PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Telehealthcare for patients suffering from chronic obstructive pulmonary disease: Effects on health-related quality of life - Results from the Danish "TeleCare North" cluster-randomised trial
AUTHORS	Lilholt, Pernille; Witt Udsen, Flemming; Ehlers, Lars; Hejlesen, Ole

VERSION 1 - REVIEW

REVIEWER	Professor Stephen Walters University of Sheffield
REVIEW RETURNED	18-Nov-2016

GENERAL COMMENTS	Since this is a cluster randomised trial (CRCT) the abstract needs to report the number of clusters randomised per group as well as the number of participants in each group. The abstract also needs to report the number of clusters analysed and the number of participants in each group with valid primary outcome data (SF-36 PCS or MCS scores). My reading of the paper is that there was considerable attrition in the sample by 12 months: at the start there were 1225 participants and by 12 month follow-up there were only 258 + 316 = 574 participants with valid primary outcome data i.e. 574/1225 = 47%. There was 53% attrition in the sample with over half of participants not providing follow-up data. This is an important and significant limitation of the study. This attrition rate should also be mentioned in the main body of the text in the results. Finally, I think the abstract should report the observed mean changes for PCS and MCS scores over time for each group; this should be based on the 574 with valid baseline and 12 month follow-up data. The CONSORT flow diagram, in its present format is misleading, and should be changed to reflect that only 574 participants had valid primary outcome data. From the CONSORT flow diagram it appears that 50 +53 = 103 COPD patients died during the 12 month follow-up. It does not make sense and in my opinion is not appropriate to impute missing PCS and MCS scores for these trial participants. Recruitment bias is a major problem in cRCT designs. In cRCTs, the comparability of groups is challenged because groups of trial

	participants rather than the participants themselves are randomised. The timelines for running cluster randomised controlled trials usually make allocation concealment difficult or impossible. This is because clusters are recruited and randomised and then participants are recruited. This can potentially introduce recruitment of participants with different characteristics to the clusters which can then lead to a quantitative and qualitative imbalance the between groups. The applicants should justify their strategy and approach to preventing biased recruitment at the clusters/centre; since they will be recruiting new or incident COPD patients at the clusters; and the staff at the centres will already know their treatment allocation. Three ways to minimise recruitment bias: 1) If possible, individual random allocation to the groups should be used. 2) If cluster random allocation is required, then ideally participants should be identified before random allocation of the clusters. 3) If prior identification of participants is not possible, before the randomisation of clusters, then an independent recruiter should be used to recruit participants. The primary analysis should be based on the observed outcome data seen in the trial; any other analyses should be treated as secondary or as part of a sensitivity analysis. Imputation of missing data, although a valid statistical strategy, is effectively just “making up“ the missing data. The applicants should justify why they are using the change (CHANGE) in SF-36 PCS and MCS score from randomisation to follow-up; rather than follow-up or post-treatment score (POST) adjusted for baseline score? Since you have an RCT design then randomisation of the participants should ensure that all have similar baseline levels of the outcome. In these circumstances it is well known (Frison and Pocock, 1992) that the most powerful method of statistical analysis of the outcome data is ANCOVA (i.e. post-treatment score adjusted for baseline score and treatment group) rather than analysis of change from baseline scores (CHANGE) or comparison of post-treatment means (POST). Frison L. & Pocock S.J. Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and Its Implications for Design. Statistics in Medicine 1992; 11: 1685-1704. Similarly there is considerable debate in the statistical literature about whether or is appropriate to include the baseline score as covariate when testing the effect of an independent variable (e.g. randomised treatment group) on change scores. The authors need to justify their analytical approach. The high 53% attrition rate of participants with valid primary outcome data (PCS and MCS scores) should be mentioned as limitation of the study in the strengths and limitations section; as well as in the discussion section. Did the authors use norm based scoring (NBS) for the SF-36 rather than 0 to 100 scale scoring as stated in the text? I think they used the former, NBS, but they need to be clear. With NBS scoring the scores as standardised to have a mean of 50 and SD of 10 the same as the reference population. What reference population was used for the NBS? Was it a Danish reference population?

	Not enough information is given in the sample size section to replicate the sample size calculation. We need to know the assumed SD for the outcome as well as the average cluster size. Did the sample size use a one-sided or two-sided significance level for alpha? Where did the ICC estimate of 0.05 that was used in the sample size calculation come from? Where did the estimate of 5 points for the MID used in the sample size calculation come from? Does the study have one primary outcome the PCS scale of the SF-36 or two both the PCS and MCS scales of the SF-36? Are these co-primary outcomes? Although given the results it does not really matter; how would the authors have dealt with the issue of co-primary outcomes? What if one outcome (say PCS) showed a significant result and the other (MCS) did not? How did they propose to deal with the co-primary outcomes? The data is sometimes presented with spurious numerical precision which adds no value to the thesis and even detracts from its readability and credibility. For example the PCS and MCS outcomes are standardised to have a mean score of 50 and a SD 10 the same as the reference population. For example, in Table 2 it not necessary to report mean MCS and PCS scores (and mean differences) to four decimal places; one decimal place is probably sufficient. Also for the various analyses it is not clear how many of the 1225 patients randomised or the 574 followed-up were included in the analysis. In the text the mean differences and confidence intervals are quoted to a precision of 7 decimal places! The primary analysis should be the based on the observed PCS and MCS scores (and changes at 12 months). The authors should replace table 2 with another table which includes the results for the complete case analysis of n=574 as well as the imputed analysis the sample size both overall and by group used in each of these analyses should be clearly stated. It would be good to see the mean (SD) PCS and MCS scores at baseline; and the mean (SD) PCS and MCS scores at 12 months follow-up as well as the change over time in the difference between the groups in the change scores as well as the associated 95% confidence interval for the difference. When imputing missing data in RCTS there is considerable debate on whether or not to include the randomised treatment group as a factor in the imputation model. The authors need to justify why they imputed missing data separately for each group and the affect this may have had on the results and conclusions. Also since this was a cluster RCT did the imputation model include the cluster or allow for the correlation of missing data within a cluster? The manuscript needs to make explicit the amount of missing outcome data. Table 4 reports the results of a series of sub-group analyses and comparisons. These were not pre-specified apriori in the attached protocol. To test for sub-group effect you need to fit a statistical model with a randomised group x sub-group interaction term, as well as terms for

	the randomised group and sub-group; and report the interaction term and its associated confidence interval.

REVIEWER	Brian Mckinstry University of Edinburgh I conduct research in Telehealth
REVIEW RETURNED	21-Nov-2016

GENERAL COMMENTS	Thank you for asking me to review this interesting paper which describes a cluster RCT of the impact on health related quality of life of life of a telehealth intervention for COPD. The study is well designed but suffered from a very high drop-out rate (much higher than is usual for this type of intervention) particularly among the intervention group. A list of reasons are given but no sense of which of these were the most important. The paper and protocol make no mention of the impact on workload of the intervention. Given that there appeared to be no impact on QOL It would be important to establish if the intervention took up less resource. Is this going to be the subject of another paper? Some detail on this would really help the reader Likewise, it is not clear if a qualitative process evaluation to shed light on the reasons for the failure of the intervention (and the drop out rate) had taken place. While I realise that also this may be the topic of another paper even some broad brush results would be helpful. The authors question the sensitivity of their measure of QOL and I agree with them. There is evidence that people with COPD become accustomed to the limitations of their illness and do not register its impact on QOL. Impact on hospitalisation and other clinical resource use may have been more sensitive measures. Nonetheless the results are in keeping both with large individually randomised RCTs such as Telescot and another cluster RCT (the whole system demonstrator) and provide yet more evidence for the ineffectiveness of telehealth for COPD.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

Comment no. 1: Since this is a cluster randomised trial (CRCT) the abstract needs to report the number of clusters randomised per group as well as the number of participants in each group.

Response to comment no. 1: Thank you for your suggestions. We have added information to the abstract about the number of clusters randomized per groups and the number of patients in each group. In total, there were 26 municipality districts with 13 municipality districts in each treatment arm (intervention and control arm). The intervention group consisted of 578 patients and the control group consisted of 647 patients.

Comment no. 2: The abstract also needs to report the number of clusters analysed and the number of participants in each group with valid primary outcome data (SF-36 PCS or MCS scores).

Response to comment no. 2: All 26 municipality districts were analysed and the number of participants in each group, which represents complete cases was (intervention: 258 and control: 316). This information has been added to the abstract.

Comment no. 3: My reading of the paper is that there was considerable attrition in the sample by 12 months: at the start there were 1225 participants and by 12 month follow-up there were only 258 + 316 = 574 participants with valid primary outcome data i.e. 574/1225 = 47%. There was 53% attrition in the sample with over half of participants not providing follow-up data. This is an important and significant limitation of the study. This attrition rate should also be mentioned in the main body of the text in the results.

Response to comment no. 3: We thank the reviewer for drawing attention to the attrition rate. To comment on this, the 53% attrition in the sample represents not only lost-to-follow-up patients but also incomplete cases (cases that are not lost-to-follow-up but have missing values on items in either PCS and MCS at baseline or follow-up). PCS and MCS scores could only be calculated on those patients who have responded to all items in the questionnaire. However, the attrition rate has been added to the results section, strengths and limitations section and the discussion section.

Comment no. 4: Finally, I think the abstract should report the observed mean changes for PCS and MCS scores over time for each group; this should be based on the 574 with valid baseline and 12 month follow-up data.

Response to comment no. 4: The idea purposed by the reviewer is very relevant, and we agree that reporting mean differences for PCS and MCS scores can be useful and have added them to the abstract and Table 2. But we strongly disagree that primary results should be based on a complete case analysis. See later comment.

Comment no. 5: The CONSORT flow diagram, in its present format is misleading, and should be changed to reflect that only 574 participants had valid primary outcome data.

Response to comment no. 5: We agree that more details could be included in the CONSORT diagram. Figure 2 has been changed to allow the reader to see the sensitivity analysis that has been conducted on the data with complete case data, n =574 and imputed data, n=1,225. Figure 2 is also commented in the results section. Hope that it is more informative now.

Comment no. 6: From the CONSORT flow diagram it appears that 50 +53 = 103 COPD patients died during the 12 month follow-up. It does not make sense and in my opinion is not appropriate to impute missing PCS and MCS scores for these trial participants.

Response to comment no. 6: This must be a misunderstanding. We did not impute missing PCS and MCS scores for the patients that died during the trial period. To account for death, we transformed the scores from people that died into values of 0. This information has been added to the statistical analysis section. Coding deaths as 0 has been used as a strategy in:

Ware J, Bayliss M, Rogers W, et al. Differences in 4-Year Health Outcomes for Elderly and Poor, Chronically III Patients Treated in HMO and Fee-for-Service Systems: Results from the Medical Outcomes Study. J Am Med Assoc 1996;276:1039–47. doi:doi:10.1001/jama.1996.03540130037027

Ware et al. had unchanged conclusions by alternative analyses where death was assigned the lowest observed score and by analyses where death was assigned the value two standard deviations below the minimum observed score value.

Comment no. 7: Recruitment bias is a major problem in cRCT designs. In cRCTs, the comparability of groups is challenged because groups of trial participants rather than the participants themselves are randomised. The timelines for running cluster randomised controlled trials usually make allocation

concealment difficult or impossible. This is because clusters are recruited and randomised and then participants are recruited. This can potentially introduce recruitment of participants with different characteristics to the clusters which can then lead to a quantitative and qualitative imbalance the between groups.

The applicants should justify their strategy and approach to preventing biased recruitment at the clusters/centre; since they will be recruiting new or incident COPD patients at the clusters; and the staff at the centres will already know their treatment allocation.

Three ways to minimise recruitment bias:
1) If possible, individual random allocation to the groups should be used.
2) If cluster random allocation is required, then ideally participants should be identified before random allocation of the clusters.
3) If prior identification of participants is not possible, before the randomisation of clusters, then an independent recruiter should be used to recruit participants.

Response to comment no. 7: We agree that we should clarify the strategy and approach to prevent biased recruitment. As the reviewer has mentioned, there are ways to minimize recruitment bias and to respond on that, we have already used no. 2 of his points as strategy:

“If cluster random allocation is required, then ideally participants should be identified before random allocation of the clusters”.

To be specific - the subjects from the trial were first identified and thereafter were the allocation of municipality districts performed (this information has been added to section 2.2 Participants. We have also mentioned our strategy as a strength in the strengths and limitations section.

Explanation: Subjects were recruited between May-November 2013 and after this date, it was not possible for subjects to participate in the trial - new subjects were not continuously included in the trial. The randomization was performed by an external person who was not affiliated with the TeleCare North trial. In contrast to the WSD study by Cartwright et al 2013, the randomization of subjects was not at the general practice level but rather at the municipality district level. By doing that, it was possible to prevent healthcare providers in the same municipality district from providing care for patients in both the intervention group and control group. The randomization divided the subjects into two comparable groups, in which no differences in case distribution were found.

Cartwright, M., Hirani, S. P., Rixon, L., Beynon, M., Doll, H., Bower, P., … Newman, S. P. (2013). Effect of telehealth on quality of life and psychological outcomes over 12 months (Whole Systems Demonstrator telehealth questionnaire study): Nested study of patient reported outcomes in a pragmatic, cluster randomised controlled trial. BMJ (Online), 346(7897)

Comment no. 8: The primary analysis should be based on the observed outcome data seen in the trial; any other analyses should be treated as secondary or as part of a sensitivity analysis. Imputation of missing data, although a valid statistical strategy, is effectively just “making up“ the missing data.

Response to comment no. 8: We disagree. What scientific grounds is the reviewer referring to? For a complete case analysis to be unbiased, strict assumptions has to be taken and several antecedents have to be in place, see e.g. Sterne et. al. 2009. Missing data must only occur in an outcome variable that is measured once in each individual, provided that all variables associated with the outcome being missing can be included as covariates (under a missing at random assumption). Missing data in predictor variables also do not cause bias in analyses of complete cases if the reasons for the missing data are unrelated to the outcome, which they rarely are (it is reasonable that drop-out are associated

with e.g. reduced health-related quality of life). And only when it is plausible that data are missing at random, but not completely at random, analyses based on complete cases may be biased.

By only assuming that outcomes are missing at random (an assumption that can be increased by e.g. including a lot of relevant covariates in explaining missingness), much more powerful analyses can be made by multiple imputation, which is also the stated analysis strategy for handling missing data in the trial protocol (Udsen et. al. 2014).

Presenting primary results based on multiple imputation is also in accordance with other research published in BMJ on the topic, e.g. the Whole System Demonstrator QoL results, see Cartwright et. al 2013.
Having said that, we agree that although results can be biased, it could be informative to conduct a complete case analysis for comparison to check the robustness of conclusions and have included a complete case analysis as a sensitivity analysis in Table 3.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, 339(July), 157–160. http://doi.org/10.1136/bmj.b2393

Udsen, F., Lilholt, P., Hejlesen, O., & Ehlers, L. (2014). Effectiveness and cost-effectiveness of telehealthcare for chronic obstructive pulmonary disease: study protocol for the Danish “TeleCare North” pragmatic cluster-randomized trial. Trials, 15(178)

Cartwright, M., Hirani, S. P., Rixon, L., Beynon, M., Doll, H., Bower, P., … Newman, S. P. (2013). Effect of telehealth on quality of life and psychological outcomes over 12 months (Whole Systems Demonstrator telehealth questionnaire study): Nested study of patient reported outcomes in a pragmatic, cluster randomised controlled trial. BMJ (Online), 346(7897)

Comment no. 9: The applicants should justify why they are using the change (CHANGE) in SF-36 PCS and MCS score from randomisation to follow-up; rather than follow-up or post-treatment score (POST) adjusted for baseline score? Since you have an RCT design then randomisation of the participants should ensure that all have similar baseline levels of the outcome. In these circumstances it is well known (Frison and Pocock, 1992) that the most powerful method of statistical analysis of the outcome data is ANCOVA (i.e. post-treatment score adjusted for baseline score and treatment group) rather than analysis of change from baseline scores (CHANGE) or comparison of post-treatment means (POST). Frison L. & Pocock S.J. Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and Its Implications for Design. Statistics in Medicine 1992; 11: 1685-1704.

Response to comment no. 9: We agree. Thank you. The analysis strategy has been changed to reflect an ANCOVA approach as described by Vickers et al. 2001 and results updated accordingly. Results are the same.

Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. BMJ 2001;323:1123–4. doi:10.1136/bmj.323.7321.1123

Comment no. 10: Similarly there is considerable debate in the statistical literature about whether or is appropriate to include the baseline score as covariate when testing the effect of an independent variable (e.g. randomised treatment group) on change scores. The authors need to justify their analytical approach.

Response to comment no. 10: We agree. Since the analysis strategy has been changed, this comment is no longer relevant.

Comment no. 11: The high 53% attrition rate of participants with valid primary outcome data (PCS and MCS scores) should be mentioned as limitation of the study in the strengths and limitations section; as well as in the discussion section.

Response to comment no. 11: The idea proposed by the reviewer about mentioning the high attrition rate in the strength and limitation section and in the discussion section has been met.

Comment no. 12: Did the authors use norm based scoring (NBS) for the SF-36 rather than 0 to 100 scale scoring as stated in the text? I think they used the former, NBS, but they need to be clear. With NBS scoring the scores as standardised to have a mean of 50 and SD of 10 the same as the reference population. What reference population was used for the NBS? Was it a Danish reference population?

Response to comment no. 12: This is certainly some good questions. We have used the QualityMetric Health Outcomes Scoring Software to convert all scores to a single metric. The standardized scores were transformed to the T-score Based scoring (NBS) metric which is based on 2009 US general population norms.

The US general population norms are comparable with the Danish population norms. Studies have demonstrated that the variation in SF-36 scales in the Danish population are comparable with the US population (Danish data has slightly lower variance but also greater disparity than the US), (Bjørner et al.). Information about the scoring procedure has been added to the statistical analysis section and all tables.

Bjørner JB, Damsgaard MT, Watt T, Bech P, Rasmussen NK, Kristensen TS, Modvig J, T. K. el al. (1997). Dansk manual til SF-36 : et spørgeskema om helbredsstatus.

Comment no. 13: Not enough information is given in the sample size section to replicate the sample size calculation. We need to know the assumed SD for the outcome as well as the average cluster size.
Did the sample size use a one-sided or two-sided significance level for alpha?

Where did the ICC estimate of 0.05 that was used in the sample size calculation come from? Where did the estimate of 5 points for the MID used in the sample size calculation come from?

Response to comment no. 13: This is certainly a good observation. The assumed SD for the outcome was 10 and the average cluster size was hypothesized to be 50 with a coefficient of variation of 0.5. The sample size used a two-sided significance level for alpha. The ICC estimate of 0.05 and 5 point for the MCID used in the sample size calculation was based on a recent Norwegian study by Bentsen et al 2013. The lack of information to replicate the sample size have been added to the manuscript.

Bentsen SB, Rokne B, Wahl AK. Comparison of health-related quality of life between patients with chronic obstructive pulmonary disease and the general population. Scand J Caring Sci 2013;27:905–12. doi:10.1111/scs.12002

Comment no. 14: Does the study have one primary outcome the PCS scale of the SF-36 or two both the PCS and MCS scales of the SF-36? Are these co-primary outcomes? Although given the results it does not really matter; how would the authors have dealt with the issue of co-primary outcomes? What if one outcome (say PCS) showed a significant result and the other (MCS) did not? How did they propose to deal with the co-primary outcomes?

Response to comment no. 14: We thank the reviewer for the question about co-primary outcomes. This information is addressed in the study protocol (Udsen et al. 2014). The sample size calculation is based on the study protocol's primary outcome, PCS. We have only one primary outcome for effectiveness: PCS scale of the SF-36 and this is also the reason why we have not discussed how to deal with co-primary outcomes if one of the outcomes (PCS or MCS) were significant, and the other was insignificant.

Udsen, F., Lilholt, P., Hejlesen, O., & Ehlers, L. (2014). Effectiveness and cost-effectiveness of telehealthcare for chronic obstructive pulmonary disease: study protocol for the Danish "TeleCare North" pragmatic cluster-randomized trial. Trials, 15(178)

Comment no. 15: The data is sometimes presented with spurious numerical precision which adds no value to the thesis and even detracts from its readability and credibility. For example the PCS and MCS outcomes are standardised to have a mean score of 50 and a SD 10 the same as the reference population. For example, in Table 2 it not necessary to report mean MCS and PCS scores (and mean differences) to four decimal places; one decimal place is probably sufficient. Also for the various analyses it is not clear how many of the 1225 patients randomised or the 574 followed-up were included in the analysis. In the text the mean differences and confidence intervals are quoted to a precision of 7 decimal places!

Response to comment no. 15: This is certainly a good observation. The decimals have been removed from the manuscript and tables and only two decimals are now represented in the manuscript.

Comment no. 16: The primary analysis should be the based on the observed PCS and MCS scores (and changes at 12 months). The authors should replace table 2 with another table which includes the results for the complete case analysis of n=574 as well as the imputed analysis the sample size both overall and by group used in each of these analyses should be clearly stated. It would be good to see the mean (SD) PCS and MCS scores at baseline; and the mean (SD) PCS and MCS scores at 12 months follow-up as well as the change over time in the difference between the groups in the change scores as well as the associated 95% confidence interval for the difference.

Response to comment no. 16: We agree. We have updated Table 2 (in the revised manus named table 3) to reflect the primary analysis (based on multiple imputation) as well as a sensitivity analysis based on complete cases. Baseline difference between groups is misplaced here but can be found in Table 1 (new table named table 2 in the manus). The rest of the comments are mostly referring to the analysis strategy that we have changed now.

Comment no. 17: When imputing missing data in RCTS there is considerable debate on whether or not to include the randomised treatment group as a factor in the imputation model. The authors need to justify why they imputed missing data separately for each group and the affect this may have had on the results and conclusions.

Response to comment no. 17: We have followed good practice imputation procedures published by Faria et. al. which states the following: "The imputation should be implemented separately by randomized treatment allocation. This explicitly recognizes in the imputation model that imputations are different between treatment groups, hence that the posterior distribution of the missing data given the observed may be different between treatment groups. Imputing the treatment groups together but including all possible interactions would only recognize differential means by treatment group and not a differential covariance structure" (p.1161).
The reference and main argument from above has been added to the manuscript.

Faria, R., Gomes, M., Epstein, D., & White, I. R. (2014). A guide to handling missing data in cost-

effectiveness analysis conducted within randomised controlled trials. PharmacoEconomics, 32(12), 1157–70. http://doi.org/10.1007/s40273-014-0193-3

Comment no. 18: Also since this was a cluster RCT did the imputation model include the cluster or allow for the correlation of missing data within a cluster?

Response to comment no. 18: Yes. The imputation models included the cluster-variable which have been added to the manuscript (we forgot) along with more details on the imputation model.

Comment no. 19: The manuscript needs to make explicit the amount of missing outcome data.

Response to comment no. 19: We appreciate the reviewer's comment and have added information to the results section about the amount of missing outcome data.

Comment no. 20: Table 4 reports the results of a series of sub-group analyses and comparisons. These were not pre-specified apriori in the attached protocol.

To test for sub-group effect you need to fit a statistical model with a randomised group x sub-group interaction term, as well as terms for the randomised group and sub-group; and report the interaction term and its associated confidence interval.

Response to comment no. 20: We agree with the reviewer that the subgroups analysis was not clarified in the study protocol. In the results section, we have made it more clear that we have performed a posteriori-defined subgroup analyses. We have also added this as a limitation in the discussion section.

We agree with the analysis strategy, which has also been done in the original manuscript. However, the description of the procedure was missing from the description of the analysis, which has been added now (paragraph 2.7).

Reviewer 2:

Comment no. 1: Thank you for asking me to review this interesting paper which describes a cluster RCT of the impact on health related quality of life of life of a telehealth intervention for COPD. The study is well designed but suffered from a very high drop-out rate (much higher than is usual for this type of intervention) particularly among the intervention group. A list of reasons are given but no sense of which of these were the most important.

Response to comment no. 1: We appreciate the question raised by the reviewer. Unfortunately, we have not explored which of the reasons that are most important, and we can therefore not answer his question.

Comment no. 2: The paper and protocol make no mention of the impact on workload of the intervention. Given that there appeared to be no impact on QOL It would be important to establish if the intervention took up less resource. Is this going to be the subject of another paper? Some detail on this would really help the reader. Likewise, it is not clear if a qualitative process evaluation to shed light on the reasons for the failure of the intervention (and the drop out rate) had taken place. While I realise that also this may be the topic of another paper even some broad brush results would be helpful.

Response to comment no. 2: The reviewer's questions are very relevant and these questions will be answered through other papers of the trial in the future. Unfortunately, we cannot report the results of

these topics in this paper. However, the described research will later be published – none of the results are available yet.

Comment no. 3: The authors question the sensitivity of their measure of QOL and I agree with them. There is evidence that people with COPD become accustomed to the limitations of their illness and do not register its impact on QOL. Impact on hospitalisation and other clinical resource use may have been more sensitive measures.
Nonetheless the results are in keeping both with large individually randomised RCTs such as Telescot and another cluster RCT (the whole system demonstrator) and provide yet more evidence for the ineffectiveness of telehealth for COPD.

Response to comment no. 3: We thank the reviewer for his reflections about the sensitivity of the HRQoL measure that we have discussed in the discussion section. We also appreciate that he agrees that the trial provides further evidence for the ineffectiveness of telehealth for COPD.

VERSION 2 – REVIEW

REVIEWER	Professor Stephen Walters University of Sheffield
REVIEW RETURNED	22-Dec-2016

GENERAL COMMENTS	The authors have responded satisfactorily to the majority of my earlier comments. The manuscript is much improved. Minor comments For the abstract in the results section the authors could clarify whether the means and confidence intervals are based on the complete case sample or the larger imputed sample by including the sample used in the analysis i.e. N=574 etc. Again for the abstract ..... The data is sometimes presented with spurious numerical precision which adds no value to the manuscript and even detracts from its readability and credibility. For example the PCS and MCS outcomes are standardised to have a mean score of 50 and a SD 10 the same as the reference population. For example, in the abstract and text and tables it not necessary to report mean MCS and PCS scores (and mean differences) to two decimal places; one decimal place is probably sufficient. In the strengths list of bullet points make clear the attrition rate is 53% - saves the reader having to calculate the percentage. The authors have decided to give a PCS and MCS score of 0 to the 103 COPD patients who died during the 12 month follow-up. This does not make sense and in my opinion is not appropriate to impute missing PCS and MCS scores for these trial participants. Can the authors provide a citation from the SF-36 v2 scoring manual or elsewhere to say that this is appropriate for the SF-36?

	For example, in Table 2 it not necessary to report mean MCS and PCS scores (and mean differences) to two decimal places; one decimal place is probably sufficient.

REVIEWER	Brian McKinstry University of Edinburgh United Kingdom
REVIEW RETURNED	22-Dec-2016

GENERAL COMMENTS	Thank you. The changes you have made have greatly strengthened the paper. This was a huge achievement although I realise a frustrating one for you.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Comment no. 1: For the abstract in the results section the authors could clarify whether the means and confidence intervals are based on the complete case sample or the larger imputed sample by including the sample used in the analysis i.e. N=574 etc.

Response to comment no. 1: The idea purposed by the reviewer is very relevant. We have in the abstract clarified that means and SD are based on the imputed data (n=1,225).

Comment no. 2: Again for the abstract .....
The data is sometimes presented with spurious numerical precision which adds no value to the manuscript and even detracts from its readability and credibility. For example the PCS and MCS outcomes are standardised to have a mean score of 50 and a SD 10 the same as the reference population. For example, in the abstract and text and tables it not necessary to report mean MCS and PCS scores (and mean differences) to two decimal places; one decimal place is probably sufficient.

Response to comment no 2: We thank the reviewer for the comment about numerical precision. We have now presented data with one decimal in the manuscript (abstract, text and tables). However, we do not understand the reviewer's argument for why the data only should be presented with one decimal?

Comment no. 3: In the strengths list of bullet points make clear the attrition rate is 53% - saves the reader having to calculate the percentage.

Response to comment no. 3: We appreciate the reviewer's suggestion. We have added the 53% attrition rate to the bullet point in the "strengths and limitations section".

Comment no. 4: The authors have decided to give a PCS and MCS score of 0 to the 103 COPD patients
who died during the 12 month follow-up. This does not make sense and in my opinion is not appropriate to impute missing PCS and MCS scores for these trial participants.

Response to comment no. 4: Thank you for your comment. This must be a misunderstanding. We have not imputed missing PCS and MCS scores for patients (n=103) that died during the trial. Patients were given a value of 0. We have added a sentence in the statistical analysis section, which clarify that we have performed single imputation by replacing missing values with 0 for patients that

died during the trial period.

Comment no. 5: Can the authors provide a citation from the SF-36 v2 scoring manual or elsewhere to say that this is appropriate for the SF-36?

Response to comment no. 5: The Danish SF-36 manual suggests replacing summary scores with 0 for deceased patients

See

Bjørner JB, Damsgaard MT, Watt T, Bech P, Rasmussen NK, Kristensen TS, Modvig J, T. K. el al. (1997). Dansk manual til SF-36 : et spørgeskema om helbredsstatus.

This strategy is also consistent with the work of Diehr and colleagues that suggest coding patients with 0 in summary scores as one strategy for deceased patients along with other strategies (citations below):

Diehr P, Patrick D, Hedrick S, Rothman M, Grembrowski S, Raghunathan TE, B. S. (1995). Including Deaths When Measuring Health Status Over Time. Med Care, 33, 164–172.

Diehr P, Patrick DL, Spertus J, Kiefe CI, McDonell M, Fihn SD. Transforming self-rated health and the SF-36 scales to include death and improve interpretability. Med Care. 2001;39(7):670–680.

Diehr P, Patrick DL, McDonell MB, Fihn SD. Accounting for deaths in longitudinal studies using the SF-36: the performance of the physical component scale of the short form 36-item health survey and the PCTD. Med Care. 2003;41(9):1065–1073

Comment no. 6: For example, in Table 2 it not necessary to report mean MCS and PCS scores (and mean differences) to two decimal places; one decimal place is probably sufficient.

Response to comment no. 6: Thank you for your advice. We have changed all tables, including table 2 to present data with only one decimal.

Reviewer: 2

Comment no. 1: Thank you. The changes you have made have greatly strengthened the paper. This was a huge achievement although I realise a frustrating one for you

Response to comment no. 1: We are grateful for the kind words, we have received from the reviewer no 2.

VERSION 3 – REVIEW

REVIEWER	Professor Stephen Walters University of Sheffield
REVIEW RETURNED	06-Feb-2017

GENERAL COMMENTS	The authors have now responded satisfactorily to my initial comments.