PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cost-effectiveness of total knee replacement in addition to non-surgical treatment: 2-year outcome from a randomized trial in secondary care in Denmark
AUTHORS	Skou, Søren; Roos, Ewa; Laursen, Mogens; Arendt-Nielsen, Lars; Rasmussen, Sten; Simonsen, Ole; Ibsen, Rikke; Larsen, Arendse; Kjellberg, Jakob

VERSION 1 – REVIEW

REVIEWER	Jamie Collins Brigham and Women's Hospital, USA
REVIEW RETURNED	28-Aug-2019

GENERAL COMMENTS	This manuscript presents a cost-effectiveness analysis of total knee replacement vs. usual care. It finds that TKR is not cost-effective at a willingness to pay threshold of 22,665 Euros. This is in contrast to many previous studies, which have found TKR to be cost-effective. The manuscript is well written, clear, and sufficiently detailed to understand the analyses. However, I have some major concerns about the study design. The authors note in the discussion section that the findings are not robust (different results depending on covariate adjustment, imputation, inclusion of deaths), and given that this study only included 100 subjects, I would recommend tempering the conclusions considerably. It seems the conclusion is not that "TKR plus non-surgical treatment is not cost-effective compared to non-surgical treatment alone", but that more research is needed and under certain circumstances TKR may not be cost-effective. My major concern is with the time horizon, and I have a few additional questions (see below). 1) The major concern here is that the time horizon is only 2 years. Certainly one would expect the benefits of TKR to last more than 2 years, thus the current study dramatically underestimates the benefits associated with TKR. The authors note this in the discussion section of the manuscript, but do not account for TKR subjects continuing to accumulate QALYs as they live (many pain-free) with their joint replacement. I would expect TKR recipients' QoL to continue to be higher beyond the 2 year timeframe. If that's true, then TKR recipients would continue to accrue more QALYs than the non-surgical group after the end of 2 years. While at the same time, the non-TKR group will continue to accumulate costs as more subjects randomized to non-surgical treatment convert to TKR. The 2nd panel on Cost-Effectiveness in Health and Medicine recommends that the time horizon considered be long enough to capture all differences in costs and benefits between scenarios. I do not think that an analysis of the cost-effectiveness of TKR after 2 years is meaningful and I cannot recommend this manuscript for publication without an extended time horizon.

	2) Patients with severe pain (>60/100 on VAS) were excluded from the trial. The results may not be generalizable to all patients with knee OA. This is noted in the discussion but warrants further attention. What percentage of otherwise eligible knee OA patients were excluded from the trial due to severe pain? Are patients with severe pain less likely to benefit from non-operative therapy? More detail is needed so that readers can better understand this limitation. 3) Given that almost one-third of the "non-surgical" group underwent TKR within 2 years of randomization, is it really fair to call this group non-surgical? Could it instead be framed as a non-surgical intervention with an option for delayed TKR? How were the costs and QALYs for these cross-overs incorporated? 4) Cost-effectiveness acceptability curves could help the reader the uncertainty in the probabilistic sensitivity analyses.

REVIEWER	Rafael Pinedo-Villanueva University of Oxford, UK
REVIEW RETURNED	10-Oct-2019

GENERAL COMMENTS	Thanks to the authors for an interesting manuscript. Below my comments and suggestions: In the abstract, a distinction is made between cost-effectiveness results in the unadjusted compared to the adjusted analysis. Please specify what such adjustment considered. The conclusion assumes generalisability, which I doubt can be taken for granted with a sample of 100 whose selection (before randomisation) has not even been described to this point. I would either make explicit and justify the assumption of full generalisability, or tone down the conclusion to make it specifically relevant to that group of 100 patients. In any case, the conclusion should also be limited to the geographical area where the trial was conducted, or at a minimum to Denmark, considering that clinical practice and costs vary greatly amongst countries even within Europe. This applies to the respective section(s) in the manuscript. The term "data" is used in this paper as the aggregation of specific pieces of information, hence it is a plural noun and the verb that follows should reflect that. In some cases it does, in others it doesn't; please correct when necessary. In the Methods section, when reporting the exclusion criteria for the trial, some context as to the significance and interpretation of the 60 mm over 100 mm VAS for mean pain the previous week is necessary. Could this be read as the trial excluding patients at the severe end of symptomatic knee OA spectrum? If so, then discussion of results and conclusions should reflect that, much beyond the comment made in the limitations section. Please define "high-volume orthopedic specialists". It is not clear to me why "Non protocol-driven resources, e.g. costs of recruitment, were included" if they were generated by each group. Please explain. Under "Measurement of effectiveness" you seem to be confusing the concept of QALY with that of health utility, or additional details are

	needed. QALYs do not have a “maximum achievable” of 1 (whilst health utility does), unless it is restricted to one year. Please refine. You indicate that “QALY values were imputed at 24 months”. I want to think this is a mistake as imputing QALYs is likely to generate greater bias than imputing only the missing health utilities. If you actually meant that it was health utilities what were imputed, then please correct. If you indeed imputed QALYs, then you should consider imputing health utility values and calculate QALYs based on the observed and imputed scores. I find the QALY gain calculation (formula) confusing. Why was it not done by calculating area under the curve for each year, and then discounting that of year 2? If there is a good reason do follow the formula shown, please explain and justify as I am not convinced it is the most appropriate way to calculate QALYs in this case. If Table 3 excludes deaths, then what is being reported is essentially health utilities as survival is assumed for everyone. This is linked to the confusing presentation of health utilities and QALYs. Please revise to add clarity. Table 6: can you please clarify in the note/legend that the base-case analysis excludes deaths? I -however- believe it should include them. I disagree with your statement that “If TKR plus non-surgical treatment was to become cost-effective in the longer term, the decrease in QALY in the non-surgical group would need to continue.” I don’t believe that is correct. This would be most clearly shown if you report health utilities for each arm at each point. There it would become clear, I would expect, that (leaving deaths to the side for the moment, as you id in your primary base-case analysis), the additional improvement achieved by the TKR group would be maintained in the time, with costs growing more rapidly for the non-surgical arm, hence leading inexorably to a reduction of the ICER until it would eventually fall under the threshold. This is likely why the two studies identified using decision-analytic models over longer time horizons found the opposite of what you did. The sensitivity of results to the time horizon must be highlighted in the abstract, discussion, limitations and conclusion. In the conclusion, you may want to emphasise that you mean “[not] imputing missing values”.

REVIEWER	Ross Wilson University of Otago, New Zealand I collaborate with Prof Skou on an unrelated research grant application.
REVIEW RETURNED	16-Oct-2019

GENERAL COMMENTS	This is a well-conducted trial investigating a clinically and economically important question. My comments are generally concerned with the presentation and interpretation of the cost-effectiveness analysis. 1. The cost perspective is reported as 'a limited societal perspective (i.e. health care costs and public transfer payments)'. The data on public transfer payments do not seem to be used in any cost-utility

	analyses, have not been valued in monetary terms (i.e. 'costs'), and appear to be available for only one third of the sample, so I am not clear as to the value of including these at all; perhaps a health system perspective only would be more appropriate? 2. Do the registers from which cost data were extracted include the actual costs of care, or just resource use data? If the latter, how were reference costs obtained and what were these unit costs? 3. While the NICE threshold is a useful reference, there seems no particular reason to believe that it is the appropriate threshold for use in Denmark (or any other context in which these results may be interpreted). Presenting the sensitivity of the cost-effectiveness finding to the choice of threshold (i.e. a cost-effectiveness acceptability curve across a range of threshold values) would improve the interpretation of results as compared to using the single NICE decision threshold. 4. I was confused by the terminology regarding the measurement of health gains over time. The values 'QALY baseline', 'QALY 3 months', etc on the RHS of the equation on line 261 (and likewise in Tables 3 and 5) are presumably the instantaneous HRQoL utility values as measured by the EQ-5D, not QALYs (which are a cumulative measure of HRQoL experienced over time) If this is the case, I don't think the statements regarding potential cost-effectiveness over a longer time horizon follow from the observed data as suggested in lines 377-387. The gain in HRQoL of TKR compared to non-surgical treatment was 0.09 at 3 months, 0.133 at 6 months, 0.085 at 12 months, and 0.142 at 24 months (0.139 with discounting), if I have understood Table 3 correctly. If anything, this seems to be increasing over time. Meanwhile the costs of TKR are all incurred in the first year, so with a longer time horizon, we would see increasing QALY gains with no increase in incremental costs, so improving cost-effectiveness ratios over time. 5. The method of multiple imputation should be described. In addition, it was reported that 'missing values occurred on utilities at 24 months' - were these the only missing data, or was multiple imputation also used for other variables? 6. Given the large number of cross-overs, the true effect (and cost) of TKR is likely to be understated. While this is likely unavoidable, given the ethical issues with denying TKR to those for whom it is considered necessary, it is an important limitation that needs to be acknowledged. A secondary per-protocol or compliers analysis might be informative to (partially) address this.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1
Reviewer Comment:
This manuscript presents a cost-effectiveness analysis of total knee replacement vs. usual care. It finds that TKR is not cost-effective at a willingness to pay threshold of 22,665 Euros. This is in contrast to many previous studies, which have found TKR to be cost-effective. The manuscript is well written, clear, and sufficiently detailed to un-derstand the analyses. However, I have some major concerns about the study design. The authors note in the discussion section that the findings are not robust (different results depending on covariate adjustment, imputation, inclusion of deaths), and

given that this study only included 100 subjects, I would recommend tempering the conclu-sions considerably. It seems the conclusion is not that “TKR plus non-surgical treat-ment is not cost-effective compared to non-surgical treatment alone”, but that more research is needed and under certain circumstances TKR may not be cost-effective. My major concern is with the time horizon, and I have a few additional questions (see below).

Author response:
Thank you for your valuable and important feedback, that have helped improve the manuscript further.
We do agree that our findings are in contrast to several previous studies1, but on the other hand aligns with a recent study that found that TKR was not cost-effective at a group level over 8 years2. Importantly, all previous studies were either based on a Mar-kov model or cohort-based studies. Our manuscript presents the first RCT-based cost-effectiveness analysis of TKR in addition to recommended non-surgical treatment. Due to the nature of an RCT, as opposed to the other study designs previously applied, this is therefore the first comparison of two groups that at baseline are comparable across characteristics, expect for the treatment that they were randomized to. Further-more, another strength of the RCT is that interventions and assessments were highly standardized and that the non-surgical intervention consisted of a comprehensive package of the recommended non-surgical treatments, individualized to the patient, again in opposition to previous attempts to evaluated the cost-effectiveness of TKR. This highlights the relevance and importance of our cost-effectiveness analysis.
We have addressed each of your comments, point-by-point, below and have tempered the conclusions as recommended.

Author action:
Abstract, lines 45-49: “Conclusions: From a 24-month perspective, TKR plus non-surgical treatment does not appear to be cost-effective compared to non-surgical treatment with the option of later TKR if needed in patients with moderate to severe knee osteoarthritis and moderate intensity pain in secondary care in Denmark. Results were sensitive to changes, highlighting the need for further confirmatory research also assessing the long-term cost-effectiveness of TKR.”

Lines 383-390: “TKR plus non-surgical treatment appear to be more expensive, but also more effective than non-surgical treatment after 24 months in patients with knee OA eligible for TKR and moderate intensity pain. The cost-utility analysis suggested that TKR plus non-surgical treatment was not cost-effective compared to non-surgical treatment with the option of later TKR if needed from a 24-month health system per-spective in secondary care in Denmark when adjusting for covariates and imputing missing values. Results were sensitive to changes, as the treatment was cost-effective in the unadjusted scenario, highlighting the need for further research with 5 to 10-year time horizons.”

Lines 467-472: “From a 24 months perspective in secondary care in Denmark, TKR plus non-surgical treatment does not appear to be cost-effective compared to non-surgical treatment with the option of later TKR if needed in patients with moderate to severe osteoarthritis and moderate intensity pain, eligible for TKR. However, as TKR plus non-surgical treatment was just cost-effective when not adjusting for covariates and not imputing missing values, further confirmatory studies with longer follow-up are need-ed.”

Reviewer Comment:
1) The major concern here is that the time horizon is only 2 years. Certainly one would expect the benefits of TKR to last more than 2 years, thus the current study dramatical-ly underestimates the benefits associated with TKR. The authors note this in the dis-cussion section of the manuscript, but do not account for TKR subjects continuing to accumulate QALYs as they live (many pain-free) with their joint replacement. I would expect TKR recipients’ QoL to continue to be higher beyond the 2 year timeframe. If that’s true, then TKR recipients would continue to accrue more QALYs than the non-surgical group after the end of 2 years. While at the same time, the non-TKR group will continue to accumulate costs as more subjects randomized to non-surgical treatment convert to TKR. The 2nd panel on Cost-Effectiveness in Health and Medicine recom-mends that the time horizon considered be long enough to capture all differences in costs and benefits between scenarios. I do not think that an analysis of the cost-effectiveness of TKR after 2 years is meaningful and I cannot recommend this manu-script for publication without an extended time horizon.

Author response:
Although we acknowledge that some patients will experience sustained benefits of TKR after 1 and 2 years, on a group level this is in fact not necessarily the case. Across a range of outcomes, including pain, function and quality of life (SF-36) patients under-going TKR report the largest improvements after 6-12 months with a decline in out-comes at 5 years.3 These findings are supported by the quality of life data of our study, as highlighted in the manuscript. Altogether, this questions the assumption that patients will continue to benefit from the TKR in the long-term.
We do agree, that longer-term follow-ups of this study is needed, preferably at 5 or even 10 years, which we also plan to conduct. However, the 2-year follow-up is also important as it highlights the middle-term cost-effectiveness of a highly frequent or-thopedic procedure based on two highly comparable groups that has been tested a treated in a standardized fashion.
To adapt to the comments from the reviewer, we have emphasized throughout the manuscript, that this is cost-effectiveness at 24 months (i.e. short- to middle-term), and that 5-10-year follow-ups are essential to be able to evaluate the long-term cost-effectiveness of TKR. In addition to the response and changes made based on your previous comment, we believe that this will ensure that the readers are aware of the study limitations and thereby are able to interpret the results with caution.

Author action:
Manuscript adapted throughout to reflect that this is cost-effectiveness at 24 months (i.e. short- to middle-term) and that further long-term evaluations are needed. Further-more, we have added a bullet point in the Strengths and limitations section and a sen-tence in the discussion in the limitations section:
Strengths and limitations of this study, lines 62-64: "•The 24-month time horizon and the selected population included limit conclusions on the long-term cost-effectiveness of total knee replacement and the generalizability to other populations"

Lines 458-461: "The short time horizon and the different findings in the analysis without adjustment for covariates and imputation of missing values and the sub-analysis in-cluding deaths emphasize the susceptibility of the results and highlight the need for further analyses in the field including follow-ups at 5-10 years."

Reviewer Comment:
2) Patients with severe pain (>60/100 on VAS) were excluded from the trial. The results may not be generalizable to all patients with knee OA. This is noted in the discussion but warrants further attention. What percentage of otherwise eligible knee OA patients were excluded from the trial due to severe pain? Are patients with severe pain less likely to benefit from non-operative therapy? More detail is needed so that readers can better understand this limitation.

Author response:
Evidence from more than 50 RCTs suggests that pain intensity is not associated with the effects from non-surgical treatments such as exercise4, suggesting that non-surgical treatment would also be effective in patients with more severe pain. However, we included a cut-off of 60 mm for eligibility since the surgeons felt that it was un-ethical not to offer surgery to patients in severe pain, defined as pain above 60 mm on a 0-100 VAS.
As presented in Fig 1, less than 8% of patients were excluded from the study because of a pain score of more than 60 mm when assessed for eligibility.
Furthermore, 42% of patients eligible for TKR in our trial reported pain higher than 60 mm when asked about worst pain during the previous 24 hours at baseline and 22% reported at least severe pain during activities of daily living in the previous week.
The KOOS subscales are scored from 0-100, with 0 representing extreme pain and 100 representing no pain. By definition, KOOS 50 represents on average moderate pain and KOOS 75 represents on average mild pain in the pain subscale. The mean KOOS Pain (SD) score in our study of 49.1 (15.4; 0-100, worst to best) means that 95 % of the patients had KOOS Pain scores within the range of 18.3-79.9, corresponding to severe to mild pain pre-TKR.
Pre-TKR scores are known to vary5. However, several other clinical studies evaluating pain pre-TKR (examples include 6,7,8) have reported similar pre-TKR pain scores as in our study, highlighting the comparability of our study cohort with patients previously studied in clinical practice.
We have indicated the number of patients excluded due to the pain intensity exclusion criterion,

discussed it as a potential limitation, included the selected population as a limitation in the Strengths and limitations section, compared the pain intensity to that of other TKR cohorts and highlighted the evidence on the comparable effects of exercise across different pain severities. Furthermore, we have also highlighted that the mean pain intensity in the group was moderate in both the discussion and the conclusion of the abstract and the manuscript.

Author action:
Abstract, lines 45-49: “Conclusions: From a 24-month perspective, TKR plus non-surgical treatment does not appear to be cost-effective compared to non-surgical treatment with the option of later TKR if needed in patients with moderate to severe knee osteoarthritis and moderate intensity pain in secondary care in Denmark. Results were sensitive to changes, highlighting the need for further confirmatory research also assessing the long-term cost-effectiveness of TKR.”

Strengths and limitations of this study, lines 62-64: “•The 24-month time horizon and the selected population included limit conclusions on the long-term cost-effectiveness of total knee replacement and the generalizability to other populations”

Lines 303-304: “Below 8% (n=117) of patients assessed for eligibility were excluded due to pain intensity above 60mm out of 100mm.”

Lines 383-390: “TKR plus non-surgical treatment appear to be more expensive, but also more effective than non-surgical treatment after 24 months in patients with knee OA eligible for TKR and moderate intensity pain. The cost-utility analysis suggested that TKR plus non-surgical treatment was not cost-effective compared to non-surgical treatment with the option of later TKR if needed from a 24-month health system per-spective in secondary care in Denmark when adjusting for covariates and imputing missing values. Results were sensitive to changes, as the treatment was cost-effective in the unadjusted scenario, highlighting the need for further research with 5 to 10-year time horizons.”

Lines 415-417: “In contrast, we did not find that TKR was cost-effective in addition to non-surgical treatment after 24 months in patients with moderate intensity pain.”

Lines 450-457: “Furthermore, as one of the exclusion criteria was mean pain the previ-ous week above 60 mm on a 100-mm visual analogue scale, our results might not be generalizable to patients with more severe pain at baseline. However, 42% of the pa-tients reported pain higher than 60 mm when asked about worst pain during the previ-ous 24 hours and the mean pain intensity in our trial of 49 on a 0-100 worst to best scale is comparable to a range of previous clinical studies evaluating pain severity prior to TKR6,9,10. Additionally, the effects from non-surgical treatments, such as exer-cise, does not seem to be associated with pain severity at baseline4, suggesting that the non-surgical treatment might be as effective in patients with more severe pain.”

Lines 467-472: “From a 24 months perspective in secondary care in Denmark, TKR plus non-surgical treatment does not appear to be cost-effective compared to non-surgical treatment with the option of later TKR if needed in patients with moderate to severe osteoarthritis and moderate intensity pain, eligible for TKR. However, as TKR plus non-surgical treatment was just cost-effective when not adjusting for covariates and not imputing missing values, further confirmatory studies with longer follow-up are need-ed.”

Reviewer Comment:
3) Given that almost one-third of the “non-surgical” group underwent TKR within 2 years of randomization, is it really fair to call this group non-surgical? Could it instead be framed as a non-surgical intervention with an option for delayed TKR? How were the costs and QALYs for these cross-overs incorporated?

Author response:
We have now changed the way we frame the two groups to reflect that 1/3 from the non-surgical group crossed over and had TKR during the 2-year follow-up. The data were analyzed according to the intention-to-treat principle, i.e. all patients were ana-lyzed in the group that they were randomized to. If patients crossed over to surgery, any cost associated with the surgery (and QALY) was thereby included as a cost in the non-surgical arm in the analysis. This has been added to the manuscript.

Author action:
“Non-surgical treatment alone” changed to “Non-surgical treatment with the option of later TKR if needed” throughout the manuscript.

Lines 252-253: “(…) and data were analyzed in accordance with intention-to-treat princi-ple.”

Reviewer Comment:
4) Cost-effectiveness acceptability curves could help the reader the uncertainty in the probabilistic sensitivity analyses.

Author response:
We have added two figures: One cost-effectiveness plane showing the uncertainty around the ICER and one cost-effectiveness acceptability curve.

Author action:
Figure 2: Cost-effectiveness acceptability curve and supp. figure 1: cost-effectiveness plane added.

Reviewer 2

Reviewer Comment:
Thanks to the authors for an interesting manuscript. Below my comments and sugges-tions:

Author response:
Thank you for this, we believe that the comments have helped improve the manuscript.

Reviewer Comment:
In the abstract, a distinction is made between cost-effectiveness results in the unad-justed compared to the adjusted analysis. Please specify what such adjustment con-sidered. The conclusion assumes generalisability, which I doubt can be taken for granted with a sample of 100 whose selection (before randomisation) has not even been described to this point. I would either make explicit and justify the assumption of full generalisability, or tone down the conclusion to make it specifically relevant to that group of 100 patients. In any case, the conclusion should also be limited to the geo-graphical area where the trial was conducted, or at a minimum to Denmark, considering that clinical practice and costs vary greatly amongst countries even within Europe. This applies to the respective section(s) in the manuscript.

Author response:
We have now indicated in the abstract, what we adjusted for (age, sex and baseline value). We have also toned down the conclusion in the abstract and in the main text and also in the first sentence of the discussion as suggested. Furthermore, we have included the selected population as a limitation in the Strengths and limitations section.

Author action:
Abstract, lines:39-43: “While cost-effective in the unadjusted scenario (ICER of 18,497 Euros/QALY), TKR plus non-surgical treatment was not cost-effective compared to non-surgical treatment with the option of later TKR if needed in the adjusted (age, sex and baseline values), base-case scenario (ICER of 32,611 Euros/QALY) with a proba-bility of cost-effectiveness of 23.2%.”

Abstract, lines 45-49: “Conclusions: From a 24-month perspective, TKR plus non-surgical treatment does not appear to be cost-effective compared to non-surgical treatment with the option of later TKR if needed in patients with moderate to severe knee osteoarthritis and moderate intensity pain in secondary care in Denmark. Results were sensitive to changes, highlighting the need for further confirmatory research also assessing the long-term cost-effectiveness of TKR.”

Strengths and limitations of this study, lines 62-64: “•The 24-month time horizon and the selected population included limit conclusions on the long-term cost-effectiveness of total knee replacement and the generalizability to other populations”

Lines 383-390: “TKR plus non-surgical treatment appear to be more expensive, but also more effective than non-surgical treatment after 24 months in patients with knee OA eligible for TKR and moderate intensity pain. The cost-utility analysis suggested that TKR plus non-surgical treatment was not cost-effective compared to non-surgical treatment with the option of later TKR if needed from a 24-month health system per-spective in secondary care in Denmark when adjusting for covariates and imputing missing values. Results were sensitive to changes, as the treatment was cost-effective in the unadjusted scenario, highlighting the need for further research with 5 to 10-year time horizons.”

Lines 467-472: “From a 24 months perspective in secondary care in Denmark, TKR plus non-surgical treatment does not appear to be cost-effective compared to non-surgical treatment with the option of later TKR if needed in patients with moderate to severe osteoarthritis and moderate intensity pain, eligible for TKR. However, as TKR plus non-surgical treatment was just cost-effective when not adjusting for covariates and not imputing missing values, further confirmatory studies with longer follow-up are need-ed.”

Reviewer Comment:
The term “data” is used in this paper as the aggregation of specific pieces of infor-mation, hence it is a plural noun and the verb that follows should reflect that. In some cases it does, in others it doesn’t; please correct when necessary.

Author response:
We have now changed it to plural throughout.

Author action:
Changed throughout the manuscript.

Reviewer Comment:
In the Methods section, when reporting the exclusion criteria for the trial, some context as to the significance and interpretation of the 60 mm over 100 mm VAS for mean pain the previous week is necessary. Could this be read as the trial excluding patients at the severe end of symptomatic knee OA spectrum? If so, then discussion of results and conclusions should reflect that, much beyond the comment made in the limitations section.

Author response:
Evidence from more than 50 RCTs suggests that pain intensity is not associated with the effects from non-surgical treatments such as exercise4, suggesting that non-surgical treatment would also be effective in patients with more severe pain. However, we included a cut-off of 60 mm for eligibility since the surgeons felt that it was un-ethical not to offer surgery to patients in severe pain, defined as pain above 60 mm on a 0-100 VAS.
As presented in Fig 1, less than 8% of patients were excluded from the study because of a pain score of more than 60 mm when assessed for eligibility.
Furthermore, 42% of patients eligible for TKR in our trial reported pain higher than 60 mm when asked about worst pain during the previous 24 hours at baseline and 22% reported at least severe pain during activities of daily living in the previous week.
The KOOS subscales are scored from 0-100, with 0 representing extreme pain and 100 representing no pain. By definition, KOOS 50 represents on average moderate pain and KOOS 75 represents on average mild pain in the pain subscale. The mean KOOS Pain (SD) score in our study of 49.1 (15.4; 0-100, worst to best) means that 95 % of the patients had KOOS Pain scores within the range of 18.3-79.9, corresponding to severe to mild pain pre-TKR.
Pre-TKR scores are known to vary5. However, several other clinical studies evaluating pain pre-TKR (examples include 6,7,8) have reported similar pre-TKR pain scores as in our study, highlighting the comparability of our study cohort with patients previously studied in clinical practice.
We have indicated the number of patients excluded due to the pain intensity exclusion criterion, discussed it as a potential limitation, included the selected population as a limitation in the Strengths and limitations section, compared the pain intensity to that of other TKR cohorts and highlighted the evidence on the comparable effects of exercise across different pain severities. Furthermore, we have also highlighted that the mean pain intensity in the group was moderate in both the discussion and the conclusion of the abstract and the manuscript.

Author action:
Abstract, lines 45-49: “Conclusions: From a 24-month perspective, TKR plus non-surgical treatment does not appear to be cost-effective compared to non-surgical treatment with the option of later TKR if needed in patients with moderate to severe knee osteoarthritis and moderate intensity pain in secondary care in Denmark. Results were sensitive to changes, highlighting the need for further confirmatory research also assessing the long-term cost-effectiveness of TKR.”

Strengths and limitations of this study, lines 62-64: “•The 24-month time horizon and the selected population included limit conclusions on the long-term cost-effectiveness of total knee replacement and the generalizability to other populations”

Lines 303-304: “Below 8% (n=117) of patients assessed for eligibility were excluded due to pain intensity above 60mm out of 100mm.”

Lines 383-390: “TKR plus non-surgical treatment appear to be more expensive, but also more effective than non-surgical treatment with the option of later TKR if needed after 24 months in patients with knee OA eligible for TKR and moderate intensity pain. The cost-utility analysis suggested that TKR plus non-surgical treatment was not cost-effective compared to non-surgical treatment with the option of later TKR if needed from a 24-month health system perspective in secondary care in Denmark when adjust-ing for covariates and imputing missing values. Results were sensitive to changes, as the treatment was cost-effective in the unadjusted scenario, highlighting the need for further research with 5 to 10-year time horizons.”

Lines 415-417: “In contrast, we did not find that TKR was cost-effective in addition to non-surgical treatment after 24 months in patients with moderate intensity pain.”

Lines 450-457: “Furthermore, as one of the exclusion criteria was mean pain the previ-ous week above 60 mm on a 100-mm visual analogue scale, our results might not be generalizable to patients with more severe pain at baseline. However, 42% of the pa-tients reported pain higher than 60 mm when asked about worst pain during the previ-ous 24 hours and the mean pain intensity in our trial of 49 on a 0-100 worst to best scale is comparable to a range of previous clinical studies evaluating pain severity prior to TKR6,9,10. Additionally, the effects from non-surgical treatments, such as exer-cise, does not seem to be associated with pain severity at baseline4, suggesting that the non-surgical treatment might be as effective in patients with more severe pain.”

Lines 467-472: “From a 24 months perspective in secondary care in Denmark, TKR plus non-surgical treatment does not appear to be cost-effective compared to non-surgical treatment with the option of later TKR if needed in patients with moderate to severe osteoarthritis and moderate intensity pain, eligible for TKR. However, as TKR plus non-surgical treatment was just cost-effective when not adjusting for covariates and not imputing missing values, further confirmatory studies with longer follow-up are need-ed.”

Reviewer Comment:
Please define “high-volume orthopedic specialists”.

Author response:
This has now been defined.

Author action:
Lines 147-150: “A total cemented prosthesis with patellar resurfacing (NexGen, CR-Flex, fixed bearing or LPS-Flex, fixed bearing, Zimmer, Warsaw, Indiana, USA) was inserted by high-volume orthopedic specialists (a surgeon performing +100 TKRs/year) using the surgical methods recommended by the manufacturer17.”

Reviewer Comment:
It is not clear to me why “Non protocol-driven resources, e.g. costs of recruitment, were included” if they were generated by each group. Please explain.

Author response and action:
Thank you for this comment. It should read “Non protocol-driven resources, e.g. costs of recruitment, were not included.”. This has now been changed.

Reviewer Comment:
Under “Measurement of effectiveness” you seem to be confusing the concept of QALY with that of health utility, or additional details are needed. QALYs do not have a “maxi-mum achievable” of 1 (whilst health utility does), unless it is restricted to one year. Please refine.

Author response: That is true. We have rephrased this and are now using utility and QALY consistently throughout.

Author action:
Lines 223-225: “The maximum achievable health utility is 1 and hence, a QALY value of 1 reflects one year of full health, whereas a QALY value of 0 reflects death.”

Reviewer Comment:
You indicate that “QALY values were imputed at 24 months”. I want to think this is a mistake as imputing QALYs is likely to generate greater bias than imputing only the missing health utilities. If you actually meant that it was health utilities what were imput-ed, then please correct. If you indeed imputed QALYs, then you should consider im-puting health utility values and calculate QALYs based on the observed and imputed scores.

Author response: Sorry for the lack of clarity. We did use health utilities during imputa-tion and not QALYs.

Author action:
Lines 245-247: “Missing utility values occurred at 3, 6, 12 and 24 months, and thus, utilities were imputed at these time points using utilities from available time points.”

Reviewer Comment:
I find the QALY gain calculation (formula) confusing. Why was it not done by calculat-ing area under the curve for each year, and then discounting that of year 2? If there is a good reason do follow the formula shown, please explain and justify as I am not con-vinced it is the most appropriate way to calculate QALYs in this case.

Author response We have deleted the calculation of QALYs gained, as it was difficult to understand. QALY gain was calculated as the area under the curve and this has re-placed the calculation.

Author action:
Lines 260-261: “QALY gains or losses were calculated as the area under the curve, i.e. taking changes in utility over time into account.”

Reviewer Comment:
If Table 3 excludes deaths, then what is being reported is essentially health utilities as survival is assumed for everyone. This is linked to the confusing presentation of health utilities and QALYs. Please revise to add clarity.

Author response and action:
Correct, it is health utilities. We have now indicated that in Table 3. We thank the re-viewer for highlighting the confusion concerning health utility and QALYs and hope that our clarification of the terms throughout the manuscript has helped.

Reviewer Comment:
Table 6: can you please clarify in the note/legend that the base-case analysis excludes deaths? I - however- believe it should include them.

Author response: The base-case scenario of the sub-analysis does not exclude deaths but only refers to the analysis adjusted for age, sex and baseline value. Deaths are only excluded in the primary analysis (all scenarios). To give room for a cost-effectiveness plane showing the uncertainty around the ICER and a cost-effectiveness acceptability curve, table 6 has been moved to the supplementary appendix and is now called Supp. table 2.

Reviewer Comment:
I disagree with your statement that “If TKR plus non-surgical treatment was to become cost-effective in the longer term, the decrease in QALY in the non-surgical group would need to continue.” I don’t believe that is correct. This would be most clearly shown if you report health utilities for each arm at each point. There it would become clear, I would expect, that (leaving deaths to the side for the moment, as you id in your prima-ry base-case analysis), the additional improvement achieved by the TKR group would be maintained in the time, with costs growing more rapidly for the non-surgical arm, hence leading inexorably to a reduction of the ICER until it would eventually fall under the threshold. This is likely why the two studies identified using decision-analytic mod-els over longer time horizons found the opposite of what you did. The sensitivity of results to the time horizon must be highlighted in the abstract, discussion, limitations and conclusion.

Author response: We have rephrased the section to accommodate to the comment from the reviewer. Table 3 present the health utilities at the different time points. Fur-thermore, we have highlighted the sensitivity of the results to the time horizon in all sections.

Author action:
Lines 420-429: “One could argue that extending the time horizon might have led to a different conclusion. If the positive effect of the surgery persists beyond the 24 months, TKR plus non-surgical treatment might eventually end up being a cost-effective option. Though the mean utility fluctuates slightly over time in both groups, there seems to be an overall improvement in the TKR plus non-surgical group as com-pared to non-surgical treatment only. Assuming that this between-group difference is at least maintained and a potential increased cost in the non-surgical group due to future TKR surgery, this could improve the cost-effectiveness ratios in favor of TKR plus non-surgical treatment. However, as indicated by a previous report32, improvements in symptoms might decline from 1 to 5 years after TKR, questioning the assumptions underlining a potential long-term cost-effectiveness of TKR.”

Manuscript adapted throughout to reflect that this is cost-effectiveness at 24 months (i.e. short- to middle-term) and that further long-term evaluations are needed. Further-more, we have added a bullet point in the Strengths and limitations section and a sen-tence in the discussion in the limitations section:
Strengths and limitations of this study, lines 62-64: “•The 24-month time horizon and the selected population included limit conclusions on the long-term cost-effectiveness of total knee replacement and the generalizability to other populations”

Lines 458-461: “The short time horizon and the different findings in the analysis without adjustment for covariates and imputation of missing values and the sub-analysis in-cluding deaths emphasize the susceptibility of the results and highlight the need for further analyses in the field including follow-ups at 5-10 years.”

Author action:

Reviewer Comment:
In the conclusion, you may want to emphasise that you mean “[not] imputing missing values”.

Author response: We have added ’not’ to the sentence.

Author action:
Lines 470-471: “However, as TKR plus non-surgical treatment was just cost-effective when not adjusting for covariates and not imputing missing values (…)”

Reviewer 3

Reviewer Comment:
This is a well-conducted trial investigating a clinically and economically important ques-tion. My comments are generally concerned with the presentation and interpretation of the cost-effectiveness analysis.

Author response:
Thank you for your valuable comments, which has helped us improve the manuscript further.

Reviewer Comment:
1. The cost perspective is reported as 'a limited societal perspective (i.e. health care costs and public transfer payments)'. The data on public transfer payments do not seem to be used in any cost-utility analyses, have not been valued in monetary terms (i.e. 'costs'), and appear to be available for only one third of the sample, so I am not clear as to the value of including these at all; perhaps a health system perspective only would be more appropriate?

Author response and action:
We have now changed the way we frame the perspective to a health system perspec-tive as recommended by the reviewer.

Reviewer Comment:
2. Do the registers from which cost data were extracted include the actual costs of care, or just resource use data? If the latter, how were reference costs obtained and what were these unit costs?

Author response: The National Patient Registry holds information on contacts – and linking with the Danish Case Mix System enables us to calculate costs. We have clari-fied this in the manuscript.

Author action:
Lines 197-201: “Data on inpatient and outpatient services are available from the Nation-al Patient Registry (NPR), which contains information on all kinds of patient contacts including diagnoses and diagnostic and treatment procedures. Linking data on re-source use from NPR with the Danish Case Mix System (Diagnosis-Related Groups) enabled estimation of associated costs.”

Reviewer Comment:
3. While the NICE threshold is a useful reference, there seems no particular reason to believe that it is the appropriate threshold for use in Denmark (or any other context in which these results may be interpreted). Presenting the sensitivity of the cost-effectiveness finding to the choice of threshold (i.e. a cost-effectiveness acceptability curve across a range of threshold values) would improve the interpretation of results as compared to using the single NICE decision threshold.

Author response: This is a good suggestion. We have added a figure showing the probability of cost-effectiveness at different thresholds and a cost-effectiveness plane illustrating the uncertainty around the ICER.

Author action:
Figure 2 and supp. figure 1 added.
Lines 348-354: “Cost-effectiveness acceptability curve showing the probability of TKR plus non-

surgical treatment being cost-effective at different thresholds is presented in Figure 2. The probability of cost-effectiveness was below 60% up until a threshold of approx. 40,000 Euros/QALY. To reach a probability of cost-effectiveness greater than 90%, a threshold of minimum 60,000 Euros/QALY was needed.

Cost-effectiveness plane illustrating the uncertainty around the ICER is presented in supp. figure 1."

Reviewer Comment:
4. I was confused by the terminology regarding the measurement of health gains over time. The values 'QALY baseline', 'QALY 3 months', etc on the RHS of the equation on line 261 (and likewise in Tables 3 and 5) are presumably the instantaneous HRQoL utili-ty values as measured by the EQ-5D, not QALYs (which are a cumulative measure of HRQoL experienced over time)
If this is the case, I don't think the statements regarding potential cost-effectiveness over a longer time horizon follow from the observed data as suggested in lines 377-387. The gain in HRQoL of TKR compared to non-surgical treatment was 0.09 at 3 months, 0.133 at 6 months, 0.085 at 12 months, and 0.142 at 24 months (0.139 with discounting), if I have understood Table 3 correctly. If anything, this seems to be in-creasing over time. Meanwhile the costs of TKR are all incurred in the first year, so with a longer time horizon, we would see increasing QALY gains with no increase in incre-mental costs, so improving cost-effectiveness ratios over time.

Author response: Thank you for clarifying this. We have corrected the terminology to ensure that health utilities and QALYs are used appropriately. We have also rephrased the section in the discussion to accommodate to the comment from the reviewer. Table 3 present the health utilities at the different time points as does Table 5 (now termed supp. table 1). Furthermore, we have highlighted the sensitivity of the results to the time horizon.

Author action:
Lines 420-429: "One could argue that extending the time horizon might have led to a different conclusion. If the positive effect of the surgery persists beyond the 24 months, TKR plus non-surgical treatment might eventually end up being a cost-effective option. Though the mean utility fluctuates slightly over time in both groups, there seems to be an overall improvement in the TKR plus non-surgical group as com-pared to non-surgical treatment only. Assuming that this between-group difference is at least maintained and a potential increased cost in the non-surgical group due to future TKR surgery, this could improve the cost-effectiveness ratios in favor of TKR plus non-surgical treatment. However, as indicated by a previous report32, improvements in symptoms might decline from 1 to 5 years after TKR, questioning the assumptions underlining a potential long-term cost-effectiveness of TKR."

Manuscript adapted throughout to reflect that this is cost-effectiveness at 24 months (i.e. short- to middle-term) and that further long-term evaluations are needed. Further-more, we have added a bullet point in the Strengths and limitations section and a sen-tence in the discussion in the limitations section:
Strengths and limitations of this study, lines 62-64: "•The 24-month time horizon and the selected population included limit conclusions on the long-term cost-effectiveness of total knee replacement and the generalizability to other populations"

Lines 458-461: "The short time horizon and the different findings in the analysis without adjustment for covariates and imputation of missing values and the sub-analysis in-cluding deaths emphasize the susceptibility of the results and highlight the need for further analyses in the field including follow-ups at 5-10 years."

Reviewer Comment:
5. The method of multiple imputation should be described. In addition, it was reported that 'missing values occurred on utilities at 24 months' - were these the only missing data, or was multiple imputation also used for other variables?

Author response: Sorry for the lack of clarity. A few health utilities were missing at all time points. Utility values from available time points were used for imputation. Cost data was available for all

patients at all time points and no imputation was therefore needed. No other imputation took place. We used the MiAnalyse method in SAS to imputate and pool the regressions. More information has now been provided in the manuscript.

Author action:
Lines 241-247: “Missing data were handled by using multiple imputation, which enables individuals with incomplete data to be included in the analysis. The underlying assump-tion when using multiple imputation is that data are missing at random, i.e. the proba-bility of missing values is not dependent on unobserved data. Missing data were im-puted using multiple imputation in SAS, and the assumption of data missing at random was also tested and confirmed in SAS. Missing utility values occurred at 3, 6, 12 and 24 months, and thus, utilities were imputed at these time points using utilities from available time points.”

Reviewer Comment:
6. Given the large number of cross-overs, the true effect (and cost) of TKR is likely to be understated. While this is likely unavoidable, given the ethical issues with denying TKR to those for whom it is considered necessary, it is an important limitation that needs to be acknowledged. A secondary per-protocol or compliers analysis might be informative to (partially) address this.

Author response:
We have added this as a limitation in the discussion and in the Strengths and limita-tions section. Furthermore, we have now changed the way we frame the two groups to reflect that 1/3 from the non-surgical group crossed over and had TKR during the 2-year follow-up.
We do agree, that a per-protocol analyses would be relevant in future studies of this cohort. Due to Danish regulations on accessibility to the national registries, we are not able to conduct such analysis, unless we start all over, applying for permission to use the data, uploading the data to Statistics Denmark, and re-run all analyses.

Author action:
Strengths and limitations of this study, lines 65-67: “• Since nearly 1 out of 3 from the non-surgical group had TKR surgery during the 24 months, it is likely that the true addi-tional effect and cost of TKR in addition to non-surgical treatment have been underes-timated in the study.

“Non-surgical treatment alone” changed to “Non-surgical treatment with the option of later TKR if needed” throughout the manuscript.”

Lines 448-450: “As 32% from the non-surgical group had TKR surgery during the 24 months, it is likely that the true additional effect and cost of TKR have been underesti-mated in the study.”

REFERENCES
1. Kamaruzaman H, Kinghorn P, Oppong R. Cost-effectiveness of surgical interventions for the management of osteoarthritis: a systematic review of the literature. BMC Musculoskelet Disord 2017;18:183. doi: 10.1186/s12891-017-1540-2
2. Ferket BS, Feldman Z, Zhou J, Oei EH, Bierma-Zeinstra SM, Mazumdar M. Impact of total knee replacement practice: cost effectiveness analysis of data from the Osteoarthritis Initiative. Bmj 2017;356:j1131. doi: 10.1136/bmj.j1131
3. Nilsdotter AK, Toksvig-Larsen S, Roos EM. A 5 year prospective study of patient-relevant outcomes after total knee replacement. Osteoarthritis and cartilage 2009;17:601-6. doi: 10.1016/j.joca.2008.11.007; 10.1016/j.joca.2008.11.007
4. Juhl C, Christensen R, Roos EM, Zhang W, Lund H. Impact of exercise type and dose on pain and disability in knee osteoarthritis: A systematic review and meta-regression analysis of randomized controlled trials. Arthritis Rheumatol 2014;66:622-36. doi: 10.1002/art.38290;
5. Ackerman IN, Dieppe PA, March LM, Roos EM, Nilsdotter AK, Brown GC, et al. Variation in age and physical status prior to total knee and hip replacement surgery: a comparison of centers in Australia and Europe. Arthritis and Rheumatism 2009;61:166-73. doi: 10.1002/art.24215; 10.1002/art.24215

6. Gossec L, Paternotte S, Maillefert JF, Combescure C, Conaghan PG, Davis AM, et al. The role of pain and functional impairment in the decision to recommend total joint replacement in hip and knee osteoarthritis: an international cross-sectional study of 1909 patients. Report of the OARSI-OMERACT Task Force on total joint replacement. Osteoarthritis and cartilage 2011;19:147-54. doi: 10.1016/j.joca.2010.10.025
7. Terwee CB, van der Slikke RM, van Lummel RC, Benink RJ, Meijers WG, de Vet HC. Self-reported physical functioning was more influenced by pain than performance-based physical functioning in knee-osteoarthritis patients. J Clin Epidemiol 2006;59:724-31. doi: 10.1016/j.jclinepi.2005.11.019
8. Escobar A, Quintana JM, Bilbao A, Azkarate J, Guenaga JI, Arenaza JC, et al. Effect of patient characteristics on reported outcomes after total knee replacement. Rheumatology (Oxford, England) 2007;46:112-9. doi: kel184 [pii]
9. Keurentjes JC, Fiocco M, So-Osman C, Onstenk R, Koopman-Van Gemert AW, Poll RG, et al. Patients with severe radiographic osteoarthritis have a better prognosis in physical functioning after hip and knee replacement: a cohort-study. PLoS One 2013;8:e59500. doi: 10.1371/journal.pone.0059500
10. Wise BL, Niu J, Felson DT, Hietpas J, Sadosky A, Torner J, et al. Functional Impairment Is a Risk Factor for Knee Replacement in the Multicenter Osteoarthritis Study. Clinical orthopaedics and related research 2015;473:2505-13. doi: 10.1007/s11999-015-4211-3 [doi]

VERSION 2 – REVIEW

REVIEWER	Jamie Collins Brigham and Women's Hospital
REVIEW RETURNED	27-Nov-2019

GENERAL COMMENTS	Thank you for the careful and thoughtful response to review. I am satisfied with how the updated manuscript addresses the study limitations.

REVIEWER	Rafael Pinedo-Villanueva University of Oxford, UK
REVIEW RETURNED	23-Dec-2019

GENERAL COMMENTS	The authors have satisfactorily addressed my comments to the previous version as well as those of other reviewers, I believe. There remains a question as to whether the added emphasis about cost-effectiveness results being reported at 24 months is sufficient to counterbalance the need to conduct analyses of what is likely end-stage OA and total knee replacements over a time horizon shorter than lifetime considering the chronic nature of the disease and the likelihood of relevant events (such as revision surgeries) happening within not 24 months bur rather 10 to 20 years. Conclusions may hence be misleading if not read carefully, but the authors do stress this limitation and provide useful insights into findings at of the trial at 24 months.

REVIEWER	Ross Wilson University of Otago, New Zealand I collaborate with Prof Skou on an unrelated research grant funding application.
REVIEW RETURNED	04-Dec-2019

GENERAL COMMENTS	Thank you for giving me the opportunity to review this revised submission. The changes the authors have made have addressed my concerns and improved the clarity of the manuscript.