PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Effectiveness and cost-utility of a multifaceted eHealth strategy to improve back pain beliefs of patients with nonspecific low back pain: a cluster randomised trial
AUTHORS	Suman, Arnela; Schaafsma, F; van Dongen, Johanna M.; Elders, Petra; Buchbinder, Rachelle; Tulder, Maurits; Anema, Johannes

VERSION 1 – REVIEW

REVIEWER	David Blanco Universitat Politècnica de Catalunya
REVIEW RETURNED	11-Apr-2019

GENERAL COMMENTS	This report shows the results of an evaluation of the consistency between the CONSORT checklist you submitted and the information that was reported in the manuscript. The examples or cites included in the report were extracted from the CONSORT E&E Document (https://www.bmj.com/content/340/bmj.c869), the CONSORT extension for cluster trials (https://www.bmj.com/content/bmj/345/bmj.e5661), and the CONSORT extension for stepped wedge randomised cluster trials (https://www.bmj.com/content/bmj/363/bmj.k1614.full.pdf). Please, make the following revisions: • For CONSORT Item 9a ("*Mechanism used to implement the random allocation sequence, describing any steps taken to conceal the sequence until interventions were assigned*") and its extension to cluster trials ("*Specification that allocation was based on clusters rather than individuals and whether allocation concealment (if any) was at the cluster level, the individual participant level or both*"), please explain (i) how the allocation system was set up so that the person enrolling participants did not know in advance which treatment the next cluster was going to get, (ii) whether concealment was at the cluster level, the participant level,

	or both, and (iii) please specify that allocation was based on cluster rather than individuals. • For CONSORT Item 11a ("*If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how*") please report the blinding status of the different parties involved in the study (e.g. participants, outcome assessors…). You could include this information in the *Randomisation* subsection and rename it as *Randomisation and blinding.* • For CONSORT Item 13a ("*For each group, the numbers of participants who were randomly assigned, received intended treatment and were analysed for the primary outcome*") and its extension for cluster trials ("*For each group, the numbers of clusters that were randomly assigned, received intended treatment, and were analysed for the primary outcome*"), please consider redoing the flow diagram in order to capture the nuances of a stepped wedged cluster randomised cluster trials according to the flow diagram shown in Fig. 4 of the CONSORT extension for these trials (https://www.bmj.com/content/bmj/363/bmj.k1614.full.pdf). As your flow diagram stands now, it seems to be a two arm parallel trial and it is difficult for the readers to understand what means that a participant is randomised to the control or the intervention group as all clusters get are at some point the control and the intervention treatments. Please, include time periods in the flow diagram – your figure 1 mentions T0, T1, T2, T3 and T4 but it is not clear what these correspond to. In summary, please to comply with the structure and details proposed in the flow diagram of the extension for stepped wedge randomised trials. A precise, clear and transparent flow diagram is crucial to understand the study. • For CONSORT Item 13b ("*For each group, losses and exclusions after randomisation, together with reasons*"), please include in the flow diagram I previously proposed the number of lost to follow-up participants for each sequence

	and each cluster and provide the reasons why this happened. ○ An example of adequate reporting of losses and exclusions after randomisation can be found in Fig. 3 of the CONSORT E&E document (http://www.consort-statement.org/Media/Default/Downloads/CONSORT%202010%20Explanation%20and%20Elaboration%20Document-BMJ.pdf) • For CONSORT Item 17a ("*For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)*"), please include in Table 2 the effect size and its confidence interval for all study outcomes and each evaluation time points. Also, consider merging Table 2 and 3 so that it is easier to compare these effect sizes to the adjusted ones. Please, remember to include the effect sizes for the outcome "quality of life", which is shown in Table 2 but not in Table 3. Furthermore, provide a coefficient of intracluster correlation (ICC or k) for the primary outcome.

REVIEWER	Allan Riis Aalborg University
REVIEW RETURNED	17-Apr-2019

GENERAL COMMENTS	Thank you for the opportunity to review this relevant and interesting study. I have some suggestions to the authors. My comments are listed below in the order, they appear in the manuscript. 1. In the abstract, the sentence: '779 patients participated in this study, of which 331 were randomised to the intervention group (multifaceted eHealth strategy), and 448 were randomised to the control group (usual care)', is better placed in the result-section of the abstract. 2. Even though it is indirectly mentioned in the aim of the abstract, the authors should mention the primary outcome: 'back pain beliefs' in the methods-section of the abstract'. . 3. I miss a sentence about blinding in the methods-section of the abstract. 4. The background section introduces the purposes and aim of this study very well.

	5. A number of exclusion criteria are listed in the methods-section of manuscript. However, all are listed as serious comorbidities! I think “confirmed pregnancy” needs to be written in a separate sentence. 6. At page 10, line 183, the secondary outcomes are mentioned. However, patients’ levels of pain is not found in the paper. Pain was included as an outcome in the protocol and consequently leaving pain out in the reporting should be commented upon in this paper. 7. In the discussion, the authors state that results needs to be interpreted with caution because of a higher degree of drop out in the intervention group. Is it possible to investigate whether patients dropping out in the study were different in baseline characteristics than patients followed up? Furthermore, is it possible to explore differences in baseline characteristics among patients lost to follow-up in the intervention group and patients lost to follow-up in the control group? This could lead to selection bias. For instance, if patients dropping out in the intervention group were scoring better at baseline than patients dropping out of the control group. In this case, this could lead to an underestimation of the effect of the intervention. 8. At page 22, line 357 the manuscript reads: ‘This is in line with a very similar recent implementation study for the management of LBP. In that study, patients in the intervention group had higher LBP-related costs for inpatient secondary care’. Do the author mean: ‘This in contrast to a very similar..’? The study referred to [45] found ‘Results showed that costs associated with primary health care were higher, whereas secondary health care costs were lower for the intervention group when compared with the control group’. I acknowledge the great effort this implementation study has required and hope you will find my comments helpful.

REVIEWER	Donald Murphy Alpert Medical School of Brown University
REVIEW RETURNED	02-May-2019

GENERAL COMMENTS	A job well done. I am concern about the issues related to the low disability level, low absenteeism level and high drop out. This significantly decrease the usefulness of the study. However, the authors do acknowledge those.

REVIEWER	J'W Geurts Rijnstate, the Netherlands
REVIEW RETURNED	25-Jun-2019

GENERAL COMMENTS	This is a (cost)effectiveness study in which a e-health intervention is evaluated with the primary outcome questionnaire ' back pain beliefs' in a primary care setting. This paper is well written and well evaluated statistically. I compliment the authors. I have only minor issues and few suggestions to make. The program seems to stress compliance. No suggestions were made to improve compliance to the intervention or program or implementation. Also, the e-health intervention is not explained in

	this paper, it would be nice if you could spend a few lines in this paper about this program so the reader can understand the compliance issues. A minor issue is that this study is impossible to repeat for other institutions because of lacking access to the e-health program used in this study. This should be discussed in the discussion part. The abstract could contain some information mentioned in the results about the patient population i.e. ↑QOL and physical function, and in the discussion the suggestion that this program should be evaluated in a back pain population with lower health states. That is, if you think the compliance problems could be dealt with.

REVIEWER	Sarah Paganini University of Freiburg, Germany
REVIEW RETURNED	09-Jul-2019

GENERAL COMMENTS	Comments to authors The present manuscript aims to investigate the effectiveness and cost-effectiveness of multifaceted eHealth strategy aiming at improving back pain beliefs and disability in comparison with usual care for back pain patients. Data were drawn from 779 patients recruited in general or ohysiotherapy practices. This evaluation of multifaceted eHealth is of importance with regard to shortages in our health care systems and the high disability for individuals suffering from low back pain. The strength of this study is the implementation of an eHelath strategy in routine care. However there are several limitations and not all information is provided (according to the CONSORT statement and CHEERS guidelines). One major shortcoming is that a cost-effectiveness analysis and a cost-utility analysis are stated, but only a cost-utility analysis (CUA) is performed. This CUA is not discussed in the discussion section and there are contradicting statements. The listed shortcomings should be clarified. The manuscript would benefit from English proof reading. Major Comments Titel: 1. In the paper it is stated that a cost-effectiveness and a cost-utility analysis will be performed. However, only a cost-utility analysis is provided. This should be clearly mentioned in the title (“cost-utility” instead of “cost-effectiveness analysis”) Abstract: 1. Please provide information of inclusion criteria 2. In the methods section there is no information about the outcomes, assessment and statistical methods 3. There is no information about the cost-effectiveness analysis 4. Results: Statistical information for the main results and the exact costs should be provided. No quality of life measures/results are given 5. Line 50: Results: Why did 37% of the participants did not have back pain at baseline (an inclusion criterium)? This sentence needs further explanation. The guidelines of the CONSORT Statement and the CHEERS Guidelines should be followed Background:

	1. Line 93-94: There is a distinction between “indirect costs due to absenteeism” and “productivity losses due to disability”. All of those costs are indirect costs. There should be a clarification of this distinction and a definition of indirect costs in this paper. 2. Line 110-111: A reference to interventions that specifically aim at (back) pain would be helpful. 3. The rational for the multifaceted strategy of the eHealth program should be mentioned. Methods: 1. It is not stated, how exclusion criteria were assessed 2. There is no information from when until when participants were recruited. 3. The design of the stepped-wedge cluster randomised controlled trial and the procedure should be described in more detail. Figure 1 seems not to be enough for clarification. 4. There is no information about blinding/masking 5. Line 162: How was this continuing medical education operationalized? 6. Line 167: It is not clear, what the professional based intervention is. This should be clarified for the reader, even though there is a study protocol with further information. 7. Line 170: Is the description correct? “The BBQ is designed to measure the inevitable consequences of LBP”. Or does it measure the “belief” about these consequences? 8. The sample size seems quite low. Could you provide further information, how you applied the ICC in your calculation (not necessarily in the manuscript. Only for clarification)? 9. Line 184-186: Validity and reliability criteria for the RDQ-24 and the EQ-5D-3L is missing. 10. Was there an assessment of negative effects of the treatment (“harms” in the COSNORT statement)? 11. Line 189: The correct name of the TIC-P is “Trimbos/iMTA questionnaire for Costs associated with Psychiatric Illness”. The time period of the TIC-P should be mentioned (last three months) 12. Line 203, 204: A reference is missing. What is meant by “professional organisations”? 13. Line 222: What is the rational for imputing data separately for the intervention and control group? 14. Line 228: Information should be provided of how the data/results of the different imputation data sets were aggregated and how many imputations have been done. 15. Line 230: For clarification: Were there different imputations for the effectiveness and the cost-effectiveness analysis and when yes, why? 16. The main outcome(s) for the CEA and the CUA should be clarified. 17. In the study protocol a budget impact analysis was planned. If this was not done it should be stated as exception to the protocol. 18. In the CHEERS checklist it is stated that “the choice of model” was described in page 10-12. As this was a CUA alongside a clinical trial it can be assumed that no decision- analytical model was specified. 19. No information about the discount rate is given (even if data were not discounted this should be stated; see CHEERS guidelines). 20. Currency and price date should be stated (see CHHEERS guidelines).

	21. It should be mentioned whether the trial followed the CHEERS guidelines. Results: 1. For clarification: The only inclusion criterium was diagnosed low back pain? How did you test for the exclusion criteria (e.g. psychiatric disorder)? How many patients were excluded due to specific exclusion criteria? Reasons for exclusion should be listed precisely in figure 2 (according to the CONSORT guidelines). 2. Table 1: • It should be clarified, why there is no complete baseline information for all study participants (n varies for each baseline characteristic) • Please state why only 63% suffered from back pain at baseline, as this was the only inclusion criterium. • Please clarify, how the categories of educational level were defined • Please clarify, how “activity” was defined. How was physical activity and physical demanding work assessed? 3. Table 2 and 3: • For better readability mean values and effects should be presented in one table • The abbreviations should be defined (M, F, CI) • What is the rational for making different adjustments for back pain beliefs and disability? • It is not clear why there is a separate analysis for men and women for disability and not for back pain beliefs • It is not clear whether the results refer to imputed data or not 4. It should be stated if missing data differed significantly between groups 5. Table 4: • Abbreviations and “Δ” should be clarified • Why are SEMs reported and not SD? • Is “unpaid productivity” the same as “Informal care”? The same wording should be used or both categories should be mentioned and described earlier. • It would be even more informative, if the cost categories would be presented in more detail (what kind of primary, secondary, alternative care was used? What kind of medication? Only back pain medication or other as well? • How are the total societal costs calculated? Adding all costs results in higher costs. 6. Mean values of the QALYs in each group should be stated. The mean QALYs at baseline should also be stated. If they differ, there should be adjustment for baseline QALYs as well 7. Line 293: Negative ICERs should not be interpreted. If there are negative costs the ICER gets more and more negative with smaller QALY health gains, suggesting a high amount of money saved. A better description is that “the intervention dominated standard care”. However, it has to be pointed out that the QALY gain was very low and not statistically significant. 8. CI of the ICER should be mentioned 9. Line 295: Is 79% the correct value? On the Cost-effectiveness plane (figure 3) it seems like there are less cost effect pairs in the South East quadrant.

	10. Line 297: It should be clarified from which data this statement comes from (“The uncertainty around the cost-effectiveness estimate was large”) 11. There are only results for a CUA. In the method section a CEA was mentioned as well (no outcome was defined). This should urgently be clarified. 12. The CUA calculations should be checked thoroughly. There are cost savings but there is no significant difference for QALYs and very small QALY health gains for the intervention group. Therefore, it is surprising, that the CEAC reveals such good probabilities of being cost-effective. Discussion: 1. Line 319: Is a BBQ score of 26.5 (on a scale from 9 to 45) meaningfully higher than in the current study (24.7). It is questionable if this could be a possible explanation for the results. 2. Line 323-324: The result of the study (attitudes of elderly) is not discussed fully. How can this study results explain the current results? 3. There is literature on how adherence can be improved in eHealth strategies. There is also literature on self-help interventions that are more structured and guided (and effective). This could be included in the discussion of the question: What do LBP patients need to improve their beliefs and disability? 4. Line 370: It is stated that no cost-effectiveness is yielded (again in line 396). In line 306 it was stated that “The probability of cost-effectiveness was high”. These are contradicting statements. The results of the cost-utility analysis are not discussed. This is urgently needed. Why is there such a good probability of being cost-effective for an intervention that yields no significant QALY health gains? What could be the reason for the high cost savings in the intervention group for absenteeism and presenteeism, when adherence to the intervention was not high and no difference in other outcomes could be found? 5. Limitations for the cost-utility analysis should be stated (e.g. concerning power) 6. Implications for future research are that participants with LBP should be included in further studies. As this was the aim of the current trail it should be critically discussed why baseline assessment was not only done for individuals that suffer from back pain. Minor Comments Abstract: - Line 41, 42: Please check the syntax of the sentence: “Four clusters of general and physiotherapy practices and occupational physicians were randomised…” - Line 106: Do “economic” and “societal burden” mean different aspects? Background: -Line 92: To clarify the time horizon for these costs “per year” could be added. Methods: - The reference for the BBQ is missing.

	Results: 1. The wording should be consistent (follow-up vs. follow up) 2. Table 1: • "(SD)" should be added after "Back pain beliefs", "Disability", "absenteeism" and "Quality of life" • It should be mentioned, what the first information stands for (sometimes "mean", sometimes "number of participants"?) 3. Table 3: "Back beliefs" □□"Back pain beliefs" 4. Line 290 and 292: abbreviations should be clarified 5. Figure 3: • In the caption there should be more precise information (imputed data, bootstrapping, QALYs based on EQ-5D-3L). • There is no labelling for the x-axis. The labelling for the y-axis is not correct • It would be helpful to see more of the whole plane (all 4 quadrants) 6. Figure 4: • In the caption there should be more precise information (what means "CEAC", imputed data, bootstrapping, QALYs based on EQ-5D-3L). • The labelling for the x-axis is not correct (€) Discussion 1. Line 317: Wording ("back beliefs") 2. Line 332: This information should be stated in the results chapter 3. Line 334-335: Grammar of the sentence should be checked

REVIEWER	Jesse Kigozi University of Birmingham, United Kingdom
REVIEW RETURNED	06-Aug-2019

GENERAL COMMENTS	This is a well-written paper, but with comments below that would need addressing or commenting on. 1. The authors did not perform any sensitivity analysis. It would be good to see at least a sensitivity analysis using the complete-case to assess the impact of the missing data. 2. How did the authors determine all the variables that were included in the imputation model? How was the imputation done? At what level was the imputation done? More information is needed to know what was actually done. Also, did the imputation model take into consideration the effect of clustering? Similarly, was clustering taken into consideration for the models used to generate QALY and cost differences? 3. How were the intervention costs actually calculated? This detail was not included and it's not clear how these costs were calculated and apportioned to the patients? 4. Unit costs – could the authors include a table of the unit costs that have been used to estimate the costs and the corresponding resource in detail. This can be included in the appendix if not available but would need to be seen. 5. The QALY scores are currently reported rounded to 2 decimal places – need to extend these to at least 3. 6. Need to include how long resource use data were collected for at each time point in the text (direct and indirect costs) 7. Need to include a statement about the need or not for discounting in the text.

	8. What threshold have the authors used to determine a lack of cost-effectiveness – for most of the WTP values at least over 20,000 there is an over 80% chance of being cost-effective? 9. In the discussion, the authors report that the probability of the intervention being cost-effective was high but then go on to conclude that it was not cost-effective? What informed this conclusion?

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1
Reviewer Name: David Blanco
Institution and Country: Universitat Politècnica de Catalunya
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
This report shows the results of an evaluation of the consistency between the CONSORT checklist you submitted and the information that was reported in the manuscript. The examples or cites included in the report were extracted from the CONSORT E&E Document (https://www.bmj.com/content/340/bmj.c869), the CONSORT extension for cluster trials (https://www.bmj.com/content/bmj/345/bmj.e5661), and the CONSORT extension for stepped wedge randomised cluster trials (https://www.bmj.com/content/bmj/363/bmj.k1614.full.pdf).
Please, make the following revisions:
• For CONSORT Item 9a ("Mechanism used to implement the random allocation sequence, describing any steps taken to conceal the sequence until interventions were assigned") and its extension to cluster trials ("Specification that allocation was based on clusters rather than individuals and whether allocation concealment (if any) was at the cluster level, the individual participant level or both"), please explain (i) how the allocation system was set up so that the person enrolling participants did not know in advance which treatment the next cluster was going to get, (ii) whether concealment was at the cluster level, the participant level, or both, and (iii) please specify that allocation was based on cluster rather than individuals.
(i) We have now stated that "Randomisation was performed by means of computer-generated allocation, using specific software " (Methods section, paragraph 'Randomisation '). I.e. a computer software programme generated the allocation and sequence of clusters, and staff had no influence on this process.
(ii) + (iii) We have now added the following statement to the Methods section, paragraph 'Randomisation ': " randomisation and allocation were performed on cluster level. However, patients were blinded and not aware of group allocation, and thus concealment was on individual level. "

• For CONSORT Item 11a ("If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how") please report the blinding status of the different parties involved in the study (e.g. participants, outcome assessors…). You could include this information in the Randomisation subsection and rename it as Randomisation and blinding.
We have now added the following statement to the Methods section, paragraph 'Randomisation ': "Outcome assessors were blinded to individual patient allocation. " This is in addition to our previous addition stating that patients were blinded and not aware of group allocation.

• For CONSORT Item 13a (“For each group, the numbers of participants who were randomly assigned, received intended treatment and were analysed for the primary outcome”) and its extension for cluster trials (“For each group, the numbers of clusters that were randomly assigned, received intended treatment, and were analysed for the primary

outcome”), please consider redoing the flow diagram in order to capture the nuances of a stepped wedged cluster randomised cluster trials according to the flow diagram shown in Fig. 4 of the CONSORT extension for these trials (https://www.bmj.com/content/bmj/363/bmj.k1614.full.pdf). As your flow diagram stands now, it seems to be a two arm parallel trial and it is difficult for the readers to understand what means that a participant is randomised to the control or the intervention group as all clusters get are at some point the control and the intervention treatments. Please, include time periods in the flow diagram – your figure 1 mentions T0, T1, T2, T3 and T4 but it is not clear what these correspond to. In summary, please to comply with the structure and details proposed in the flow diagram of the extension for stepped wedge randomised trials. A precise, clear and transparent flow diagram is crucial to understand the study.
While the larger trial was in fact a cluster randomised trial, the patient study we are reporting on de facto was a parallel arm trial: patients were either in the intervention group or in the control group. Patients did not cross over to the other group and remained in their allocation even though their healthcare providers at some point received the professional-based
intervention (which is what the ‘cluster ’ in our study refers to). Therefore, the f lowchart depicted is a correct representation of the participant flow in this study.
We have now added specific time-points in Figure 1, but we wish to emphasize that these time points only refer to the time-points where the clusters received interventions and patients were recruited, but are not related to patient follow-up measurements.

• For CONSORT Item 13b (“For each group, losses and exclusions after randomisation, together with reasons”), please include in the flow diagram I previously proposed the number of lost to follow-up participants for each sequence and each cluster and provide the reasons why this happened.
o An example of adequate reporting of losses and exclusions after randomisation can be found in Fig. 3 of the CONSORT E&E document (http://www.consort-statement.org/Media/Default/Downloads/CONSORT%202010%20Explanation%20and%20Elaboration%20Document-BMJ.pdf)
Our flow chart (Figure 2) depicts the numbers of patients that completed the study, and thus the number of exclusions. Patients were allowed to quit participation without reasons, and reasons for loss-to-follow up are therefore not available. As stated, patients were followed individually and not on cluster level.

• For CONSORT Item 17a (“For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)”), please include in Table 2 the effect size and its confidence interval for all study outcomes and each evaluation time points. Also, consider merging Table 2 and 3 so that it is easier to compare these effect sizes to the adjusted ones. Please, remember to include the effect sizes for the outcome “quality of life”, which is shown in Table 2 but not in Table 3. Furthermore, provide a coefficient of intracluster correlation (ICC or k) for the primary outcome.
Effect sizes and confidence intervals are included for each study outcome in Table 3. We have used multilevel mixed-models statistical analyses to account for clustering effects in the analyses. Therefore, no ICC is provided. Quality of life is separate from the outcome measures, as it was used for the economic evaluation (QALYs), and therefore is presented as utility values and in the QALYs.

Reviewer: 2
Reviewer Name: Allan Riis

Institution and Country: Aalborg University
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
Thank you for the opportunity to review this relevant and interesting study. I have some suggestions to the authors. My comments are listed below in the order, they appear in the manuscript.

1. In the abstract, the sentence: ‘779 patients participated in this study, of which 331 were randomised to the intervention group (multifaceted eHealth strategy), and 448 were randomised to the control group (usual care)’, is better placed in the result-section of the abstract.
Thank you for your suggestion. We have now restructured the abstract completely to be more aligned with the headings of a structured abstract.

2. Even though it is indirectly mentioned in the aim of the abstract, the authors should mention the primary outcome: ‘back pain beliefs’ in the methods-section of the abstract’. We have now restructured the abstract and included this.

3. I miss a sentence about blinding in the methods-section of the abstract.
We have now restructured the abstract and included this.

4. The background section introduces the purposes and aim of this study very well.
Thank you!

5. A number of exclusion criteria are listed in the methods-section of manuscript. However, all are listed as serious comorbidities! I think “confirmed pregnancy” needs to be written in a separate sentence.
Thank you for your suggestion, we have taken pregnancy out of the list of serious comorbidities and written it in a separate sentence.

6. At page 10, line 183, the secondary outcomes are mentioned. However, patients’ levels of pain is not found in the paper. Pain was included as an outcome in the protocol and consequently leaving pain out in the reporting should be commented upon in this paper. Thank you for raising this important difference. We have now added information about why pain was no longer measured and reported in the Methods section.

7. In the discussion, the authors state that results needs to be interpreted with caution because of a higher degree of drop out in the intervention group. Is it possible to investigate whether patients dropping out in the study were different in baseline characteristics than patients followed up? Furthermore, is it possible to explore differences in baseline characteristics among patients lost to follow-up in the intervention group and patients lost to follow-up in the control group? This could lead to selection bias. For instance, if patients dropping out in the intervention group were scoring better at baseline than patients dropping out of the control group. In this case, this could lead to an underestimation of the effect of the intervention.

This is an interesting question. We have compared the patients that completed the study and the patients that were lost to follow-up in both the intervention and control groups. In both groups, patients that completed the study were more likely to have a high educational level compared to patients who were lost to follow up. Additionally, in the intervention group, patients that completed the study were more likely to not be employed (i.e. involved in paid work) than patients who were lost to follow-up.
We have now stated this in the Discussion
section, paragraph ‘Study limitations’.

8. At page 22, line 357 the manuscript reads: ‘This is in line with a very similar recent implementation study for the management of LBP. In that study, patients in the intervention group had higher LBP-related costs for inpatient secondary care’. Do the author mean: ‘This in contrast to a very similar..’? The study referred to [45] found ‘Results showed that costs associated with primary health care were higher, whereas secondary health care costs were lower for the intervention group when compared with the control group’.
Thank you for noticing this error, we have corrected it.

I acknowledge the great effort this implementation study has required and hope you will find my comments helpful.
Thank you for your helpful comments!

Reviewer: 3
Reviewer Name: Donald Murphy
Institution and Country: Alpert Medical School of Brown University
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
A job well done. I am concern about the issues related to the low disability level, low absenteeism level and high drop out. This significantly decrease the usefulness of the study. However, the authors do acknowledge those.
Thank you for your comments. We do agree that these issues pose threats to our study, and we have tried to discuss this to the best possible extent. We thank you for agreeing with us on this!

Reviewer: 4
Reviewer Name: J'W Geurts
Institution and Country: Rijnstate, the Netherlands
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
This is a (cost)effectiveness study in which a e-health intervention is evaluated with the primary outcome questionnaire ' back pain beliefs' in a primary care setting.
This paper is well written and well evaluated statistically. I compliment the authors. I have only minor issues and few suggestions to make.
Thank you for your compliments on our work!

The program seems to stress compliance. No suggestions were made to improve compliance to the intervention or program or implementation. Also, the e-health intervention is not explained in this paper, it would be nice if you could spend a few lines in this paper about this program so the reader can understand the compliance issues.
W e have provided an extensive description of the eHealth intervention, i.e. our “multifaceted eHealth strategy” in t he Methods section, under the paragraph ‘Intervention and control’. Also, due to maximum word count for this paper, we have referred to other publications in which the strategy has been discussed in detail. We have added the following statement to the description to clarify the compliance issue: “Patients were requi red to use pre-set usernames and passwords t o enter the intervention website.”

A minor issue is that this study is impossible to repeat for other institutions because

of lacking access to the e-health program used in this study. This should be discussed in the discussion part.
This is a very good point. While we no longer have access to the entire website, we do have the materials used available (i.e. videos, written information, exercises, etc.) We have added the following statement about this to the Discussion section, paragraph Study limitations:
"Unfortunately, the eHealth strategy is no longer accessible, which makes repeating of this study difficult. As the strategy was financed through the funding for the trial, no financial resources were available to keep the eHealth strategy functioning after the trial ended and funding stopped. Materials and screenshots are still available for future use."

The abstract could contain some information mentioned in the results about the patient population i.e. ↑QOL and physical function, and in the discussion the suggestion that this program should be evaluated in a back pain population with lower health states. That is, if you think the compliance problems could be dealt with.
We have restructured our abstract completely to comply with the sub headings of a structured abstract. We were unable to include all information in the abstract due to the maximum word count for the Abstract, but we hope that the restructured Abstract provides a better overview of our paper.

Reviewer: 5
Reviewer Name: Sarah Paganini
Institution and Country: University of Freiburg, Germany
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
The present manuscript aims to investigate the effectiveness and cost-effectiveness of multifaceted eHealth strategy aiming at improving back pain beliefs and disability in comparison with usual care for back pain patients. Data were drawn from 779 patients recruited in general or ohysiotherapy practices.

This evaluation of multifaceted eHealth is of importance with regard to shortages in our health care systems and the high disability for individuals suffering from low back pain. The strength of this study is the implementation of an eHelath strategy in routine care. However there are several limitations and not all information is provided (according to the CONSORT statement and CHEERS guidelines). One major shortcoming is that a cost-effectiveness analysis and a cost-utility analysis are stated, but only a cost-utility analysis (CUA) is performed. This CUA is not discussed in the discussion section and there are contradicting statements. The listed shortcomings should be clarified. The manuscript would benefit from English proof reading.
We have included some more information according to the statements and adjusted the supplementary files with the statements accordingly.
We have changed the definition of our economic analysis in the manuscript to state more specifically that we performed a cost-utility analysis instead of a cost-effectiveness analysis.

Major Comments Titel:
1. In the paper it is stated that a cost-effectiveness and a cost-utility analysis will be performed. However, only a cost-utility analysis is provided. This should be clearly mentioned in the title ("cost-utility" instead of "cost-effectiveness analysis").
We have changed our title to state that we performed a cost-utility analysis instead of a cost-effectiveness analysis.

Abstract:

1. Please provide information of inclusion criteria

2. In the methods section there is no information about the outcomes, assessment and statistical methods
3. There is no information about the cost-effectiveness analysis
4. Results: Statistical information for the main results and the exact costs should be provided. No quality of life measures/results are given
5. Line 50: Results: Why did 37% of the participants did not have back pain at baseline

(an inclusion criterium)? This sentence needs further explanation.
We have now completely restructured our abstract so that it is in line with the side headings of a structured abstract. We have included as much as possible of the information suggested that we were able to fit into the maximum word count for the Abstracts. We hope that this restructured abstracts provides a better overview of our paper.

The guidelines of the CONSORT Statement and the CHEERS Guidelines should be followed

We have followed these guidelines and have included checklists for both guidelines as supplementary files.

Background:

1. Line 93-94: There is a distinction between “indirect costs due to absenteeism” and “productivity losses due to disability”. All of those costs are indirect costs. There should be a clarification of this distinction and a definition of indirect costs in this paper.
We agree with the reviewer. Therefore, a clearer distinction has been added to the Introduction.
2. Line 110-111: A reference to interventions that specifically aim at (back) pain would be helpful.
W e have cited the study titled “ Self-management program for chronic low back pain: A systematic review and meta-analysis ” by Du et al. (citation number 15) that provides an overview of intervention aimed at low back pain.
3. The rational for the multifaceted strategy of the eHealth program should be mentioned.

The rationale for our strategy is based on previous studies that showed promising in improving health and self-management in patients with physical disease. We have included the following rationale in our Background: “However, eHealth, which is the provision of (personalised) health care at a distance (e.g. through internet and thus digital), has shown promise with regards to its ’ eff ectiveness and cost -effectiveness in improving outcomes such as patient health, patient satisfaction, self-management and healthcare costs in patients with physical diseases.”

Methods:

1. It is not stated, how exclusion criteria were assessed
We have included an explanation of how exclusion criteria were assessed.

2. There is no information from when until when participants were recruited.

We have included dates for the trial.

3. The design of the stepped-wedge cluster randomised controlled trial and the procedure should be described in more detail. Figure 1 seems not to be enough for clarification.

Due to the maximum word count for this paper, we have chosen to refer to our protocol paper for more detail regarding the trial and procedures. Figure 1 is added for illustration.

4. There is no information about blinding/masking

We have included this information.

5. Line 162: How was this continuing medical education operationalized?

Due to the maximum word count for this paper, and as this was not the main part for the current study, we have chosen to refer to our other (published) paper for more detail on this.

6. Line 167: It is not clear, what the professional based intervention is. This should be clarified for the reader, even though there is a study protocol with further information.

Due to the maximum word count for this paper, and as this was not the main part for the current study, we have chosen to refer to our other (published) paper for more detail on this.

7. Line 170: Is the description correct? “The BBQ is designed to measure the inevitable consequences of LBP”. Or does it measure the “belief” about these consequences? Thank you for noticing this error. We have adjusted this.

8. The sample size seems quite low. Could you provide further information, how you applied the ICC in your calculation (not necessarily in the manuscript. Only for clarification)?

The sample size calculation was done by a biostatistician who provided us with the following information: “The sample size calculation w as based on a hypothesized 10% improvement in back pain beliefs as measured by the BBQ, based on an observed mean improvement of 9.6% between three successive surveys in the Australian campaign.[19] An intra-class correlation coefficient (ICC) of 0.05 was applied to adjust for the cluster randomisation design. Assuming a 10 % improvement from a mean score of 26.5 (95% Confidence Interval (CI) 26.1-26.8, SD 6) on the BBQ, and applying an ICC of 0.05, the necessary sample size was estimated to be 500 patients. This calculation takes into account a dropout-rate of 20%, power (1-beta) of 0.90 and an alpha of 0.05. ”

9. Line 184-186: Validity and reliability criteria for the RDQ-24 and the EQ-5D-3L is missing. We have provided citations throughout our manuscript for these questionnaires that should provide this information. We have not discussed these criteria in our paper due to the maximum word count.

10. Was there an assessment of negative effects of the treatment (“harms” in the COSNORT statement)?

No, this assessment was not done because of the nature of our intervention studied, i.e. a voluntary website containing objective information. No treatments were forced upon participants.

11. Line 189: The correct name of the TIC-P is “Trimbos/iMTA questionnaire for Costs associated with Psychiatric Illness”. The time period of the TIC-P should be mentioned (last three months).

We have changed the name of the TIC-P and included that it was measured over the past three months.

12. Line 203, 204: A reference is missing. What is meant by “professional organisations”?

We mean healthcare professionals ’ associations, and w e have changed this accordingly.

13. Line 222: What is the rational for imputing data separately for the intervention and control group?

Data were imputed separately for the intervention and control group to account for the possibility that the association between observed factors and the missingness of data differs across groups.

14. Line 228: Information should be provided of how the data/results of the different imputation data sets were aggregated and how many imputations have been done.
Variables associated with t he “missingness ” of data, outcomes and potential confounders were included in the imputation model. Cost and effect measure values were imputed per time point, costs were imputed at the cost category level and effects were imputed at the outcome level. A total of 10 complete data sets were generated in order for the loss of efficiency to be below 5% and pooled estimated were calculated according to Rubin ’s rules. This information has been added to the manuscript.
15. Line 230: For clarification: Were there different imputations for the effectiveness and the cost-effectiveness analysis and when yes, why?
Different strategies were used for dealing with missing data in the effectiveness and the cost-eff ectiveness analysis; mixed models / MLE in the effectiveness analyses… because data were analysed longitudinally and multiple imputation in the cost-effectiveness analysis so that data could be imputed at a lower level (cost category level instead of the total cost level).
16. The main outcome(s) for the CEA and the CUA should be clarified.

Only a CUA in terms of QALYs was performed. This has been indicated more clearly.
17. In the study protocol a budget impact analysis was planned. If this was not done it should be stated as exception to the protocol.
A budget impact analysis will not be performed because of the lack of effectiveness on any of the outcome measures of the study.
18. In the CHEERS checklist it is stated that “the choice of model” was described in page 10-12. As this was a CUA alongside a clinical trial it can be assumed that no decision- analytical model was specified.
That is correct. We have adjusted the checklist accordingly.

19. No information about the discount rate is given (even if data were not discounted this should be stated; see CHEERS guidelines).
Information on discounting has been added to the manuscript.

20. Currency and price date should be stated (see CHHEERS guidelines).

Information on the currency and price has been added to the manuscript.

21. It should be mentioned whether the trial followed the CHEERS guidelines.
We have included a statement on this.

Results:

1. For clarification: The only inclusion criterium was diagnosed low back pain?
Yes, patients diagnosed with non-specific low back pain were included in this study. Nonspecific LBP was defined as LBP with or without motor and/or sensory deficits in one or both legs, including sciatica and radiculopathy, that is not caused by underlying specific pathology (red flags), i.e. a tumour, (osteoporotic) vertebral fracture, ankylosing spondylitis, and cauda equina syndrome.
How did you test for the exclusion criteria (e.g. psychiatric disorder)? How many patients were excluded due to specific exclusion criteria? Reasons for exclusion should be listed precisely in figure 2 (according to the CONSORT guidelines).
Exclusion criteria were built into the software program that selected patients from the patient files of the general practitioner. The general practitioner and physiotherapists also manually checked for exclusion. As we used software to select patients, no numbers and reasons on specific exclusion criteria are available.

2. Table 1: It should be clarified, why there is no complete baseline information for all study participants (n varies for each baseline characteristic)
As the questionnaire was voluntary, patients were free to not provide information on certain aspects if they wished to not do so, which is the reason why the n varies for different measures.
Please state why only 63% suffered from back pain at baseline, as this was the only inclusion criterium.
We asked patients if they had back pain at baseline, i.e. the moment they filled in the questionnaire. This moment could have been (much) later than their back pain began, and some patients already recovered from their back pain before the start of the study. This was because of the time lag between patient selection and patients actually filling in the questionnaire. We have discussed this in our Discussion section, in the second paragraph.
. Please clarify, how the categories of educational level were defined
We have added the categorizations for educational level into the table. Lower education was characterized as only having finished primary school. Vocational level is having a college degree, and higher education level is having a degree from university or university of applied sciences.
. Please clarify, how "activity" was defined. How was physical activity and physical demanding work assessed?
Both were self reported. Activity included any physical activity during a week in minutes per week. Physically demanding work was asked as binary question (yes/no).

3. Table 2 and 3:

. For better readability mean values and effects should be presented in one table
We have chosen to present in separate tables, as combining in one table would be difficult for the effect size from table 3 that measure the effect over all time points, while table 2 shows the means for each specific timepoint.
. The abbreviations should be defined (M, F, CI)

We have fully written out Male and Female these words instead of using abbreviations. The abbreviation for CI has been written out fully in the Methods section.
. What is the rational for making different adjustments for back pain beliefs and disability? Both were separate statistical models in which other variables showed confounding effects. We have adjusted for relevant confounding in each model.
. It is not clear why there is a separate analysis for men and women for disability and not for back pain beliefs
Both were separate statistical models in which other variables showed interaction. We have adjusted for relevant interaction terms in each model, gender was only relevant in the model for disability.
It is not clear whether the results refer to imputed data or not.
Only data for economic evaluation were imputed, as stated in our methods.

4. It should be stated if missing data differed significantly between groups
A complete missing-case analysis was not performed as this was not part of our study, but the groups did not differ significantly on other data.

5. Table 4:

. Abbreviations and "." should be clarified
We have fully written out the previously abbreviated words.

. Why are SEMs reported and not SD?

SEMs were reported instead of SDs because the mean costs per group were pooled estimates over the 10 data sets.
. Is “unpaid productivity” the same as “Informal care”? The same wording should be used or both categories should be mentioned and described earlier.
No it is not the same. The Methods section has been adjusted to indicate this more clearly.
. It would be even more informative, if the cost categories would be presented in more detail (what kind of primary, secondary, alternative care was used? What kind of medication? Only back pain medication or other as well?
Unfortunately, there is too much data to provide this information in a comprehensible manner and too little room (i.e. word limits) to provide such detailed information. As stated in our data sharing agreement, data can be requested from the corresponding author upon reasonable request.

. How are the total societal costs calculated? Adding all costs results in higher costs.

Total costs were the sum of all cost categories. The Methods section has been adapted to indicate this more clearly.

6. Mean values of the QALYs in each group should be stated. The mean QALYs at baseline should also be stated. If they differ, there should be adjustment for baseline QALYs as well.
Mean QALYs per group are now reported. We did not have information on the difference in QALYs at baseline. However, as the mean difference in utility values at baseline was relatively large, albeit not statistically significant, we decided to rerun the analyses, and have now adjusted for baseline.
7. Line 293: Negative ICERs should not be interpreted. If there are negative costs the ICER gets more and more negative with smaller QALY health gains, suggesting a high amount of money saved. A better description is that “the intervention dominated standard care”. However, it has to be pointed out that the QALY gain was very low and not statistically significant.
We agree with the reviewer. Therefore, the interpretation has been removed and it is now only stated that the intervention dominates usual care.
8. CI of the ICER should be mentioned

We respectfully disagree with the reviewer with regards to the provision of 95%CIs surrounding ICERs. ICERs in itself are hard to interpret. To illustrate, negative ICERs might represent reduced costs and positive effects indicating a win–win situation or increased costs and negative effects indicating a lose–lose situation. With participant-level data, we agree with the reviewer that it is natural to consider representing the uncertainty surrounding ICERs using 95% CIs. Nevertheless, as a ratio measure, estimating 95% CIs around ICERs is not straightforward and, more importantly, 95% CIs around ICERs suffer from the same interpretation problem as ICERs. As such, the provision of 95%CIs surrounding ICERs is currently discouraged (Briggs et al. 2002, Gray et al. 2011, van Dongen et al. 2014). As an alternative, we constructed cost-effectiveness acceptability curves, which provide an estimation of the joint uncertainty surrounding costs and effects (and thus incorporate an estimation of the statistical significance of the effect differences).
9. Line 295: Is 79% the correct value? On the Cost-effectiveness plane (figure 3) it seems like there are less cost effect pairs in the South East quadrant.
We checked this number and it is correct.

10. Line 297: It should be clarified from which data this statement comes from (“The uncertainty around the cost-effectiveness estimate was large”)

We agree with the reviewer that the statement was a bit confusing. Therefore, we decided to remove it from the manuscript.

11. There are only results for a CUA. In the method section a CEA was mentioned as well (no outcome was defined). This should urgently be clarified.

The term CEA was incorrect. This has now been removed from the manuscript.

12. The CUA calculations should be checked thoroughly. There are cost savings but there is no significant difference for QALYs and very small QALY health gains for the intervention group. Therefore, it is surprising, that the CEAC reveals such good probabilities of being cost-effective.

The differences in QALYs were significant when adjusted for baseline and when not adjusted for baseline. Stating that there were no significant differences in QALYs was a mistake in the text of the previous version of our manuscript, (although the numbers shown were correct).

This lead to the CEAC looking questionable, but this is explainable by this textual error. From a health economic perspective, the intervention was cost-effective, but due to the lack of effectiveness on our primary or secondary outcomes, we conclude that this intervention should not further be implemented.

Discussion:

1. Line 319: Is a BBQ score of 26.5 (on a scale from 9 to 45) meaningfully higher than in the current study (24.7). It is questionable if this could be a possible explanation for the results.

We agree that this is quite a difference, but we do not know if this difference is meaningful. We state these numbers in our discussion, to compare our results to the results of the study that we based our intervention on. However, this is not an explanation for our results.

2. Line 323-324: The result of the study (attitudes of elderly) is not discussed fully. How can this study results explain the current results?

W e compare our study to t his study, as our population ’s mean age was relatively on the higher side, and as many (almost half) of our participants already exited the workforce (retired). It is interesting therefore to see that the elderly population in that study had comparably low back beliefs as our population.

3. There is literature on how adherence can be improved in eHealth strategies. There is also literature on self-help interventions that are more structured and guided (and effective). This could be included in the discussion of the question: What do LBP patients need to improve their beliefs and disability?

We agree this is interesting literature. We have performed a process-evaluation among the participants of our study to look into their needs. We have published these results elsewhere and have cited that study in our current paper for reference and further discussion.

4. Line 370: It is stated that no cost-effectiveness is yielded (again in line 396). In line 306 it was stated that “The probability of cost-effectiveness was high”. These are contradicting statements. The results of the cost-utility analysis are not discussed. This is urgently needed. Why is there such a good probability of being cost-effective for an intervention that yields no significant QALY health gains? What could be the reason for the high cost savings in the intervention group for absenteeism and presenteeism, when adherence to the intervention was not high and no difference in other outcomes could be found?

We have removed these statements and have added more information on the cost-utility analysis. The differences in QALYs were significant when adjusted for baseline and when not adjusted for baseline. Stating that there were no significant differences in QALYs was a mistake in the text of the previous version of our manuscript, (although the numbers shown were correct). This lead to the CEAC looking questionable, but this is explainable by this textual error. From a health economic perspective, the intervention was cost-effective, but

due to the lack of effectiveness on our primary or secondary outcomes, we conclude that this intervention should not further be implemented.
With regards to the reason for the high cost savings in the intervention group for absenteeism and presenteeism in the light of low adherence, one possible explanation could be the high educational level of participants, or the fact that the intervention had a large component related to work and encouraging (return to) work.
5. Limitations for the cost-utility analysis should be stated (e.g. concerning power)

W e have added the following statement to our discussion: “Lastly, as for the lack of significant cost differences in light of the cost-utility analysis, it is known that cost data are highly skewed and therefore require large sample sizes to detect statistically significant differences.[53] In this study, the sample size calculation was based on back beliefs, which
may have underpowered it to detect significant cost differences.”

6. Implications for future research are that participants with LBP should be included in further studies. As this was the aim of the current trail it should be critically discussed why baseline assessment was not only done for individuals that suffer from back pain.
We have included in our discussion that participants with poorer health states should be included in future research. At baseline, some of our patients may have already recovered from their back pain, as there was a time lag between patient selection for the study and patients starting the study and filling out our baseline questionnaires. We have included this in our discussion as well.

Minor Comments Abstract:
- Line 41, 42: Please check the syntax of the sentence: “Four clusters of general and physiotherapy practices and occupational physicians were randomised…”
- Line 106: Do “economic” and “societal burden” mean different aspects?
We have now completely restructured our abstract so that it is in line with the side headings of a structured abstract. We have included as much as possible of the information suggested that we were able to fit into the maximum word count for the Abstracts. We hope that this restructured abstracts provides a better overview of our paper.

Background:
-Line 92: To clarify the time horizon for these costs “per year” could be added.

W e have stated that these costs are “annually ” (i.e. per year) .

Methods:

- The reference for the BBQ is missing.

Two references to the BBQ are included at the end of the paragraph that describes the BBQ.

Results:

1. The wording should be consistent (follow-up vs. follow up)
We have made changes to the manuscript to be more consistent in wording.

2. Table 1:

. “(SD)” should be added after “Back pain beliefs”, “Disability”, “absenteeism” and “Quality of life”
We have added this to the table.

. It should be mentioned, what the first information stands for (sometimes “mean”, sometimes “number of participants”?)
We have included this information in the table for the outcomes where it was necessary.

3. Table 3: “Back beliefs” . “Back pain beliefs”

We have changed this.

4. Line 290 and 292: abbreviations should be clarified
We have written out the abbreviations.

5. Figure 3:

. In the caption there should be more precise information (imputed data, bootstrapping, QALYs based on EQ-5D-3L).
For the purpose of readability of the figures, we have chosen to keep the captions concise. This information is provided extensively in the Methods section of the paper.
. There is no labelling for the x-axis. The labelling for the y-axis is not correct
We have changed this.

. It would be helpful to see more of the whole plane (all 4 quadrants)

We have changed this.

6. Figure 4:

. In the caption there should be more precise information (what means “CEAC”, imputed data, bootstrapping, QALYs based on EQ-5D-3L).
For the purpose of readability of the figures, we have chosen to keep the captions concise. This information is provided extensively in the Methods section of the paper.
. The labelling for the x-axis is not correct (€)
We have changed this.

Discussion

1. Line 317: Wording (“back beliefs”)

We have changed this.

2. Line 332: This information should be stated in the results chapter
This information is not included in the results section, as it is not a result of the current study. It is a result from the previously published process-evaluation, which is cited at the end of this given information.
3. Line 334-335: Grammar of the sentence should be checked
We have checked this sentence and changed it.

Reviewer: 6
Reviewer Name: Jesse Kigozi
Institution and Country: University of Birmingham, United Kingdom.
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
This is a well-written paper, but with comments below that would need addressing or commenting on.
1. The authors did not perform any sensitivity analysis. It would be good to see at least a sensitivity analysis using the complete-case to assess the impact of the missing data.
A sensitivity analysis based on complete cases has now been performed and added to the manuscript.

2. How did the authors determine all the variables that were included in the imputation model? How was the imputation done? At what level was the imputation done? More information is needed to know what was actually done. Also, did the imputation model take into consideration the effect of clustering? Similarly, was clustering taken into consideration for the models used to generate QALY and cost differences?
Variables associated with t he "missingness " of data, outcomes and potential confounders were included in the imputation model. Cost and effect measure values were imputed per time point, costs were imputed at the cost category level and effects were imputed at the outcome level. A total of 10 complete data sets were generated in order for the loss of
eff iciency to be below 5% and pooled estimated were calculated according t o Rubin 's rules. This information has been added to the manuscript.
Clustering was not taken into account in the economic evaluation, as the level of clustering was rather low. We therefore preferred to correct for the correlation between costs and effects using a bivariate analysis (SUR), which cannot be combined with a mixed model in a frequentistic framework.
Clustering has been taken into account in the effect evaluation by using mixed-models multilevel analysis.

3. How were the intervention costs actually calculated? This detail was not included and it's not clear how these costs were calculated and apportioned to the patients?
Intervention costs were micro-costed. More detailed information about this procedure has been added to the manuscript.

4. Unit costs – could the authors include a table of the unit costs that have been used to estimate the costs and the corresponding resource in detail. This can be included in the appendix if not available but would need to be seen.
We have included a Supplementary File containing the unit cost table.

5. The QALY scores are currently reported rounded to 2 decimal places – need to extend these to at least 3.
We have extended these to 3 decimals.

6. Need to include how long resource use data were collected for at each time point in the text (direct and indirect costs)
Resource use data was collected using 3-month recall periods. This information has been added to the manuscript.

7. Need to include a statement about the need or not for discounting in the text.

A statement about discounting has been added to the manuscript.

8. What threshold have the authors used to determine a lack of cost-effectiveness – for most of the WTP values at least over 20,000 there is an over 80% chance of being cost- effective?
We agree that stating there is a lack of cost-effectiveness is not entirely true. From a health economic perspective and looking at the cost-utility analysis, the intervention was cost- effective, but due to the lack of effectiveness on our primary or secondary outcomes, we conclude that this intervention should not further be implemented.
9. In the discussion, the authors report that the probability of the intervention being cost-effective was high but then go on to conclude that it was not cost-effective? What informed this conclusion?
This conclusion was based on the lack of effectiveness on our primary or secondary outcomes, although the results of the cost-utility analysis based on QALYs was promising. We have reworded our conclusion to show this nuance.

VERSION 2 – REVIEW

REVIEWER	Allan Riis Center for General Practice at Aalborg University
REVIEW RETURNED	25-Oct-2019

GENERAL COMMENTS	Thanks again for the opportunity to review this interesting paper. My comments are all sufficiently addressed.

REVIEWER	J.W. Geurts Rijnstate Hospital Arnhem, the Netherlands
REVIEW RETURNED	04-Nov-2019

GENERAL COMMENTS	Dear Authors, My answers and remarks are sufficiently covered. Kind regards