the main Conclusion suggest

	'....acceptable and MAY lead to improvements...'. Care over this wording needs to be taken, bearing in mind the acknowledged limitations in the main text(e.g. 25% of eligible patients entered trial; and not knowing how important the observed differences were to participants). MINOR Abstract: - line 6 suggest '...with or without PHARMACIST prescribing' - lines 32-37 I would remove reference to the remote telephone randomisation as it could create confusion regarding the nature of the interventions Intro - It would be helpful in the second para to note that NSAIDs are commonly used to treat chronic pain Method - Line 31 Reference is made to an 'independent pharmacist prescriber' here and then on Line 49 'supplementary prescribing'. The relevant para in the Introduction needs some more detail on the types of 'non-medical prescribers' - Lines 33 and 34 - more detail would be helpful regarding both the pain diary and the pharmaceutical care plan - and could the templates for both be made available? - Line 18 the wording and referencing in this last sentece of the para needs clarifying RESULTS - Line 38 'generally conducted by phone' - would be better described numerically DISCUSSION - Line 2 suggest '...to both patients and MOST professionals' - Line 19 'assessed' not 'demonstrated' - Lines 22-24 The fact that formal training was given to the pharmacist is a strenght and could be acknowledged here - Lines 34 - 37 The meaning of the sentence beginning 'Rewording of participant recruitment documentation....' is not clear; how could this address these concerns? - Line 15 need to add '...in the UK' at the end of the sentence FIGURE 1 Although many of us are familiar with the Grampian region and its good citizens, readers further afield may be confused by use of the term 'Grampian'. Suggest reword as 'Grampian region' here and in the Methods aslo

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Alison Walker, Associate editor, BMJ Open

1. *Please discuss differences in outcomes from the trial registration website:
Paper says primary outcomes are: CPG and SF12 (HADS as a secondary outcome). Registration site says primary outcomes are: SF12 and HUI 2/3 (with CPG and HADS as secondary outcomes)

Our response:
The entry in the ISRCRT No. 06131530 was for a two staged study including an initial small feasibility

study (two practices, two prescribing pharmacists and target of 24 patients, unit of randomisation the practice) in which we developed the training package, did pre and post qualitative work with GPs, Pharmacists and patients to maximise acceptability of the intervention and select outcome measures. In this feasibility study, our planned primary outcome measures were the SF36 (subsequently replaced by the SF12 during development work on the questionnaire to reduce the participant burden) and the HUI, because based on previous work we believed that general health as opposed to a pain measure would be the better outcome. The feasibility study was designed to choose one of these for the second larger pilot RCT (6 practices, 6 pharmacists, 214 patients) which is reported in the submitted paper. We also had as secondary outcomes the Chronic Pain Grade, the WHOQOL Bref Pain and Discomfort Module (to select one of these as the pain measure), the HADS, the ICECAP-O and SF6D for use in Health Economic analyses. As a result of this feasibility work, and based on completeness and likely discrimination between participants) we selected the SF12 and the CPG as co-primary outcome measures for the second larger pilot RCT. These are as reported in our submitted paper. We excluded the WHOQOL Bref Pain and Discomfort Module. The ICECAP-O is a newly developed instrument that has not yet been validated. We were asked to include it by its authors (colleagues at the University of East Anglia) to give additional data on its use in our specific population. It was not one of our target outcomes so is not reported.

2.* Generalisability (all participants were Caucasian and most were female and over 65) That low recruitment does introduce bias.

Our response:
We acknowledge that as all participants were Caucasians, reflecting the general population in our study sites, we cannot generalise our finding to all other ethnic groups. We have added this as a limitation. The preponderance of females (overall 62% of our sample and the average age (over 65 years) reflects the population of patients with chronic pain. Most epidemiological studies report that Chronic Pain is more prevalent in women and increases with age, for example the The Lancet paper referenced in the manuscript, (Elliott AM, Smith BH, Penny KI, Smith WC, Chambers WA. The epidemiology of chronic pain in the community. Lancet 1999;354:1248-1252.) shows that women have a higher prevalence than men and that pain increases with age (highest prevalence in over 75s and second highest in 65-74s). Other commonly reported studies are Magni (Magni M, Caldieron C, Rigatti-Luchini S, Merskey H. Chronic musculo-skeletal pain and depressive symptoms in the general population. An analysis of the 1st National Health and Nutrition Examination Survey data. Pain 1990; 43: 299-307) and Verhaak (Verhaak PF, Kerssens JJ, Dekker J, Sorbi MJ, Bensing JM. Prevalence of chronic benign pain disorder among adults: a review of the literature. Pain 1989; 37: 215-222).
The following text has been added to the relevant sections of the Discussion page 12 :
'The preponderance of females (overall 62%) and average age of 65 years reflects the wider chronic pain population (1) as does the distribution of pain site (30,31)

There were, however limitations. Although high follow-up response rates were achieved at both three (86%) and six months (85%) only 25% of eligible patients entered the trial. This low initial consent rate is in line with other studies (32,33), , but may cause unknown biases including problems of generalisability, as does the solely Caucasian ethnicity.'

3.* Heterogeneity of the sample: can you compare arm pain and back pain? and pain over 10 years with pain for less than a year?

Our response:
In this pilot study we did not seek to compare the effect of pharmacist prescribing on different types or durations of pain. The reasons for collecting this data were to describe the population to show it was broadly representative of the wider population of people with chronic pain. In other studies back and limbs have been reported as the most common sites of chronic pain further confirming the

generalizability of our sample to the wider population. (M, Caldieron C, Rigatti-Luchini S, Merskey H. Chronic musculo-skeletal pain and depressive symptoms in the general population. An analysis of the 1st National Health and Nutrition Examination Survey data. Pain 1990; 43: 299-307). (Gureje O, Von Korff M, Simon GE, Gater R. Persistent pain and well-being. A World Health Organization study in primary care. Journal of the American Medical Association 1998; 280: 147-151.). Text has been added to the strengths of the paper in the Discussion section page 12 as follows:

'The preponderance of females (overall 62%) and average age of 65 years reflects the wider chronic pain population (1) as does the distribution of pain site (30,31)'

Duration of pain is not often described in other studies as it is generally accepted that once pain has become chronic, the duration doesn't really matter because the issues are the same in terms of treatment. In our study we appear to have a relatively good distribution across both shorter and longer term pain thus increasing the generalizability of the overall sample.

4. * Please be clearer about what effect size you are trying to estimate from this exploratory study which can then be used to power their larger trial.

Our response:
As this was an exploratory study, our sample size was based on several criteria. We needed to have sufficient numbers in each arm to assess variability in the outcome measures which would inform a formal sample size calculation of a full trial (Lancaster et al suggests 30 patients or greater are needed to estimate a parameter in an RCT (Lancaster GA, Dood S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract 2004;10(2):307-312). The sample needed to be large enough to include a range of practices and pharmacists to avoid undue bias (for example, the bias that would arise from only involving one or two pharmacists who might have particular expertise). Pre-specification of the effect size at the exploratory stage is only viable if previous researchers (on a similar population) have published the minimally important difference (MID) for an outcome measure. Since no MIDs were available for any of our outcome measures in this population (particularly among the non-medical prescribing arm), then we could not pre-specify any effect sizes. One of the aims of this exploratory study was to quantify variability in the outcome measures.

Reviewer 2 : Peter Watson Statistician
MRC Cognition and Brain Sciences Unit
15 Chaucer Road
Cambridge

I have no competing interests connected to the research in this paper.

1. I was confused by the rationale of the paper. The authors state that since it is an exploratory study with no previous work it is not powered for the purpose of statistical testing (page 9). In this scenario descriptive statistics and effect sizes would be appropriate. The study however features (Tables 2-5 on pages 22-25) p-values which are carried out to test apriori statistical hypotheses but giving p-values assumes (a) there are apriori hypotheses and (b) the study is adequately powered to allow differences to be found using statistical tests and preclude the 'file-drawer' problem.

Our response:
We have justified the chosen sample size of this exploratory study in our answer to comment 4 of reviewer 1. No formal power calculation was done since this was an exploratory study. However,

because there were no published MIDs available, we felt it was important to present the actual clinical magnitude of change in outcome at 6 months alongside a statistical assessment of this change in terms of a p-value. This would allow readers to assess both clinical and statistical significance simultaneously. With around 50 patients per arm, this was deemed sufficient numbers to examine the within and between group changes in outcome measures with appropriate univariate statistical tests. For example, with 50 in each arm, it is likely that the paired t-test assumption that the samples are drawn from populations with a normal distribution holds. Since non-parametric tests are generally most suitable for small samples, and don't require any assumptions about the parametric form of the distribution, using the Wilcoxon signed rank test to compare baseline to 6 month within group changes on groups of 50 patients is acceptable. With total samples sizes of around 150 for between group comparisons using ANOVA or the Kruskall Wallis test, it is more than likely that the p-values will be valid.

Given the above comments, we have added appropriate text to the paper under the limitations section as follows (page 13)

'However, because there were no published MIDs available to estimate effect size for the outcomes in this population, it was important to present the actual clinical magnitude of change in outcome at 6 months alongside a statistical assessment of this change (p-value). This allows an assessment of both clinical and statistical significance simultaneously with the caveat that this is an exploratory study. With around 50 patients per arm, this was deemed sufficient numbers to examine the change in outcome measures with appropriate within and between group univariate statistical tests'

2. I was confused by what the p-values in Tables 2 to 5 (pages 22-25) were testing. Could these also be explained clearly in the table to explicitly say what is being tested. For example the between group p-values in Tables 2 and 3 appear to test for a time by group interaction to see if there are different rates of improvement between baseline and follow-up in the various groups.

Our response:
We have added appropriate footnotes to tables 2-5 stating which test was performed.

3. I am also not clear of the reason for the 'Not Valid' p-values in Table 4 when exact tests are available and find the p-values for (I assume) the marginal homogeneity tests of the signs of change counterintuitive but this may be because it is not clear what is being tested.

Our response:
The p-value from a 3 by 3 exact test has been added to table 4. The data management section now includes an explanation of the null hypothesis for the marginal homogeneity test as follows:

'Data were entered into identical SPSS databases at each site and accuracy checks carried out on 10% before databases were merged. Descriptive statistics included means and standard deviations (SD) for normally distributed continuous data, medians (interquartile range (IQR)) for skewed continuous data and percentages (n) for categorical data. Analysis was conducted on an intention-to-treat basis for participants with complete data on relevant measures using SPSS version 18.

Exploratory analyses for parametric data included the paired t-test for within-arm comparisons of mean difference between baseline and 6 months and one-way ANOVA for between arm comparisons of mean difference. For non-parametric data it included the Wilcoxon Signed Rank test for within-arm comparisons of median difference and the Kruskal Wallis test for between arm comparisons of median difference. Categorical data was analysed using the marginal homogeneity test for within-arm comparisons (with null hypotheis that the distribution of CPG grade or HADS group does not change between baseline and 6 month follow-up) and the Chi-squared test for between arm comparisons;

analyses reported here are based on 6 month follow-up data (other than for participant experiences). Within arm effect sizes, expressed in terms of a Pearson correlation coefficient (r) have been calculated using the formulas from Rosenthal (1991) (27). Effect sizes can be directly compared using Cohen's (1988) (28) criteria of r=0.1 (small effect); r=0.3 (medium effect) and r=0.5 (large effect).'

4. The aim of this randomised control trial (page 5) is to compare responses in patients who are randomly allocated to one of three treatment groups (page 7): treated by a personalised prescription issued and followed-up by the pharmacist, a review of existing medication followed up by the GP or an 'as usual' (control group).

Questionnaires are posted (page 8) at three and six months after follow-up to compare responses with baseline. The responses (page 8) compared are the Chronic Pain grade (CPG in Table 2, page 22) disability and intensity scores, a general health index the SF-12v2 (in Table 3, page 23) with physical and mental components and the Hospital Anxiety and Depression Scale (HADS) (in Table 4 on page 24). In addition to changes in average responses over time (comparing baseline and six months) the sign of these differences is also compared across the three groups using numbers of individuals with negative, zero or positive differences.

I think the statistical tests are appropriate but find the description of the results (especially the p-values) in the tables somewhat confusing and also would like a little more focus on what the effect sizes are and at least a mention of a caveat regarding the interpretation and use of p-values in an exploratory study that has, consequently, not been powered (page 9, third paragraph). One could argue that descriptive statistics only are more meaningful in a truly exploratory analysis as there will be no apriori hypotheses to either test or power so no need for any p-values or statistical inference.

Our response:
Following on from our response to the first point, we have added a caveat to the limitations section of the discussion regarding the use and interpretation of the p-values in this exploratory study. Also please see our response to point 4 below on effect size.

5. Page 9. An aim quite correctly in any exploratory study such as this is to estimate an effect size (third paragraph). I, therefore, wondered what form(s) this effect size takes. Is this a standardised difference in means between groups or the unstandardised difference in means between groups or an eta-squared (for the one-way anova)? Would a phi coefficient be appropriate for the chi-square tests? Field (2005) also gives an easily computed effect size for nonparametric tests. On a related note do we have any idea how large a difference would need to be, irrespective of its statistical significance, to be regarded as clinically interesting? The effect size is key here as the study has not been powered which would mean any statistically non-significant results may be due to having too small a sample size to detect a clinically meaningful difference.

Our response:
The within group effect sizes have been added to tables 2-3. These have been expressed in terms of Pearson correlation coefficient using the formulas given in Rosenthal et al (Rosenthal R. Meta-analytical procedures for social research. Newbury Park, CA, Sage 1991) for the paired t-test, the Kruskall Wallis test and the marginal homogeneity test. Explanatory text has been added to the data management and analysis section. as follows (pages 9/10)

'Exploratory analyses for parametric data included the paired t-test for within-arm comparisons of mean difference between baseline and 6 months and one-way ANOVA for between arm comparisons of mean difference. For non-parametric data it included the Wilcoxon Signed Rank test for within-arm comparisons of median difference and the Kruskal Wallis test for between arm comparisons of median difference. Categorical data was analysed using the marginal homogeneity test for within-arm

comparisons (with null hypotheis that the distribution of CPG grade or HADS group does not change between baseline and 6 month follow-up) and the Chi-squared test for between arm comparisons; analyses reported here are based on 6 month follow-up data (other than for participant experiences). Within arm effect sizes, expressed in terms of a Pearson correlation coefficient (r) have been calculated using the formulas from Rosenthal (1991) (27). Effect sizes can be directly compared using Cohen's (1988) (28) criteria of r=0.1 (small effect); r=0.3 (medium effect) and r=0.5 (large effect)'.

the results section has also been modified page 11:
'In the prescribing arm, there was a statistically significant within arm improvement for the CPG intensity (p=0.002, effect size (r)=0.45) and disability (p=0.003, effect size (r)=0.43) subscales, and between arms on the intensity sub-scale (p=0.02), but not the disability subscale (p=0.55) (Table 2). There was a significant within-arm improvement in overall CPG grade in the prescribing (p=0.003) and review arm (p=0.001), but not in the TAU arm. The SF-12 Physical Component Score showed a statistically significant within arm improvement in the TAU arm (p=0.02, effect size (r)=0.35) (Table 3), but not between trial arms. The SF-12 Mental Component Score showed a statistically significant deterioration in the TAU arm (p=0.002, effect size (r)=0.45)(Table 3), as did the HADS-D (p=0.03, Table 4). Analysis was also carried out on the non-categorised HADS scores which showed a statistically significant improvement within the prescribing arm for Depression (p=0.022) and Anxiety (p=0.007). These were both significant between groups (p=0.022 and p=0.045 respectively) (Table 5)'
and a sentence has been added tot he Discussion page 14.
'The effect size of 0.45 suggests this could be an important difference.'
The CPG is a commonly used measure of pain severity and has been shown to be a valid and reliable measure for use as a self-completion postal questionnaire in a general population sample (Smith BH, Penny KI, Purves AM, Munro C, Wilson B, Grimshaw J, Chambers WA, Smith WC. The Chronic Pain Grade Questionnaire: validation and reliability in postal research. Pain 1997; 71: 141-147).
As far as we are aware there has been no work to test the clinical significance of a 1 grade move on the CPG specifically (i.e. across the four grades). The fact that there are only 4 grades suggests a move of one point would have clinical significance and such a move is hard to achieve . Original testing of the CPG was rigorous, and much of it was done on pain clinic attendees with patient samples representing sufferers of back pain, headache and temporo-mandibular disorder pain. This would point to the creation of groups that were deemed clinically significant even though it wasn't stated. In terms of a move on the sub-scale scores of disability and intensity: the following paper does state that a move of two points on an eleven point chronic pain intensity scale (which makes up 6 of the 7 CPG questions) is clinically significant (this isn't specific to the CPG) but does give us something factual to pin clinical significance onto (Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001 94: 149-158).

6. I, therefore, wonder about the usefulness of quoting p-values in Tables 2 to 5. The effect size, however, should not be influenced by sample size so should give us a good idea of any clinically large differences and should be expressed for all comparisons. If any large effect sizes are found one can then (as suggested on page 9, third paragraph) see if these effects would be replicated in a larger study which would be adequately powered to have enough people in it to detect such a difference and p-values could then be used to test for any associated hypotheses in the larger study that were suggested by this pilot study.

Our response:
Our response to points 1 and 3 address the reviewer's concern of including p-values and we have added a caveat to the limitations section of the discussion. We have now added the within group effect sizes to tables 2 and 3 which has resulted in new text under the data management and results sections.

7. Page 9. Did you consider testing for between factor (group) by within factor (time) interactions? This is not explicitly mentioned in the final (analysis) paragraph on page 9 but I think it is tested for in Table 2 by the between groups p-value (far right) testing the difference over time within group (ie a group by time interaction).

Our response:
The between group p-values are from univariate analyses of the mean or median changes from baseline to 6 months in each arm. The chi squared test was used to compare changes in CPG or HADs grade across the 3 arms. No tests of between by within factor interactions were done.

7. Pages 22 to 25. I think the p-values could be more clearly expressed in the captions and body of Tables 2 to 5. For example you could put the p-value next to the statistic it relates to e.g. difference CPG intensity=-8.0 (16.34),p=0.002 in Table 2 and the p-value for testing between groups (presumably the three differences in average CPG intensity between baseline and 6 months follow-up in Table 2) on the same row as the actual numbers it is testing (see Table 2 the first two P(between groups) p-values). I assume the between groups p-values for the counts at the bottom of Tables 2 (page 22) and 4 (page 24) are testing the group profiles of the 3x3 tables of counts for CPG grade differences (Table 2) and HADS-A so testing to see if groups improve at different rates (ie the group by time interaction).

Our response:
We have moved the between group p-values down to align with the within group changes. We have added footnotes to state what the p-values relate to.

8. I am also not clear why nonparametric tests (with medians) are used for the CPG disability and SF12 mental scales but parametric tests (with means) used for the CPG intensity and SF12 physical scales. Is one subscale in each test skewed or having a limited number of responses to motivate the use of a nonparametric test?

Our response:
The distribution of change in CPG intensity and disability scores were examined alongside their descriptive statistics. For intensity, the changes followed an approximate normal distribution and so mean (SD) changes are presented and the changes from baseline to 6 months examined using a paired t-test. However, changes in CPG disability scores were skewed and so the more appropriate median (interquartile range) summary statistics were included and the scores compared with a non-parametric Wilcoxon test.

9. Page 24. It is not clear to me what the within group p-values are comparing in Table 4. For example, I notice a p-value of 1.0 in Table 4 is given for Prescribing Difference HADS-D which I assume is testing whether the counts for the signed HADS-D difference of 5,34,5 are equal (where there seem a lot more zero differences than positive or negative ones). I therefore wondered if you were testing if the number of negative HADS-D changes was equal to the number of positive HADS-D changes which could give a p=1.00 for the prescribed group but then in the next column for the review group we have three counts of 4, 37 and 4 respectively for <=-1,0 and >=1 which has a p-value of 0.71 rather than 1.0, despite the number of positive and negative changes in the review group also being the same.

Our response:
The marginal homogeneity test was used for the within group changes in categorical factors. The null hypothesis is that there is no change in distribution of HADS category over follow-up. Effectively, in a 3 by 3 cross tabulation of HADS at baseline and HADS at 6 months, if absolutely no patient changed

HADS grade then everyone would lie on the diagonal. The marginal homogeneity test examines movement off the diagonal.

10. The “Not Valid” in Table 4 for P(between groups) suggests it is not possible to do a statistical test of the two-way frequency tabulation owing to small numbers in the cells but a Fisher exact test can still be used here (recommended by Howell (1997) when the expected count in any cell is less than 5).

Our response:
The not valid p-value has now been replaced with a p-value from an Exact test for a 3 by 3 table.

11. It might also aid interpretability if the differences in CPG (Table 2, p22) and HADS grades (Table 4, p24) was expressed as improvement (presumably <=-1) or deterioration (>=1) or no change in these tables.

Our response:
We believe the current presentation is clear

References
Field, A (2005) Discovering Statistics using SPSS. 2nd Edition. Sage:London.
Howell, DC (1997) Statistical methods for psychology. Fourth Edition. Wadsworth. Belmont,CA.

Reviewer 3: D.K. Raynor, Professor of Pharmacy Practice, University of Leeds, UK and Academic Advisor, Luto Research (www.luto.co.uk)

1. Slight rewording needed in the Abstract and Key Messages - see below

The interpretation and conclusions are in general appropriately described, but I suggest below a slight softening of the wording in the Abstract, Key Messages and Conclusion.

Our response:
Please see 2 and 3 below

2. GENERAL
This is an important paper in a field where controlled studies are few and far between.

Our response:
We thank the reviewer for this comment

3. My only significant concern is that in the Abstract and the Conclusion the choice of words could be argued to 'over-claim'. At the end of the Abstract I would suggest replicating the wording used in the Discussion later to make the wording here '...and suggests THERE MAY BE a benefit for patients....'.

Our response:
Abstract Conclusion amended as suggested
‘This is the first RCT of pharmacist-prescribing in the UK, and suggests there may be a benefit for patients with chronic pain. A larger trial is required’

4. This also applies to the first Key Message.

Our response:

Key messages amended as suggested
‘The findings suggest there may be improved pain related outcomes for patients receiving pain related care from a pharmacist prescriber’

5. Then in the main Conclusion suggest '....acceptable and MAY lead to improvements...'. Care over this wording needs to be taken, bearing in mind the acknowledged limitations in the main text(e.g. 25% of eligible patients entered trial; and not knowing how important the observed differences were to participants).

Our response:
Main conclusion amended as suggested
‘Our results suggest that pharmacist prescribing (and possibly pharmacist review alone) for patients with chronic pain is feasible, acceptable and may lead to improvements... in pain and other measures. A larger fully-powered trial is now needed to confirm these findings’

MINOR
6. Abstract:
- line 6 suggest '...with or without PHARMACIST prescribing'

Our response:
Abstract amended as suggested
‘To compare the effectiveness of pharmacist medication-review, with or without pharmacist prescribing, with standard care, for patients with chronic pain’

7- lines 32-37 I would remove reference to the remote telephone randomisation as it could create confusion regarding the nature of the interventions

Our response:
We would prefer to retain the method of randomisation in the Abstract text as it is a marker of the quality of the randomisation. We will remove of the editor advises.

8 Intro
- It would be helpful in the second para to note that NSAIDs are commonly used to treat chronic pain

Our response:
Text amended to read:
‘One study found that the most common medications involved in adverse drug reaction-related emergency admissions involved non-steroidal anti-inflammatory drugs (NSAIDs) (11) which are commonly used to manage pain.’

9 Method
- Line 31 Reference is made to an 'independent pharmacist prescriber' here and then on Line 49 'supplementary prescribing'. The relevant para in the Introduction needs some more detail on the types of 'non-medical prescribers'

Our response:

The following text has been added:
‘Pharmacists can either be qualified as supplementary prescribers, in which case they operate within an agree d clinical management plan (CMP) in partnership with the doctor and patient, or as an independent prescriber, in which case they can either prescribe completely independently or within a CMP.’

10. - Lines 33 and 34 - more detail would be helpful regarding both the pain diary and the pharmaceutical care plan - and could the templates for both be made available?

Our response:
We have inserted:
‘Copies of the pain diary and pharmaceutical care plan are available from the authors on request.’
We are also happy for these to be added as additional materials if the Editor wishes.

11. - Line 18 the wording and referencing in this last sentence of the para needs clarifying

Our response:
We are sorry but we are not clear which sentence this refers to

RESULTS
12. - Line 38 'generally conducted by phone' - would be better described numerically

Our response:
We have amended to:
‘....of which 34/37 were conducted by phone.’

DISCUSSION
13. - Line 2 suggest '...to both patients and MOST professionals'

Our response:
Amended as suggested

14. - Line 19 'assessed' not 'demonstrated'

Our response:
Amended as suggested

15. - Lines 22-24 The fact that formal training was given to the pharmacist is a strenght and could be acknowledged here

Our response:
Thank you for this suggestion. We have amended the sentence on page 12 to read:
‘Pharmacists received formal training and agreed and used a common treatment algorithm which should have increased standardisation of treatment’

16. - Lines 34 - 37 The meaning of the sentence beginning 'Rewording of participant recruitment documentation....' is not clear; how could this address these concerns?

Our response:
We have changed the sentence to to clarify the changes as follows and hope it is clearer.
‘Concerns identified by participants during the formal feedback e.g. having too many people involved in one’s care may have contributed to poor response rates and rewording of participant recruitment documentation to reassure participants of the role of the pharmacist could address this’.

18. - Line 15 need to add '...in the UK' at the end of the sentence

Our response:

We have added ‘in the UK’ into the sentence as follows;
Another evaluation of 26 patients using a medication review service provided jointly by a physiotherapist and pharmacist in the UK, reported improvement in pain control for 88% of patients (33).

19. FIGURE 1
Although many of us are familiar with the Grampian region and its good citizens, readers further afield may be confused by use of the term 'Grampian'. Suggest reword as 'Grampian region' here

Our response:
We have replaced ‘Grampian’ on the explanatory footnote to the Figure with Grampian Health Board area. We have not used the suggested Grampian Region as the Region is the local authority area. We have not replaced the text within the box of the flowchart as we feel this wold be too busy and the footnote provides sufficient explanation

20. and in the Methods also

Our response:
We now say ‘Practices in the Grampian Health Board area....’ in the Methods