PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to the Thorax but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open where it was re-reviewed and accepted for publication.

ARTICLE DETAILS

TITLE (PROVISIONAL)	THE EFFECTS OF MAINTENANCE SCHEDULES FOLLOWING PULMONARY REHABILITATION IN PATIENTS WITH CHRONIC OBSTRUCTIVE PULMONARY DISEASE: A RANDOMISED CONTROLLED TRIAL
AUTHORS	Wilson, Andrew; Browne, Paula; Olive, Sandra; Clark, Allan; Galey, Penny; Dix, Emma; Woodhouse, Helene; Robinson, Sue; Wilson, Edward; Staunton, Lindi

VERSION 1 - REVIEW

This manuscript received three reviews at Thorax but the referess declined to make their comments public.

VERSION 2 – REVIEW

REVIEWER	David Gillespie Cardiff University, Wales
REVIEW RETURNED	27-Jun-2014

GENERAL COMMENTS	This is a paper reporting an interesting randomised trial on the effects of a maintenance programme following pulmonary rehabilitation in patients with COPD. My role is to review the statistical aspects of this manuscript. However, I have several comments and queries on the general reporting of the study as well. Below I summarise my comments and queries under the section in the manuscript from which they originated. Abstract • Having two separate sentences in the abstract where the results are described as not statistically significantly different and then numerically superior may be confusing at best, and misleading at worst, to a reader. I suggest this is reworded to reflect the fact that there were small improvements in outcome, but they were not statistically significant. Strengths and limitations of this study • I suggest that the fourth bullet point is rewritten, to specify what is meant by "intensive" Introduction

	• The final three sentences belong in the methods section. The end of an introduction should be less specific. Methods • Eligibility criteria were quite broad. This was presumably for pragmatic reasons. However, with such a criteria (i.e. allowing the inclusion of non-COPD participants), I do not think the title of the paper accurately reflects the study population. This should be appropriately altered. • How similar are the PR programmes conducted in Norwich to other PR programmes across the UK? • More detail is needed on the randomisation process. What was the method of randomisation? Were any balancing/stratification factors used? • Were any process measures evaluated? To what extent were the intervention sessions delivered with fidelity? To what extent was the intervention used as an adjunct to (or instead of) standard care? • Why were statistical analyses based on an analysis of change scores (i.e. between-group differences in change from baseline scores)? The recommended approach is to use the follow-up score as the outcome and control for baseline as a covariate (i.e. analysis of covariance, or ANCOVA). For more information, see the following two articles for more details http://www.ncbi.nlm.nih.gov/pubmed/16895814 http://www.ncbi.nlm.nih.gov/pubmed/16921578 • If the CRQ was measured at baseline, 3, 6, 9 and 12 months, why did the authors not use a linear mixed model to simultaneously investigate the evolution of CRQ scores over time, between-group differences, and the interaction between treatment group and time (i.e. any between-group differential evolution of CRQ scores over time) • A lot more detail of how drop-outs were replaced using imputation is needed Results • Space should be given to briefly describing the randomised participants on their measures prior to randomisation (i.e. their baseline characteristics) • Less space should be given to describing the characteristics of participants at the start of PR, as these were not the primary focus of this study. However, a comparison should be made of the randomised participants to those who started the PR programme but were not subsequently randomised (to demonstrate how different they were and therefore how generalisable your findings are to PR patients as a whole group) • As in the abstract, the sentence describing the numerical superiority of some of the outcomes should be merged with an assessment of whether there were statistically significant between-group differences or not, so as to not confuse or mislead a reader. • Adherence to the intervention was low (approx. 52% according to the results section). Participants who adhered to the intervention were older, had less dyspnoea impairment and performed better at the ESWT at baseline. This is concerning, and a deeper exploration should be made of this. You refer to a per-protocol analysis on page 12, but do not describe the per-protocol population in your methods section. Was this population defined a priori? Per-protocol analysis is prone to selection bias. How was any potential selection biases investigated? Were any attempts made to adjust this analysis for selection bias?

	Discussion • I think a more thorough exploration of non-adherence and the impact it may have had on findings is required. An ITT analysis can only tell you the effect that offering maintenance schedules had on outcomes. A per-protocol analysis is prone to selection bias (and given how different those who adhered to intervention were on baseline characteristics, I would say that selection bias would be a significant problem here). Something like a complier average causal effect analysis may be more appropriate. • Were any assessments made of the cost of the intervention evaluated by the authors? On page 16 you suggest that other interventions are too expensive, but given that the paper does not assess the cost of their intervention, it is difficult to know what to make of this point. • I think more work should be done to demonstrate the generalisability of this study, and perhaps greater thought should be given of the population to which you are hoping to generalise. In the title, you suggest you're interested in COPD patients, but your broad inclusion criteria mean that you generalise to other groups attending PR. You then restrict your sample to those who attend at least 60% of PR sessions. In my opinion, you therefore limit your generalisability to patients attending PR who can complete more than half of PR sessions. As mentioned earlier, it would also be good to judge how similar PR programmes in Norwich are to others around the UK • The per-protocol population is defined for the first time in the discussion section, and only defined very loosely. This population should be defined in more detail in the methods section. A statement of whether the population was defined a priori or post-hoc should also be provided. Tables • Row 2 of table 1 does not look correct • Were mean (SD) appropriate for all measures? How were the measures distributed? • What are the number of participants that the data refer to in each column?

VERSION 2 – AUTHOR RESPONSE

Reviewer Name David Gillespie
Institution and Country Cardiff University, Wales
Please state any competing interests or state 'None declared': None declared

This is a paper reporting an interesting randomised trial on the effects of a maintenance programme following pulmonary rehabilitation in patients with COPD.

My role is to review the statistical aspects of this manuscript. However, I have several comments and queries on the general reporting of the study as well. Below I summarise my comments and queries under the section in the manuscript from which they originated.

Abstract
• Having two separate sentences in the abstract where the results are described as not statistically significantly different and then numerically superior may be confusing at best, and misleading at worst, to a reader. I suggest this is reworded to reflect the fact that there were small improvements in outcome, but they were not statistically significant.

Strengths and limitations of this study
• I suggest that the fourth bullet point is rewritten, to specify what is meant by “intensive”

This has been revised accordingly

Introduction
• The final three sentences belong in the methods section. The end of an introduction should be less specific.

The last two sentences have been moved to the methods section. The introduction is not specific.

Methods
• Eligibility criteria were quite broad. This was presumably for pragmatic reasons. However, with such a criteria (i.e. allowing the inclusion of non-COPD participants), I do not think the title of the paper accurately reflects the study population. This should be appropriately altered.

COPD can be considered an umbrella term for emphysema and chronic bronchitis. Including patients with a labelled diagnosis of emphysema and chronic bronchitis is appropriate.

• How similar are the PR programmes conducted in Norwich to other PR programmes across the UK?

The programmes conducted in Norwich are identical to other programmes other than our programme has one weekly supervised session rather than 2 supervised sessions per week. We have discussed the implication of this difference

• More detail is needed on the randomisation process. What was the method of randomisation? Were any balancing/stratification factors used?

No balancing/stratification factors were used

• Were any process measures evaluated? To what extent were the intervention sessions delivered with fidelity? To what extent was the intervention used as an adjunct to (or instead of) standard care?

A process evaluation was not part of the study design and was not undertaken. The number of patients attending all maintenance sessions is quoted in the results and a per-protocol analysis was undertaken comprising those patients.

• Why were statistical analyses based on an analysis of change scores (i.e. between-group differences in change from baseline scores)? The recommended approach is to use the follow-up score as the outcome and control for baseline as a covariate (i.e. analysis of covariance, or ANCOVA). For more information, see the following two articles for more details http://www.ncbi.nlm.nih.gov/pubmed/16895814 http://www.ncbi.nlm.nih.gov/pubmed/16921578

The analysis was based on the analysis of change scores, this was not used to “control” for baseline measures but was simply our choice of outcome measure. The adjusted approaches mentioned in the papers are more power under certain conditions, they are included in the “additional analysis” section

of the consort statement. We did not do these – and our analyses will still be unbiased, according to the references above, and were agreed prior to the database lock.

• If the CRQ was measured at baseline, 3, 6, 9 and 12 months, why did the authors not use a linear mixed model to simultaneously investigate the evolution of CRQ scores over time, between-group differences, and the interaction between treatment group and time (i.e. any between-group differential evolution of CRQ scores over time)

Although CRQ was measured at all these points the primary endpoint was defined as the difference between baseline and 12 months. The other time points were measured for secondary outcomes and in order to impute missing values of the outcomes.

• A lot more detail of how drop-outs were replaced using imputation is needed

As addressed above, we used Iteratively Chain Equations imputing using the values of all observed baseline and post-baseline outcome measures as well as treatment group. A total of 5 imputed datasets were constructed and the results were combined using Rubin's equation

Results
• Space should be given to briefly describing the randomised participants on their measures prior to randomisation (i.e. their baseline characteristics)

Table 1 provides data on the patients' baseline characteristics both before enrolment into the standard program and prior to randomisation.

• Less space should be given to describing the characteristics of participants at the start of PR, as these were not the primary focus of this study. However, a comparison should be made of the randomised participants to those who started the PR programme but were not subsequently randomised (to demonstrate how different they were and therefore how generalisable your findings are to PR patients as a whole group)

This information about pre-PR base-line is included in the text because it is not in a table. We have not undertaken a statistical analysis comparing those who did and did not complete the initial PR programme however the data is presented.

• As in the abstract, the sentence describing the numerical superiority of some of the outcomes should be merged with an assessment of whether there were statistically significant between-group differences or not, so as to not confuse or mislead a reader.

This has been removed

• Adherence to the intervention was low (approx. 52% according to the results section). Participants who adhered to the intervention were older, had less dyspnoea impairment and performed better at the ESWT at baseline. This is concerning, and a deeper exploration should be made of this. You refer to a per-protocol analysis on page 12, but do not describe the per-protocol population in your methods section. Was this population defined a priori? Per-protocol analysis is prone to selection bias. How was any potential selection biases investigated? Were any attempts made to adjust this analysis for selection bias?

We accept that per protocol analysis is open to selection bias and therefore we have not put much weight on this analysis in the discussion. It was undertaken to determine the maximum benefit of the intervention and other reviewers have requested this information. No attempts at dealing with

potential bias were undertaken. The results of the per protocol analysis were similar to the intention to treat analysis and do not alter the conclusion of the study.

Discussion
• I think a more thorough exploration of non-adherence and the impact it may have had on findings is required. An ITT analysis can only tell you the effect that offering maintenance schedules had on outcomes. A per-protocol analysis is prone to selection bias (and given how different those who adhered to intervention were on baseline characteristics, I would say that selection bias would be a significant problem here). Something like a complier average causal effect analysis may be more appropriate.

This is out with the scope of the study. Given the small number of patients it is probably unlikely to alter the conclusion of the study.

• Were any assessments made of the cost of the intervention evaluated by the authors? On page 16 you suggest that other interventions are too expensive, but given that the paper does not assess the cost of their intervention, it is difficult to know what to make of this point.

This analysis is ongoing and will be reported separately

• I think more work should be done to demonstrate the generalisability of this study, and perhaps greater thought should be given of the population to which you are hoping to generalise. In the title, you suggest you're interested in COPD patients, but your broad inclusion criteria mean that you generalise to other groups attending PR. You then restrict your sample to those who attend at least 60% of PR sessions. In my opinion, you therefore limit your generalisability to patients attending PR who can complete more than half of PR sessions. As mentioned earlier, it would also be good to judge how similar PR programmes in Norwich are to others around the UK

The terms emphysema and chronic bronchitis are synonymous with COPD. The reviewer has stated that the inclusion criteria are broad and therefore the results are generalisble. The aim of the study was to determine the effect of maintenance sessions on patients who have completed the initial programme and therefore non-compliance with the initial pulmonary rehabilitation is an appropriate exclusion (Eur Respir J 2002; 20: 20–29)

• The per-protocol population is defined for the first time in the discussion section, and only defined very loosely. This population should be defined in more detail in the methods section. A statement of whether the population was defined a priori or post-hoc should also be provided.

This has been included

Tables
• Row 2 of table 1 does not look correct

This refers to the number and percentage of males in each group

• Were mean (SD) appropriate for all measures? How were the measures distributed?

Some of the measures were not distributed according to the Normal distribution, but we felt – in line with the consort statement – that it was still better to provide the mean and standard deviation. The use of the t-test was justified by the central-limit theorem. It can also be justified by randomisation as the t-distribution is a good approximation to the randomisation distribution.

• What are the number of participants that the data refer to in each column?

The number of individuals to which the data refer has been added in the tables.

VERSION 3 - REVIEW

REVIEWER	David Gillespie Cardiff University, Wales
REVIEW RETURNED	24-Oct-2014

GENERAL COMMENTS	I am satisfied that most of my comments have been adequately addressed. The final point that requires further consideration is the per-protocol analysis. In the results section the authors write: "The results of the per protocol (PP) analysis were in keeping with the intention to treat (ITT) analysis except there was a significantly greater MET-minutes per week in the intervention group but more exacerbations and admissions." I think the findings of the PP analysis at the very least need displaying in a table. You describe characteristics of those patients who completed all maintenance sessions. It would make sense to contrast this against those who did not. This would go some way to assessing any selection bias. If you are willing to believe the findings from your PP analysis (and why would you perform such an analysis if you would be unwilling to believe it?), then it seems as though patients who received the intervention had increased activity, but this may have come at a cost (an increase in exacerbations and an increase in admissions). While I agree that the overall conclusion of the paper is not altered by the PP analysis (i.e. you do not recommend that your maintenance programme is adopted), I think there is some suggestion from the PP analysis that the intervention itself might even be harmful.

REVIEWER	Marla Beauchamp Spaulding Rehabilitation, Department of PM&R, Harvard Medical School Cambridge, MA, USA
REVIEW RETURNED	17-Oct-2014

GENERAL COMMENTS	The authors have improved their manuscript and the changes address the majority of my prior comments. In particular, the revised rationale for undertaking the trial and further description of the methods is very helpful. I have two additional comments related to the ESWT for the authors to consider: 1. There is recent study by Borel et al. published in ERJ (September 2014) that defines the MCID for the ESWT in patients with COPD as 56-61 sec and 70-82 metres. While I recognize it was not based on a trial of PR, I believe these results are in line with prior work on

	responsiveness of the ESWT and would be a reasonable starting point. In re-considering the results of this study in light of these new MCID estimates, it would appear that there was a clinically important difference between the intervention and control groups at baseline- the authors state in the discussion on page 20 that baseline values were corrected for, but it is not clear how this was done if an ANCOVA was not performed. 2. In addition, looking at the results in Table 2, it would appear that while there were no statistically significant differences between groups, the difference of 109 m between the groups might be clinically important in favour of the maintenance group. I would suggest that this point is worthy of mention in the discussion. As a minor comment, there are some minor typos/grammatical errors in the abstract and introduction that should be addressed.

VERSION 3 – AUTHOR RESPONSE

Reviewer Name Marla Beauchamp
Institution and Country Spaulding Rehabilitation, Department of PM&R, Harvard Medical School Cambridge, MA, USA
Please state any competing interests or state 'None declared': None declared.

The authors have improved their manuscript and the changes address the majority of my prior comments. In particular, the revised rationale for undertaking the trial and further description of the methods is very helpful. I have two additional comments related to the ESWT for the authors to consider:

1. There is recent study by Borel et al. published in ERJ (September 2014) that defines the MCID for the ESWT in patients with COPD as 56-61 sec and 70-82 metres. While I recognize it was not based on a trial of PR, I believe these results are in line with prior work on responsiveness of the ESWT and would be a reasonable starting point. In re-considering the results of this study in light of these new MCID estimates, it would appear that there was a clinically important difference between the intervention and control groups at baseline- the authors state in the discussion on page 20 that baseline values were corrected for, but it is not clear how this was done if an ANCOVA was not performed.

OUR OUTCOME MEASURE WAS THE CHANGE FROM BASELINE, HENCE ALTHOUGH THE ACTUAL VALUES OF SOME OF THE VARIABLE APPEAR DIFFERENT AT BASELINE BETWEEN THE GROUPS BECAUSE WE MEASURED THE CHANGE FROM BASELINE THE TWO GROUPS HAVE THE SAME VALUE OF THE OUTCOME AT BASELINE (I.E. ZERO).

2. In addition, looking at the results in Table 2, it would appear that while there were no statistically significant differences between groups, the difference of 109 m between the groups might be clinically important in favour of the maintenance group. I would suggest that this point is worthy of mention in the discussion.

WE HAVE ADDED A STATEMENT TO MENTION THAT THE CHANGE IN ESWT DISTANCE MAY HAVE BEEN CLNICALLY SIGNIFICANT

As a minor comment, there are some minor typos/grammatical errors in the abstract and introduction that should be addressed.

WE HAVE REVIEWED THIS FOR TYPOGRAPHICAL ERRORS

Reviewer Name David Gillespie
Institution and Country Cardiff University, Wales
Please state any competing interests or state ‘None declared’: None declared

I am satisfied that most of my comments have been adequately addressed.

The final point that requires further consideration is the per-protocol analysis. In the results section the authors write:

“The results of the per protocol (PP) analysis were in keeping with the intention to treat (ITT) analysis except there was a significantly greater MET-minutes per week in the intervention group but more exacerbations and admissions.”

I think the findings of the PP analysis at the very least need displaying in a table. You describe characteristics of those patients who completed all maintenance sessions. It would make sense to contrast this against those who did not. This would go some way to assessing any selection bias.

WE HAVE INCLUDED THE PER PROTOCOL DATA IN TABLE 2

If you are willing to believe the findings from your PP analysis (and why would you perform such an analysis if you would be unwilling to believe it?), then it seems as though patients who received the intervention had increased activity, but this may have come at a cost (an increase in exacerbations and an increase in admissions).

While I agree that the overall conclusion of the paper is not altered by the PP analysis (i.e. you do not recommend that your maintenance programme is adopted), I think there is some suggestion from the PP analysis that the intervention itself might even be harmful.

WE HAVE ADDED A STATEMENT THAT THE INTERVENTION MAY CAUSE HARM AS ADVISED BY THE REVIEWER. HOWEVER WE HAVE ALSO ADDED A STATEMENT TO EXPLAIN WHY THE RESULTS SHOULD BE INTERPRETED WITH CAUTION.