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.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Health-Related Quality of Life and Anxiety in the PAN-CAN Lung Cancer Screening Cohort
AUTHORS	Taghizadeh, Niloofar; Tremblay, Alain; Cressman, Sonya; Peacock, Stuart; McWilliams, Annette; MacEachern, Paul; Johnston, Michael; Goffin, John; Goss, Glen; Nicholas, Garth; Martel, Simon; Laberge, Francis; Bhatia, Rick; Liu, Geoffrey; Schmidt, Heidi; Khattra, Sukhinder; Tsao, Ming-Sound; Tammemagi, Martin; Lam, Stephen

VERSION 1 – REVIEW

REVIEWER	Birol Baytan Uludag Univ. Ped. Hematology Bursa, Turkey
REVIEW RETURNED	07-Aug-2018

GENERAL COMMENTS	This study was planned well and it may bring innovation to the field

REVIEWER	Neill Booth University of Tampere, Finland
REVIEW RETURNED	07-Aug-2018

GENERAL COMMENTS	The manuscript may not be that far from receiving a "Yes" response to the remaining three items in the Review Checklist. I attach my comments and suggestions for "major revision" as a pdf file. General comment: Although this manuscript makes an admirable attempt to add new information to the literature, more care should be taken to make it clear to the reader that the these results stem from a study cohort which has no control group, as well as focussing on what this study can add to the literature. The information from this single-arm, non-randomised cohort study is likely to be most important in informing the design or focus of confirmatory randomised studies. Please consider even adding some indication of this fact to the title, e.g., by adding the word "Cohort" at the end of the title, or at least by adding an appropriate MESH term or two, such as "Cohort Studies", to the keywords! More major comments: 1) Before commenting further on the other results of the study (especially those results relating to Figures 6 through 8) it would be useful to be provided with unequivocal average baseline STAI-scores for the screen-negative sub-population, to supplement the average baseline STAI scores for the whole study population (30.9) and for the screen positives (29.9). The STAI-score runs

	from 20 to 80 and as the STAI has its floor only 10.9 or 9.9 below the above average scores, which may substantively affect the correct interpretation of the main STAI-related results. Histograms with thinner ‘bins’, possibly with a separate bin for each discrete value of the STAI score would likely help the reviewing process. My reading of the STAI-related results described between baseline and “post-baseline CT” (see lines 31 through 36 on page 14 & lines 43 through 53 on page 16) is that, as per the manuscript text, for the whole study population, increased anxiety occurs in around 20% (180/937, Figure 6). However, my reading of the results depicted in Figure 7, contrary to the manuscript text (lines 43 through 47 on page 16) and Figure 6, is that decreased anxiety occurs for around 20% of the screen positives (41/213). For the whole study population decreased anxiety seems to occur in around 5% (50/937) and increased anxiety for around 10% of the screen positives (20/213). A) Is it the case that the text on lines 43 through 47 on page 16: “However, more participants experienced a clinically significant increase vs. decrease in anxiety score” would be usefully replaced by “However, more participants experienced a clinically significant decrease in anxiety score”? B) Is it the case that the text on lines 49 through 52 on page 4: “and was present in both the cohort with negative and positive examinations” should be modified? C) Is it the case that the text on lines 20 through 23 on page 18: “Higher anxiety was also more frequent in the subgroup with positive baseline scan” should be modified? 2) Questions to a statistician (and/or to the authors): A) about lines 49 through 54 on page 10: would it be more appropriate to use a chi-square test for independence, given that, rather than there being a binary category “>MCID” & “<MCID” for STAI, there are three categories: “>MCID”, “change less than MCID” and “<MCID” for STAI? The following Stata code may be informative here: tabi 20 152 41 \ 180 707 50, cchi2 exact B) The above chi-square test seems to suggest that the cell containing those participants overall (180) for whom anxiety increased more than ‘the MCID’ does not make a major contribution to the overall chi-square statistic. C) Might the chi-square test –based approach reduce the potential problems arising from regression to the mean, which may undermine the use of a Z-test approach here? 3) Might participation in this screening study in general have an effect on smoking cessation and thus on the anxiety levels of participants (given that, some people supposedly experience reduced levels of anxiety as a result of smoking)? In particular, might the trial’s smoking cessation advice impact on participants’ smoking habits as the trial protocol (1.-PanCan-Early-Detection-of-Lung-Cancer-Protocol-14JUL17.pdf, section 4.4.16) states: “…every current smoker will be provided at the very minimum, a brochure such as Clear Horizons…”? 4) Re the comparison of the results of the current study to other studies: a) If the version of the STAI used is the STAI-Y, then would the STAI-related results of this study be most comparable to reference 21? Is reference 24 only comparable on a more general level? b) Is reference 22 the most comparable in terms of EQ-5D VAS? c) Is there any other study which has reported EQ-5D index scores, if not, should more focus be placed on these results in the current study? d) Are references 22 and 23 the most comparable in terms of the SF-12’s MCS and PCS?

	5) Although statistical power is discussed in relation to subgroup analysis, it would probably be informative for the reader to be given reasons why would this study be expected to reveal statistically significant in generic HRQoL scores, given both the results of earlier RCTs and the power of this study population to reveal small changes in HRQoL. That is, a) In general, could mean generic HRQoL scores be expected to change in any major (>= MCID) way between baseline and “post-baseline CT”?, b) To what extent is the current study powered to reveal small(er) changes in generic scores? and c) Would post-hoc sample-size calculation be a useful addition here (especially for the generic measures)? and d) Should the prominence and/or analysis of the generic-HRQoL results in this manuscript, perhaps with the exception of the EQ-5D index scores, be reduced? 6) Throughout the document the wording should not oversell the ability of a cohort study to provide explanatory evidence, i.e., its ability to provide any more than initial evidence of efficacy or effectiveness. Changes in anxiety and HRQoL during the period cannot be fully attributed to the LDCT and related study interventions because, e.g., we do not know if a control group would have experienced similar changes over a similar period. Given that this is a single-arm, prospective study, these results are likely to be exploratory or descriptive, rather than explanatory. The authors should refrain from overplaying findings, e.g., “helpful in determining if a true clinically significant impact is present”. For example, the authors could consider using the verb “suggest” instead of verbs such as “determine”, “show” or “demonstrate”. Minor comments: Lines 20 through 22 on page 20 make a claim about follow-up and response rate which should either be referenced or substantiated within the paper. Preferably the dropout or unit non-response rates for this part of the PLCSS-study should be more clearly reported here. Lines 25 through 37 on page 20 make a claim about generalisability on the basis of mean scores. The age groups do not match those in the study population for the EQ-5D and the SF-12, and appear to be missing for the STAI. The mean STAI scores quoted seem to be higher than those at baseline in the study population overall and higher than those in the screen positive group at baseline. Lines 32 to 33 on page 22: Shouldn’t “robust mortality reduction“ be replaced with “robust reduction in lung-cancer mortality“ or “robust reduction in mortality from lung cancer“? The STROBE checklist items 12c and 12d do not seem to be addressed fully in the text. Even more minor comments: The term “Pack years” should be explained. The abbreviation MCID (Minimal Clinically Important Difference) could be usefully defined on its first appearance in the body text. Units (years) should be provided for smoking duration in Table 1. As the EQ-5D-3L does not calculate scores if any dimension is missing, then missing scores could be reported in preference to missing dimensions in Supplementary Table 2. Does “baseline CT scan” mean “screening CT scan”?, so would “one month after screening CT scan” be more descriptive than “1-month post baseline CT scan”? Is the version of STAI used in the current study the STAI-Y, STAI-X, or some other version?

VERSION 1 – AUTHOR RESPONSE

Reviewers' Comments to Author:

Reviewer: 1

Reviewer Name: Birol Baytan

Institution and Country: Uludag Univ. Ped. Hematology Bursa, Turkey Please state any competing interests or state 'None declared': I have no conflict of interest with the study

This study was planned well and it may bring innovation to the field *Thank you for this review.*

Reviewer: 2

Reviewer Name: Neill Booth

Institution and Country: University of Tampere, Finland Please state any competing interests or state 'None declared': None declared

The manuscript may not be that far from receiving a "Yes" response to the remaining three items in the Review Checklist. I attach my comments and suggestions for "major revision" as a pdf file. *Thank you for this review.*

General comment:

Although this manuscript makes an admirable attempt to add new information to the literature, more care should be taken to make it clear to the reader that the these results stem from a study cohort which has no control group, as well as focussing on what this study can add to the literature. The information from this single-arm, non-randomised cohort study is likely to be most important in informing the design or focus of confirmatory randomised studies. Please consider even adding some indication of this fact to the title, e.g., by adding the word "Cohort" at the end of the title, or at least by adding an appropriate MESH term or two, such as "Cohort Studies", to the keywords!

Title modified as described above in editorial requests. The term "cohort study" was added to the keywords. Both the abstract and the method section already specify that this is a single arm study and that changes in QoL are compared to baseline pre-screening values. In other words, each individual is his/her own control group, which can be a more sensitive methodology to detect changes over time.

More major comments:

1) Before commenting further on the other results of the study (especially those results relating to Figures 6 through 8) it would be useful to be provided with unequivocal average baseline STAI-scores for the screen-negative sub-population, to supplement the average baseline STAI scores for the whole study population (30.9) and for the screen positives (29.9).

The mean baseline STAI for the group who were found to be screen negative was 31.2. This has been added to the text section of the results (page 16, line 231).

The STAI-score runs from 20 to 80 and as the STAI has its floor only 10.9 or 9.9 below the above average scores, which may substantively affect the correct interpretation of the main STAI-related results. Histograms with thinner 'bins', possibly with a separate bin for each discrete value of the STAI score would likely help the reviewing process.

The width of the histogram bars was specifically selected to represent the magnitude of the MCID. Using thinner bars of discrete values of the instruments, would fail to highlight this critical component of the analysis, as well as make it difficult to visualize the proportion of participants with changes smaller or greater than the MCID. Since changes in MCID are essentially clinically insignificant, this is a critical and unique component of our analysis in contrast to other reports of mean changes of statistical, but questionable clinical significance. As such we respectfully suggest that the histogram remain as submitted.

My reading of the STAI-related results described between baseline and "post-baseline CT" (see lines 31 through 36 on page 14 & lines 43 through 53 on page 16) is that, *as per the manuscript text*, for the whole study population, increased anxiety occurs in around 20% (180/937, Figure 6). However, my reading of the results depicted in Figure 7, *contrary to the manuscript text (lines 43 through 47 on page 16) and Figure 6*, is that decreased anxiety occurs for around 20% of the screen positives (41/213).

For the whole study population decreased anxiety seems to occur in around 5% (50/937) and increased anxiety for around 10% of the screen positives (20/213).

A) Is it the case that the text on lines 43 through 47 on page 16: "However, more participants experienced a clinically significant increase vs. decrease in anxiety score" would be usefully replaced by "However, more participants experienced a clinically significant decrease in anxiety score"?
B) Is it the case that the text on lines 49 through 52 on page 4: "and was present in both the cohort with negative and positive examinations" should be modified?
C) Is it the case that the text on lines 20 through 23 on page 18: "Higher anxiety was also more frequent in the subgroup with positive baseline scan" should be modified?

The reviewer is correct that the text is discordant with the figure 7 – thank you. The figure and values in the text are correct and the text has now been corrected.

2) Questions to a statistician (and/or to the authors):

A) about lines 49 through 54 on page 10: would it be more appropriate to use a chi-square test for independence, given that, rather than there being a binary category ">MCID" & "<MCID" for STAI, there are three categories: ">MCID", "change less than MCID" and "<MCID" for STAI? The following Stata code may be informative here: tabi 20 152 41 \ 180 707 50, cchi2 exact

B) The above chi-square test seems to suggest that the cell containing those participants overall (180) for whom anxiety increased more than 'the MCID' does not make a major contribution to the overall chisquare statistic.
C) Might the chi-square test –based approach reduce the potential problems arising from regression to the mean, which may undermine the use of a Z-test approach here?
We considered that cases with changes less than the MCID were non-informative and the analysis should assess whether an excess of cases with >MCID in either direction was present. In other

words, we are not interested to see if there are more cases with changes more or less than MCID, but only in the proportion that have true worsening vs. improvement in this measure. Using the MCID as a threshold for significant change reduces the chances of regression to the mean effect, as by definition, changes of such magnitude are more than normal / non-perceivable variations in an individual.

3) Might participation in this screening study in general have an effect on smoking cessation and thus on the anxiety levels of participants (given that, some people supposedly experience reduced levels of anxiety as a result of smoking)? In particular, might the trial's smoking cessation advice impact on participants' smoking habits as the trial protocol (1.-PanCan-Early-Detection-of-Lung-Cancer-Protocol-14JUL17.pdf, section 4.4.16) states: "…every current smoker will be provided at the very minimum, a brochure such as Clear Horizons…"?

The reviewer is correct that participation in a screening program has been associated with smoking cessation rates higher than the usual baseline rate in smokers in general, even when only minimal smoking cessation assistance is offered as was the case in our study. But smoking cessation has been associated with decreased in anxiety rather than increase (see BMJ. 2014 Feb 13;348) so that any smoking cessation would have been expected to reduce the magnitude of our findings in this regard. Baseline smoking was not associated with changes in anxiety in the multivariate analysis (table 3), but smoking status was not reassessed at 1-month post baseline, so that this could not be explored further, although we expect that quit rates within this short period would be very low.

4) Re the comparison of the results of the current study to other studies: a) If the version of the STAI used is the STAI-Y, then would the STAI-related results of this study be most comparable to reference 21? Is reference 24 only comparable on a more general level? b) Is reference 22 the most comparable in terms of EQ-5D VAS? c) Is there any other study which has reported EQ-5D index scores, if not, should more focus be placed on these results in the current study? d) Are references 22 and 23 the most comparable in terms of the SF-12's MCS and PCS?

We have now clarified the version of STAI by adding "form Y" to page 11, line 150. The references cited in our paper have either used form Y or have not indicated the version, although the Y version is the most commonly used since the update from form X in 1980. In general, we have cited the most relevant studies in terms of all measurements in the lung cancer screening setting when available. The Nelson study (22) used only the VAS component of the EQ-5D but not the full score which is not recommended. To our knowledge, no other lung cancer screening trial has used the full score. This is important as this score can be used to calculate QALYs and we have now highlighted this.

5) Although statistical power is discussed in relation to subgroup analysis, it would probably be informative for the reader to be given reasons why would this study be expected to reveal statistically significant in generic HRQoL scores, given both the results of earlier RCTs and the power of this study population to reveal small changes in HRQoL. That is, a) In general, could mean generic HRQoL scores be expected to change in any major (>= MCID) way between baseline and "post-baseline CT"?, b) To what extent is the current study powered to reveal small(er) changes in generic scores? and c) Would post-hoc sample-size calculation be a useful addition here (especially for the generic measures)? and d) Should the prominence and/or analysis of the generic-HRQoL results in this manuscript, perhaps with the exception of the EQ-5D index scores, be reduced?

Screening interventions in otherwise healthy individuals has the potential to cause harm even to those who receive no benefit from the screening / do not have the disease, which is the large majority of targeted individuals. As such, understanding if the intervention has overall negative impacts of QoL seems critical. It should be noted that at the time of study design, very little published information *existed on the QoL impact of LDCT screening for lung cancer. b/c: The interpretation of results and potential differences which could be detected can best be based on confidence intervals that are reported in our study. Our 95%CI for mean changes in all of the instruments used were narrow and much smaller than the MCIDs for each instrument. As such, we can be confident in these estimates. Performing post hoc sample-size calculation is not recommended [see: Goodman SN et al (Ann Intern Med. 1994 Aug 1;121(3):200-6]. d: The study was designed to measure a broad range of HRQoL and anxiety measures. Only focusing "post-hoc" on tools which demonstrated differences would represent a type of publication bias. In addition, the lack of impact of screening on such measures is also critical in informed decision-making regarding implementation of screening on a health system level as well as individual decision making.*

6) Throughout the document the wording should not oversell the ability of a cohort study to provide explanatory evidence, i.e., its ability to provide any more than initial evidence of efficacy or effectiveness. Changes in anxiety and HRQoL during the period cannot be fully attributed to the LDCT and related study interventions because, e.g., we do not know if a control group would have experienced similar changes over a similar period. Given that this is a single-arm, prospective study, these results are likely to be exploratory or descriptive, rather than explanatory. The authors should refrain from overplaying findings, e.g., "helpful in determining if a true clinically significant impact is present". For example, the authors could consider using the verb "suggest" instead of verbs such as "determine", "show" or "demonstrate".

We have revised the wording according to the reviewer's suggestion.

Also, the lack of a control group in our study had been acknowledge as a limitation in our discussion:

"Another potential limitation is that we did not compare our results to an unscreened control group but instead used each participant's baseline scores. As such, other factors unrelated to the screening intervention such as aging or changes in smoking status, could affect the longitudinal changes (or lack thereof) noted in our study"

Minor comments:

Lines 20 through 22 on page 20 make a claim about follow-up and response rate which should either be referenced or substantiated within the paper. Preferably the dropout or unit non-response rates for this part of the PLCSS-study should be more clearly reported here.

The proportion of different levels of each instrument dimensions by study visits and also number of missing are presented in supplementary tables 2-4 that can be used as a reference for our reported response rate.

Lines 25 through 37 on page 20 make a claim about generalisability on the basis of mean scores. The age groups do not match those in the study population for the EQ-5D and the SF-12, and appear to be missing for the STAI. The mean STAI scores quoted seem to be higher than those at baseline in the study population overall and higher than those in the screen positive group at baseline.

The reviewer is correct that the age and scores are not identical, but we believe "comparable" as stated, and certainly within each test's SD and MCID. We have added the age range for the STAI reference values.

Lines 32 to 33 on page 22: Shouldn't "robust mortality reduction" be replaced with "robust reduction in lung-cancer mortality" or "robust reduction in mortality from lung cancer"?

LDCT lung cancer screening has been demonstrated to reduce <u>*all-cause*</u> *mortality in the National Lung Cancer Screening trial. This has recently been confirmed in the NELSON study (not referenced as only presented in abstract form to date, Sept 2018). As such the statement is correct.*

The STROBE checklist items 12c and 12d do not seem to be addressed fully in the text.

STROBE checklist items 12c and 12d (regarding handling missing data):

Generalized linear mixed model that have been used in our study automatically handle missing data by maximum likelihood (Ibrahim JG, Chen M-H, Lipsitz SR, Herring AH. Missing data methods in generalized linear models: a comparative review. J Am Stat Assoc. 2005;100:332–346.). We have added this to the method section.

Even more minor comments:

The term "Pack years" should be explained.

We have now added the following definition: "Number of cigarettes per day / 20 x number of years of smoking".

The abbreviation MCID (Minimal Clinically Important Difference) could be usefully defined on its first appearance in the body text.

This has been added.

Units (years) should be provided for smoking duration in Table 1. *This has been added.*

As the EQ-5D-3L does not calculate scores if any dimension is missing, then missing scores could be reported in preference to missing dimensions in Supplementary Table 2. *The frequency of missing EQ-5D-3L scores has been added to the table.*

Does "baseline CT scan" mean "screening CT scan"?, so would "one month after screening CT scan" be more descriptive than "1-month post baseline CT scan"?

All of the scans are screening, but our reference time point is the initial or baseline scan. Theoretically, QoL impacts of subsequent annual scans could be different than after a baseline scan. As such, the "baseline" term is preferred to the less precise and descriptive than simply "screening CT scan".

Is the version of STAI used in the current study the STAI-Y, STAI-X, or some other version? *It is STAI-Y, and we have now added the version to the method section.*