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)	Education level and health-related quality of life after oesophageal cancer surgery - a nationwide cohort study
AUTHORS	Schandl, Anna; Johar, Asif; Mälberg, Kalle; Lagergren, Pernilla

VERSION 1 – REVIEW

REVIEWER	Jens Klein University Medical Center Hamburg-Eppendorf, Department Of Medical Sociology, Germany
REVIEW RETURNED	14-Dec-2017

GENERAL COMMENTS	Dear authors, Thank you for your interesting paper regarding the important research field of HRQOL among cancer patients. Nevertheless, I made some comments and suggestions. Methods/Discussion: Educational level (p. 5-6): The variable for educational level is only binary. So, it is not impossible to show a potential social gradient. As I am not a specialist regarding the Swedish educational system, wasn't there any other possibility in coding the variable? This could be discussed in the Discussion section. Reference population/Statistical analyses (p. 7-8) and Discussion: As the authors write in the discussion, to use normative data as a proxy for baseline/preoperative HRQOL is a potential and significant limitation. The so called “response shift” is a potential bias and could be discussed (e.g. Ubel et al. 2003/Sprangers et al. 1999). Moreover, HRQOL after 6 months is not included in the analysis when HRQOL after 3 years is the outcome variable (Table 4). Strictly speaking, the publication is about two cross-sectional analyses. So, it is not an analysis of individual data over the course of 3 years/over a three-year period (which, for instance, could be conducted with mixed models). The authors do not suggest using longitudinal analysis approach, but a “cohort study”/”longitudinal data collection” could potentially imply this. This could be clarified in Methods or Discussion. On page 7, the introduction of the statistical analyses is a bit misleading. Firstly, the calculation of mean score differences is mentioned to assess HRQOL. Afterwards, the conduction of linear regression to assess the association between the level of education and HRQOL is introduced. The reader expects in a second step regression estimates, but it is only about the already mentioned mean scores (which - I guess - are the adjusted estimated marginal means). This could be clarified (the wording in the abstract is

	clearer). Health-related quality of life assessment (p. 6-7) and Discussion: To estimate inequalities, the authors rely on the cut-off of ≥10 points based on Osoba et al. 1998 for defining clinical relevance. There is an ongoing debate about the interpretation of clinical significance regarding cross-sectional mean differences and scores over time (Binenbaum et al. 2014, Cocks et al. 2011, 2012). The latter recommend a more sensitive assessment for each subscale and distinguish between improving and declining scores. Norman et al. (2003) report in their review about the remarkable universality of half a standard deviation as threshold. Finally, Fayers (2001) resumes that clinical significance remains qualitative, is subjective and a matter of opinion. This lack of clarity could be discussed.

REVIEWER	Erin Kent and Maria Rincon, done jointly National Cancer Institute, USA
REVIEW RETURNED	25-Jan-2018

GENERAL COMMENTS	**BMJ Review: Education level and HRQoL after oesophageal cancer surgery—A nationwide cohort study** **I. Summary/Article overview** The study looks at the association between education level (low and high) and self-reported short and long-term HRQoL among esophageal cancer survivors after surgery in Sweden. The study consists of longitudinal follow-up (6 months and 3 years post-operative) of 90% of esophageal cancer cases diagnosed in Sweden between 2001 and 2005 who underwent surgical treatment. Reference values of HRQoL were obtained from a representative sample of non-cancer respondents; 10-point differences from the average values was considered clinically significant. Overall analysis found no clinical difference in the HRQoL cancer measure or disease-specific symptom scales. Analysis by gender reveal statistically significant poorer outcomes for women in the functional scales and symptoms, but not for men. This study provides disease-specific evidence of differences in HRQoL outcomes across gender in low education group, suggesting the importance of strategies that address post-treatment needs in this subgroup. More globally, this research suggests the need for us to understand more deeply what higher education actually means for well-being, and specifically patient well-being. It's telling that women manifest this association, but not men. Why? Perhaps more qualitative data collection is needed to understand the experiences of women of different educational strata to determine what is needed to move forward. Suggest adding more to the discussion on the possible reasons for this gender difference. **II. Questions/comments** **Abstract**

1. Page 2, conclusions (second sentence). There was no association for men between low education and HRQoL, but is this the same as stating is of “less importance”?

Introduction

2. Sentence 3 (pg. 4) – Phrasing is slightly confusing here. Do the authors mean that most patients will reach pre-operative HRQoL levels one year after surgery?
3. Though I like the phrase, “modern welfare states” may be unfamiliar to readers of medical journals. Suggest considering an alternative one.
4. In the introduction, a transition between indicating that SES has been associated cancer survival and education. Those less familiar with social determinants might not immediately make the connection that SES includes education. One sentence describing this connection would suffice.
5. Rationale, ideally informed by theory for the choice to look at stratification by gender is needed.

Methods

6. Reference population (pg. 7): baseline HRQoL data was obtained from reference 20. Paragraph states this proxy baseline was measured with the QLQ-C30 and the QLQ-OES18. A quick search did not reveal data for the use of the QLQ-OES18 in that article.
7. Comorbidity classification: defined in the article as “diabetes, cardiac, respiratory, renal or other specified condition” from the Swedish patient register (pg. 5). Number of comorbidities are classified as 0 or ≥1. Was further distribution considered, perhaps 0, 1, 2+ given that 36% of high education respondent and 51% of low educations respondents report ≥1 comorbidities?
8. Statistical analysis (pg. 8): North “Caroline” should be changed to North Carolina.
9. Commas typically belong inside the end of closed quotation marks (see. P. 6)

Results

10. Reporting of mean scores: in the tables for this article, “high” reference categories report mean scores and their 95% CIs, while “low” categories report mean difference from the reference category. Is this a standard method of reporting two groups under comparison?

Discussion

11. Limitations and strengths list (page 3) and discussion section (page 17) mention limited understanding on weak/moderate effects due to small sample size of stratified groups. Would analysis for significance in borderline scores to clinical significance help address this?

	12. Discussion: Suggest adding the phrase “in certain domains” after “However, in women, low education was associated with worse functioning and more symptoms.” 13. See overall comment re. need to further unpack gender differences in Discussion. **Figures 1 and 2 (page 24 and 25)** 14. Suggest changing the type of figure for Figure 1. Line graphs tend to communicate trends, and these are separate scales. Perhaps stacked bars would be more accurate? 15. In addition, the male groups, stratified by education, both use dashed lines (long versus short dashes), which can be difficult to differentiate, especially without different coloring **Throughout** 16. Female and male categorization referred to as “sex” instead of “gender” (pgs. 2,7,9,10,17,18). Would it be more appropriate to label this category as gender? How was sex/gender determined?

REVIEWER	Francesco Cavallin Independent statistician, Italy
REVIEW RETURNED	27-Jan-2018

GENERAL COMMENTS	The authors investigated the association between educational level and patients’ health-related quality of life (HRQL) after oesophagectomy for cancer. They found an association between lower education and worse HRQL (regarding some items) in women but not in men. The introduction is clear, the methods are appropriate and the manuscript is well-written. Overall the topic could be interesting but the study is flawed by the incorrect interpretation of the results. In fact, given the clinical significance for mean difference (MD) set at 10 points, any further investigations of statistical significance should test the null hypothesis MD=10, in order to claim that such MD is statistically significant. Thus, any statistically significant 2-sided test (i.e. $p<0.05$) would be associated with a corresponding 95% confidence interval excluding any values between -10 and 10. All significant results ($p<0.05$) in the manuscripy are associated with 95% confidence intervals including values between -10 and 10, thus the authors clearly tested the null hypothesis MD=0. The correct interpretation of the results in the manuscript is that those MDs are statistically different from 0, but we cannot exclude that the true value of the MDs can be between -10 and 10, thus including clinically non-significant values. Therefore, the authors should redo the statistical analysis, testing the null hypothesis MD=10 instead of MD=0, in order to obtain meaningful results. The discussion section should also be adjusted according to the new results. The participation rate is good (75%), but you should include a comparison of baseline characteristics between participants and

	non-participants to support your claim at the beginning of the results section; a supplementary table would be fine. I understand authors' justification but I still have some concerns about authors using HRQL measured in 4910 Swedish citizens as a proxy for patients' baseline HRQL, which is actually missing. Missing data are a well-known drawback of retrospective studies and I acknowledge the effort of the autors. However, I am not sure that using HRQL from general population can truly mirror the baseline HQRL in cancer patients, because of the low prevalence of alive esophageal cancer patients in the general population (due to low incidence and poor prognosis). Since I am investigating the postoperative change in HRQL in cancer patients, I would be interested in the HQRL during the disease and before the surgical treatment, in order to effectively measure that change. Therefore, using HQRL from general population can be a proxy of patients' HQRL before the neoplasm. Please remove the sentence "From this perspective, the use of HRQOL data from a healthy reference population in the analyses could provide more valid estimates compared to the preoperative baseline." from discussion section, because it does not add any further justification to your use of such proxy for patients' baseline HRQL (which is indeed a reasonable proxy). In Discussion, the sentence "A reasonably large sample size at 6 months provided good statistical power" is not supported by any results (you can see large confidence intervals in tables), please remove the sentence. In addition, please add the criteria for claiming statistical significance (I suppose alpha=0.05 and 2-tailed test) in the statistics section in methods.

REVIEWER	Chema Strik-Lips Radboudumc department of anesthesiology, Nijmegen, the Netherlands
REVIEW RETURNED	03-Feb-2018

GENERAL COMMENTS	I would like to congratulate the authors with this interesting paper. However, I do think there are some flaws with regard to the design of this study. Is educational level alone appropriate enough of a variable for a reduced HRQOL? Isn't there a possibility that this is also influenced by income, marital status etcetera, is there a possibility to take these variables into account? Is the distinction between 9 years or less versus 10 years or more not to crude or arbitrarily for making the statement that educational level influences HRQOL? Although I understand that the authors are limited by the number of participants but I think that both categories have patients that in the Netherlands would not have been classified as "low" or "high" educational levels. Wouldn't it be more interesting to further subdivide educational level and not stratify for men or women to better test your hypothesis? The lack of baseline measurement is a major flaw, the authors have tried to correct for this by using HRQOL of a large population as reference. I think this is incorrect, patients with esophageal carcinoma might not be similar to the general population as certain

	risk factors (ie smoking and alcohol consumption among others) might not be represented equally in this reference population. It is my recommendation to remove the HRQOL data of the reference population and only use data from the study population in order to clarify for readers that there is no true baseline measurement. Given the small sample of women in the cohort and the fact that there is no difference with regard to HRQOL and eductional level for men, isn't there a possibility that the found difference is a type 1 error? I believe it would be appropriate to perform some sort of correction for multiple hypothesis testing (Bonferroni or other) in order to strengthen your findings. Only neoadjuvant therapy is shown in the baseline table, adjuvant therapy might be more important, especially with regard to interpreting the 6-month HRQOL results, as chemotherapy can also influence taste and global HRQOL. The number of patients that required adjuvant therapy must be shown otherwise the results can't be interpreted completely. The time-span between data-collection and writing of this paper is rather long. In the discussion you state "The data collection of the study ended in 2008; even though this is not the most recent, there is no reason to believe that education level has less influence on HRQOL today compared to some years ago.", I beg to differ, the last decade things have changed in healthcare. Surgery has been performed more often minimal invasive. Additionally, due to dedicated oncologic pathways, information and guidance for patients has been improved, which might the results of this study less relevant now. Although, this would not impact the research objective directly, I believe that a more recent cohort would increase the relevance of this study. Minor suggestions: - is it possible to expand and subcategorize the variable "number of comorbidities"? Because to me, 0 comorbities or 1 or more comorbidities is not saying anything at all. - Could you also present a baseline table for the group of patients that survived for 3 years in order to appropriately interpret table 4?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

1. Methods/discussion: The variable for educational level is only binary. So, it is not impossible to show a potential social gradient. As I am not a specialist regarding the Swedish educational system, wasn't there any other possibility in coding the variable? This could be discussed in the Discussion section.
Reply: The reviewer poses a relevant question here. Initially, we classified the education variable in three categories: 1) compulsory education or ≤9 years: primary or secondary education (up to the age of 16 years), 2) moderate education or 10-12 years: upper secondary education 3) high education or ≥13 years: postsecondary education. This categorization resulted in a limited number of participants in the two first groups. To increase the statistical power we therefore decided to combine group 1 and 2. However, the analysis showed similar results irrespective of two or three groups. We have included a paragraph in the discussion to clarify this (page 18).

2. Reference population/Statistical analyses (p. 7-8) and Discussion:
As the authors write in the discussion, to use normative data as a proxy for baseline/preoperative HRQOL is a potential and significant limitation. The so called “response shift” is a potential bias and could be discussed (e.g. Ubel et al. 2003/Sprangers et al. 1999).
Reply: We thank the reviewer for this comment. To clarify this issue, we have included a paragraph regarding response shift as a limitation in the discussion section (page 19).

3. Moreover, HRQOL after 6 months is not included in the analysis when HRQOL after 3 years is the outcome variable (Table 4). Strictly speaking, the publication is about two cross-sectional analyses. So, it is not an analysis of individual data over the course of 3 years/over a three-year period (which, for instance, could be conducted with mixed models). The authors do not suggest using longitudinal analysis approach, but a “cohort study”/”longitudinal data collection” could potentially imply this. This could be clarified in Methods or Discussion.
Reply: We thank the reviewer for bringing this up. The reviewer is correct that the analyses at 6 months and 3 years are performed separately. We have therefore removed the term “longitudinal” when describing the data collection. We have also clarified this in the method section (page 7).

4. On page 7, the introduction of the statistical analyses is a bit misleading. Firstly, the calculation of mean score differences is mentioned to assess HRQOL. Afterwards, the conduction of linear regression to assess the association between the level of education and HRQOL is introduced. The reader expects in a second step regression estimates, but it is only about the already mentioned mean scores (which - I guess - are the adjusted estimated marginal means). This could be clarified (the wording in the abstract is clearer).
Reply: We apologize if we were not clear in our description of how the statistical analyses were done. We have now rewritten and clarified parts of this paragraph (page 7).

5. Health-related quality of life assessment (p. 6-7) and Discussion:
To estimate inequalities, the authors rely on the cut-off of ≥10 points based on Osoba et al. 1998 for defining clinical relevance. There is an ongoing debate about the interpretation of clinical significance regarding cross-sectional mean differences and scores over time (Binenbaum et al. 2014, Cocks et al. 2011, 2012). The latter recommend a more sensitive assessment for each subscale and distinguish between improving and declining scores. Norman et al. (2003) report in their review about the remarkable universality of half a standard deviation as threshold. Finally, Fayers (2001) resumes that clinical significance remains qualitative, is subjective and a matter of opinion. This lack of clarity could be discussed.
Reply: Thank you for this comment. Clinical significance is in one way more important than a statistical significance when it comes to analyses of patient-reported data. The choice of using the cut off of 10 mean scores for clinical significance for all subscales and items can be discussed. Since we use a disease-specific module together with the core questionnaire that do not have any studies to base the clinical significance level on we decided to use 10 as a cut off for all scales and items in the current study. We have compared the results of the QLQ-C30 scales and items with the Cocks et al reference for clinical significance in cross sectional studies and if using this reference almost all (except for a few scales and items at 6 months and 3 years assessment) would have reached the level of clinical significance and been further tested for statistical significance. Since we in priory decided the cut off of 10 we would like to keep the method as it is. We have added a paragraph about this in the discussion section of the revised version of the manuscript (page 18).

Reviewer 2

1. Page 2, conclusions. There was no association for men between low education and HRQoL, but is this the same as stating is of “less importance”.
Reply: The reviewers are right absolutely right about this. We have rephrased the sentence to “while no such association was found for men.”

2. Sentence 3, phrasing is slightly confusing here. Do the authors mean that most patients will reach pre-operative HRQoL levels one year after surgery?
Reply: We apologize if the sentence was not clear and thank the reviewers for their suggestion. We have rephrased the sentence to ”Most patients reach pre-operative HRQOL levels one year after surgery” (page 4).

3. Though I like the phrase, “modern welfare states” may be unfamiliar to readers of medical journals. Suggest considering an alternative one.
Reply: We thank the reviewers for this suggestion. We have removed the sentence to avoid possible misunderstandings

4. In the introduction, a transition between indicating that SES has been associated with cancer survival and education. Those less familiar with social determinants might not immediately make the connection that SES includes education. One sentence describing this connection would suffice.
Reply: Thank you for this comment. We have accordingly rephrased the whole paragraph to focus on education level instead of socioeconomic status (page 4).

5. Rationale, ideally informed by theory for the choice to look at stratification by gender is needed.
Reply: We agree and have added the following sentence “the perception on healthcare and health-related quality of life seems to differ in men and women” (Wessels et al. Gender-related needs and preferences in cancer care indicate the need for an individual approach to cancer patients, The Oncologist 2010;15:648-655; Tan et al. Evaluation of gender-specific aspects on quality-of-life in patients with larynx carcinoma, Acta Oto-Laryngologica 2016; 136:1201-1205) (page 4).

6. Reference population: baseline HRQol data was obtained from reference 20. Paragraph states this proxy baseline was measured with the QLQ-C30 and the QLQ-OES18. A quick search did not reveal data for the use of the QLQ-OES18 in that article.
Reply: We did not use the reference values from this reference. We have collected a reference population database which has both QLQ-C30 and QLQ-OES18 from the randomly selected background population (A portion of this reference database was used in this reference. We have further clarified the process of creating proxy database in the manuscript (page 7).

7. Comorbidity classification: defined in the article as “diabetes, cardiac, respiratory, renal or other specified condition” from the Swedish patient register. Number of comorbidities are classified as 0 or ≥1. Was further distribution considered, perhaps 0,1,2+given that 36% of high education respondent and 51% of low education respondents report ≥1 comorbidities?
Reply: We decided on the grouping in our study protocol before the data analysis. However, due to this question, we also ran the analysis with three categories and the results were very similar.

8. Statistical analysis: North “Caroline” should be changed to North Carolina.
Reply: We thank the reviewers for pointing this out and apologize for the spelling mistake. This has now been corrected.

9. Commas typically belong inside the end of closed quotation marks.

Reply: We thank the reviewer for pointing this out. The mistake has been checked and corrected throughout the manuscript.

10. Reporting of mean scores: in the tables for this article, “high” reference categories report mean scores and their 95% Cis, while “low” categories report mean difference from the reference category. Is this a standard method of reporting two groups under comparison?
Reply: We are grateful for this comment. The way of presenting is commonly used. By reporting the reference group mean score the reader can interpret where on the scale (often a scale from 0 to 100) the actual values and the difference are. If you only present the mean score differences you do not know where the differences originate from, e.g., is a difference from a high level of functions or low level of function.

11. Limitations and strengths list and discussion section mention limited understanding on weak/moderate effects due to small sample size of stratified groups. Would analysis for significance in borderline scores to clinical significance help address this?
Reply: Yes, in some cases this approach can be helpful. However, this approach requires further investigation where it can be applied and analyzed. Recommendations from this work can be useful for future research.

12. Discussion: Suggest adding the phrase “in certain domains” after “However, in women, low education was associated with worse functioning and more symptoms”.
Reply: Thank you for this suggestion, we have added this suggestion to the sentence (page 18).

13. See overall comment regarding need to further unpack gender differences in Discussion.
Reply: The literature about this topic is sparse, but we have further elaborated on gender differences in HRQOL in the discussion (page 19).

14. Suggest changing the type of figure for Figure 1. Line graphs tend to communicate trends, and these are separate scales. Perhaps stacked bars would be more accurate? In addition, the male groups, stratified by education, both use dashed lines (long versus short dashes), wish can be difficult to differentiate, especially without different colouring.
Reply: We have designed two alternative figures without lines, but leave it up to the reviewer to choose which they think would best show the results. See enclosed files plot1_version2_reply to reviewers and plot2_version2_reply to reviewers.

15. Female and male categorization referred to as “sex” instead of “gender”. Would it be more appropriate to label this category as gender? How was sex/gender determined?
Reply: We agree with the reviewer and have changed sex to gender where we found it applicable. Gender was determined in regard to the individual social security number.

Reviewer 3

1. Overall the topic could be interesting but the study is flawed by the incorrect interpretation of the results. In fact, given the clinical significance for mean difference (MD) set at 10 points, any further investigations of statistical significance should test the null hypothesis MD=10, in order to claim that such MD is statistically significant. Thus, any statistically significant 2-sided test (i.e. $p<0.05$) would be associated with a corresponding 95% confidence interval excluding any values between -10 and 10. All significant results ($p<0.05$) in the manuscript are associated with 95% confidence intervals including values between -10 and 10, thus the authors clearly tested the null hypothesis MD=0. The correct interpretation of the results in the manuscript is that those MDs are statistically different from

0, but we cannot exclude that the true value of the MDs can be between -10 and 10, thus including clinically non-significant values. Therefore, the authors should redo the statistical analysis, testing the null hypothesis MD=10 instead of MD=0, in order to obtain meaningful results. The discussion section should also be adjusted according to the new results.
Reply: Thank you for your informative comment. The description of the application of clinically relevant MD was ambiguous in the statistical analysis section which now have been rectified. We do not test the clinically relevant difference. We test the null hypothesis of MD=0 only for scales which have clinically relevant difference (this is in accordance with the reference paper by Osoba et al.) We have previously used this method in other studies e.g. (Derogar et al, Health-related quality of life among 5-year survivors of esophageal cancer surgery: a prospective population-based study. J Clin Oncol 2012;30:413-8)(page 7).

2. The participation rate is good (75%), but you should include a comparison of baseline characteristics between participants and non-participants to support your claim at the beginning of the results section; a supplementary table would be fine.
Reply: We understand the reviewer's concern, and have included the suggested tables as supplementary material (page 8).

3. I understand authors' justification but I still have some concerns about authors using HRQL measured in 4910 Swedish citizens as a proxy for patients' baseline HRQL, which is actually missing. Missing data are a well-known drawback of retrospective studies and I acknowledge the effort of the autors. However, I am not sure that using HRQL from general population can truly mirror the baseline HQRL in cancer patients, because of the low prevalence of alive esophageal cancer patients in the general population (due to low incidence and poor prognosis). Since I am investigating the postoperative change in HRQL in cancer patients, I would be interested in the HQRL during the disease and before the surgical treatment, in order to effectively measure that change. Therefore, using HQRL from general population can be a proxy of patients' HQRL before the neoplasm. Please remove the sentence "From this perspective, the use of HRQOL data from a healthy reference population in the analyses could provide more valid estimates compared to the preoperative baseline." from discussion section, because it does not add any further justification to your use of such proxy for patients' baseline HRQL (which is indeed a reasonable proxy).
Reply: We are grateful for this comment and would like to clarify. The proxy for patients´ baseline HRQOL is supposed to mirror the HRQOL patients experienced before the cancer diagnosis, i.e., not people with esophageal cancer. We have clarified this in the method section (page 7). According to the reviewer's recommendation, we have removed the suggested sentence from the discussion.

4. In Discussion, the sentence "A reasonably large sample size at 6 months provided good statistical power" is not supported by any results (you can see large confidence intervals in tables), please remove the sentence.
Reply: We agree and have removed this sentence from the discussion.

5. In addition, please add the criteria for claiming statistical significance (I suppose alpha=0.05 and 2-tailed test) in the statistics section in methods.
Reply: We thank the reviewer for observing this and have added a sentence in the method section (page 8).

Reviewer 4
1. Is educational level alone appropriate enough of a variable for a reduced HRQOL? Isn't there a possibility that this is also influenced by income, marital status etcetera, is there a possibility to take these variables into account?
Reply: The intention with this study was to study education level in relation to HRQOL. However, we agree with the reviewer that there are also other variables such as income and marital status that are

of importance for HRQOL recovery. We have included a paragraph about this in the discussion (page 19).

2. Is the distinction between 9 years or less versus 10 years or more not to crude or arbitrarily for making the statement that educational level influences HRQOL? Although I understand that the authors are limited by the number of participants but I think that both categories have patients that in the Netherlands would not have been classified as "low" or "high" educational levels. Wouldn't it be more interesting to further subdivide educational level and not stratify for men or women to better test your hypothesis?
Reply: The reviewer poses a relevant question here and the same comment was raised by reviewer 1. Please see reply to reviewer 1, comment 1.

3. The lack of baseline measurement is a major flaw, the authors have tried to correct for this by using HRQOL of a large population as reference. I think this is incorrect, patients with esophageal carcinoma might not be similar to the general population as certain risk factors (ie smoking and alcohol consumption among others) might not be represented equally in this reference population. It is my recommendation to remove the HRQOL data of the reference population and only use data from the study population in order to clarify for readers that there is no true baseline measurement.
Reply: This was also raised by reviewer 3. Please see reply to reviewer 3, comment 3.

4. Given the small sample of women in the cohort and the fact that there is no difference with regard to HRQOL and eductional level for men, isn't there a possibility that the found difference is a type 1 error? I believe it would be appropriate to perform some sort of correction for multiple hypothesis testing (Bonferroni or other) in order to strengthen your findings.
Reply: We thank the reviewer for the comment. However, to address the problem of being exposed to type 1 error when performing multiple significance tests, we chose to test only those variables with clinically relevant differences. This is a common way when analyzing HRQOL data where the actual difference for the individual patient is more important than the statistical significance. If you have a large scale study you may have nice statistical significant results that are not meaningful or noticeable for the patient.

5. Only neoadjuvant therapy is shown in the baseline table, adjuvant therapy might be more important, especially with regard to interpreting the 6-month HRQOL results, as chemotherapy can also influence taste and global HRQOL. The number of patients that required adjuvant therapy must be shown otherwise the results can't be interpreted completely.
Thank you for this comment. However, at the time point for the data collection, adjuvant therapy was rarely used in oesophageal cancer surgery patients. Moreover, no such data are available in this database. This has been included as a limitation in the manuscript (page 19).

6. The time-span between data-collection and writing of this paper is rather long. In the discussion you state "The data collection of the study ended in 2008; even though this is not the most recent, there is no reason to believe that education level has less influence on HRQOL today compared to some years ago.", I beg to differ, the last decade things have changed in healthcare. Surgery has been performed more often minimal invasive. Additionally, due to dedicated oncologic pathways, information and guidance for patients has been improved, which might the results of this study less relevant now. Although, this would not impact the research objective directly, I believe that a more recent cohort would increase the relevance of this study.
Reply: Thank you for this comment. The majority of this cohort of oesophageal cancer patients, consists of elderly people. For those patients, education level would not change even with a more recent cohort. Health care, on the other hand, has developed during the last decade, which may have led to improved HRQOL. We have included a sentence about limited generalisability on page 19.

7. Minor suggestions:
- is it possible to expand and subcategorize the variable "number of comorbidities"? Because to me, 0 comorbities or 1 or more comorbidities is not saying anything at all.
Reply: This was also raised by reviewer 2. Please see reply to reviewer 2, comment 7.

- Could you also present a baseline table for the group of patients that survived for 3 years in order to appropriately interpret table 4. Is educational level alone appropriate enough of a variable for a reduced HRQOL? Isn't there a possibility that this is also influenced by income, marital status etcetera, is there a possibility to take these variables into account?
Reply: We thank the reviewer for this suggestion and have included a baseline table for the 3 year survivors (page 9-10).

VERSION 2 – REVIEW

REVIEWER	Jens Klein University Medical Center Hamburg-Eppendorf, Germany, Department of Medical Sociology
REVIEW RETURNED	22-Mar-2018

GENERAL COMMENTS	The authors have adequately revised the manuscript and addressed the issues raised by the reviewer.

REVIEWER	Francesco Cavallin Independent statistician, Italy
REVIEW RETURNED	30-Mar-2018

GENERAL COMMENTS	I congratulate the authors for the revision of the manuscript.

REVIEWER	Erin E. Kent National Cancer Institute, Outcomes Research Branch
REVIEW RETURNED	31-Mar-2018

GENERAL COMMENTS	The author's comments and revised manuscript were responsive, thorough, and thoughtful. Only two minor comments from me remain: 1. Introduction: The revised sentence, "Moreover, the perception on healthcare and HRQOL seems to differ in men and women (10, 11)." Did you mean that the "association" of healthcare and HRQOL seem to differ "between" mean and women? 2. The authors state that there are two versions of the revised figures attached, one with lines and one with bars, and I see them listed as attachments at the bottom of the response document, but I cannot open them as to advise. Perhaps they need to be uploaded through the actual submission portal (not attached)?

REVIEWER	C. Strik-Lips Radboudumc, department of anesthesiology, the Netherlands
REVIEW RETURNED	03-Apr-2018

GENERAL COMMENTS	1. Is educational level alone appropriate enough of a variable for a reduced HRQOL? Isn't there a possibility that this is also influenced by income, marital status etcetera, is there a possibility to take these

	variables into account? Reply: The intention with this study was to study education level in relation to HRQOL. However, we agree with the reviewer that there are also other variables such as income and marital status that are of importance for HRQOL recovery. We have included a paragraph about this in the discussion (page 19). Reviewer reply: Just because education level is stored in a national database doesn't mean that it's a more important determinant for health-related quality of life. The absence of these variables is a serious flaw of this study. 2. The lack of baseline measurement is a major flaw, the authors have tried to correct for this by using HRQOL of a large population as reference. I think this is incorrect, patients with esophageal carcinoma might not be similar to the general population as certain risk factors (ie smoking and alcohol consumption among others) might not be represented equally in this reference population. It is my recommendation to remove the HRQOL data of the reference population and only use data from the study population in order to clarify for readers that there is no true baseline measurement. Reply: This was also raised by reviewer 3. Please see reply to reviewer 3, comment 3. Reviewer reply; The authors used a case-matching analysis to justify a HRQOL measurement in the general population as a by proxy baseline measurement for the cohort study. Although the authors matched patients based on age, gender, number of comorbidities and education level, I believe this is wrong because these variables are very limited in predicting HRQOL and probably do not reflect the study population. Furthermore, the authors use this "baseline measurement" in their logistic regression analysis to study the effect of other variables on HRQOL, this introduces more bias in the study than interpreting the results without a baseline measurement. It is my strong suggestion to remove the "baseline measurement" from the study. 3. Given the small sample of women in the cohort and the fact that there is no difference with regard to HRQOL and eductional level for men, isn't there a possibility that the found difference is a type 1 error? I believe it would be appropriate to perform some sort of correction for multiple hypothesis testing (Bonferroni or other) in order to strengthen your findings. Reply: We thank the reviewer for the comment. However, to address the problem of being exposed to type 1 error when performing multiple significance tests, we chose to test only those variables with clinically relevant differences. This is a common way when analyzing HRQOL data where the actual difference for the individual patient is more important than the statistical significance. If you have a large scale study you may have nice statistical significant results that are not meaningful or noticeable for the patient. Reviewer reply: I believe the authors do not fully understand the basics of statistical reasoning. Normally, studies are powered on 1 outcome variable, if one would test significance between groups on a secondary outcome, a correction for multiple hypothesis testing already would be appropriate. In the present study 31 outcome variables were tested, this means that by chance alone 1.5 variable will reach statistical significance. Hopefully, the authors now understand that only assessing variables with a "clinically relevant" difference is NOT the correct way to adress multiple hypothesis testing. A common way in analysing results is not always the correct way. Sample size and effect size are not a factor in this question as

	the authors suggest by their sentence regarding large sample sizes and small effect sizes. Furthermore, "clinically relevant" differences in quality of life is a very slippery topic which will not only differ between patients, but will also differ in the same patient over time. Therefore, a change of 10 on the HRQOL scale might be a small difference for detecting your "clinically relevant" differences. The authors should incorporate the absence of correcting for multiple hypothesis testing in the discussion. 4. The conclusion of the study is wrong, the authors only found differences on certain subscales of the questionnaires while the conclusion states 'This study indicates that for women, low education level is associated with worse HRQOL outcome', this should be specified according to the results. Minor suggestions: - The manuscript could benefit from a language editor, for example, in line 46 page 18 the authors confuse precious with previous

VERSION 2 – AUTHOR RESPONSE

Reviewer 2

1. Introduction: The revised sentence, “Moreover, the perception on healthcare and HRQOL seems to differ in men and women (10, 11).” Did you mean that the “association” of healthcare and HRQOL seem to differ “between” mean and women?

Reply: We thank the reviewer for the suggested rephrasing of the sentence. We have revised according to the suggestion.

2. The authors state that there are two versions of the revised figures attached, one with lines and one with bars, and I see them listed as attachments at the bottom of the response document, but I cannot open them as to advise. Perhaps they need to be uploaded through the actual submission portal (not attached)?

Reply: We apologize for this inconvenience. However, the submission process does not allow submission of duplicate figures. This time the revised figures are uploaded as PDFs (Figure1_version2 and Figure2_version2).

Reviewer 4

1. Is educational level alone appropriate enough of a variable for a reduced HRQOL? Isn't there a possibility that this is also influenced by income, marital status

etcetera, is there a possibility to take these variables into account?

Reply: The intention with this study was to study education level in relation to HRQOL. However, we agree with the reviewer that there are also other variables such as income and marital status that are of importance for HRQOL recovery. We have included a paragraph about this in the discussion (page 19).

Reviewer reply: Just because education level is stored in a national database doesn't mean that it's a more important determinant for health-related quality of life. The absence of these variables is a serious flaw of this study.

Reply: At an initial state we considered using income and occupation-based measures of socioeconomic position, but chose individual education level since it is robust and easy to measure. Income is rather complex to measure as one has to take the wealth and the numbers supported by the income into account.

Education level is relevant to people regardless of age or working circumstances unlike e.g. occupational status and many other socioeconomic indicators. The majority of patients with oesophageal cancer are senior citizens and we therefore chose educational level as exposure measure in the study. An explanation is included in the beginning of the discussion.

2. The lack of baseline measurement is a major flaw, the authors have tried to correct for this by using HRQOL of a large population as reference. I think this is incorrect, patients with esophageal carcinoma might not be similar to the general population as certain risk factors (ie smoking and alcohol consumption among others) might not be represented equally in this reference population. It is my recommendation to remove the HRQOL data of the reference population and only use data from the study population in order to clarify for readers that there is no true baseline measurement.

Reply: This was also raised by reviewer 3. Please see reply to reviewer 3, comment 3.

Reviewer reply: The authors used a case-matching analysis to justify a HRQOL measurement in the general population as a by proxy baseline measurement for the cohort study. Although the authors matched patients based on age, gender,

number of comorbidities and education level, I believe this is wrong because these variables are very limited in predicting HRQOL and probably do not reflect the study population. Furthermore, the authors use this "baseline measurement" in their logistic regression analysis to study the effect of other variables on HRQOL, this introduces more bias in the study than interpreting the results without a baseline measurement. It is my strong suggestion to remove the "baseline measurement" from the study.

Reply: According to the reviewer's suggestion we have removed the HRQOL data of the reference population from the multivariable regression models. The new results are presented in table 3,4,5 and figure 1 and 2. We have also removed the paragraph about the reference population in the method section. The limitation discussion starts with the lack of preoperative baseline HRQOL data.

3. Given the small sample of women in the cohort and the fact that there is no difference with regard to HRQOL and eductional level for men, isn't there a possibility that the found difference is a type 1 error? I believe it would be appropriate to perform some sort of correction for multiple hypothesis testing (Bonferroni or other) in order to strengthen your findings.

Reply: We thank the reviewer for the comment. However, to address the problem of being exposed to type 1 error when performing multiple significance tests, we chose to test only those variables with clinically relevant differences. This is a common way when analyzing HRQOL data where the actual difference for the individual patient is more important than the statistical significance. If you have a large scale study you may have nice statistical significant results that are not meaningful or noticeable for the patient.

Reviewer reply: I believe the authors do not fully understand the basics of statistical reasoning. Normally, studies are powered on 1 outcome variable, if one would test significance between groups on a secondary outcome, a correction for multiple hypothesis testing already would be appropriate. In the present study 31 outcome variables were tested, this means that by chance alone 1.5 variable will

reach statistical significance. Hopefully, the authors now understand that only assessing variables with a "clinically relevant" difference is NOT the correct way to adress multiple hypothesis testing. A common way in analysing results is not always the correct way. Sample size and effect size are not a factor in this question as the authors suggest by their sentence regarding large sample sizes and small effect sizes.

Furthermore, "clinically relevant" differences in quality of life is a very slippery topic which will not only differ between patients, but will also differ in the same patient over time. Therefore, a change of 10 on the HRQOL scale might be a small difference for detecting your "clinically relevant" differences.

The authors should incorporate the absence of correcting for multiple hypothesis testing in the discussion.

Reply: Thank you for the comment. Adjusting for the multiple testing is not an uncontroversial methodological topic (ref) as one inflates the type II error while trying to reduce the type I error. We have mentioned in the manuscript that we have not adjusted for the multiple testing.

Rothman K.J.: No adjustments are needed for multiple comparisons. Epidemiology 1990; 1: pp. 43-46

4. The conclusion of the study is wrong, the authors only found differences on certain subscales of the questionnaires while the conclusion states 'This study indicates that for women, low education level is associated with worse HRQOL outcome', this should be specified according to the results.

Reply: The conclusion has been rephrased to: “Low education was not associated with worse HRQOL after oesophageal cancer surgery. However, when data were stratified for sex, low education level was associated with worse functioning and more symptoms in certain HRQOL domains for women. For men, no such association was found.”

Minor suggestions:

- The manuscript could benefit from a language editor, for example, in line 46 page

18 the authors confuse precious with previous

Reply: We apologize for the spelling mistake. This has now been corrected.

VERSION 3 – REVIEW

REVIEWER	Erin E. Kent National Cancer Institute, Rockville, MD, United States
REVIEW RETURNED	25-May-2018

GENERAL COMMENTS	The authors were responsive to the reviewer's comments. I have no further comments at this time.