PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cost-effectiveness of an Internet-based perioperative care programme to enhance postoperative recovery in gynaecological patients: economic evaluation alongside a stepped-wedge cluster-randomised trial
AUTHORS	Bouwsma, Esther; Bosmans, J; van Dongen, Johanna M.; Brölmann, Hans; Anema, Johannes; Huirne, Judith

VERSION 1 – REVIEW

REVIEWER	Aron Onerup Institute of clinical sciences, Sahlgrenska Academy, University of Gothenburg, Sweden
REVIEW RETURNED	27-Jun-2017

GENERAL COMMENTS	The article is well-written and easy to read. However, my main concern is that I find it hard to justify a health economic analysis. Could you justify why it should be done? Health economy is based on the clinical outcome measures, mainly RTW. Given that the primary outcome measure is not significant a health economic analysis seems redundant. In this manuscript the primary and secondary effect measures are discussed once again although they are submitted for publication in a separate manuscript. Is this a manuscript on exclusively the health eonomy? In that case I suggest that the reports on other effect measures such as RTW should be excluded and only reported in the manuscript for that purpose. I find the discussion too elaborate and positive regarding the results. Results are presented as differences although they are almost exclusively not statistically significant. I think that these should be changed from e.g. "...total societal costs in the intervention group were €647 lower compared to the usual care group..." to "there was no statistically significant difference..." with the numbers within parenthesis. To hypothesise that there would be a difference if it would have been possible to use median instead of mean seems optimistic and would have to be justified more than it is in the manuscript. Here are my specific suggestions for revisions: - Use the term intervention instead of "care programme" through all of the manuscript when used in the same sentence as usual care, e.g. in the abstract under "Intervention". - p 5, row 31: Is it possible to conceal allocation until inclusion with unblinded researchers and clinicians? - p5 row 58, wrong order of references (11 before 10) - Participants: Why did the study not reach full inclusion (433 out of

	planned 454)? - p 9 row 53: It seems unlogical to assume missing at random when the intervention is computer-based and dropouts not using computer as much as participants with complete data. see also p 14, row 35. - I suggest that you change from mentioning the exact number to "did not differ statistically sginificant" or similar term on p 10, row 13, row 18, and row 31. - p 10, row 41: I find this result hard to understand: If there is no statistically significant difference in RTW between the two groups, how can it make any difference how much the society is willing to pay for each day? It is not intuitively easy to understand to me. - p11, row 38: Given that health economy is underreported as stated in the article, I find it justified to clarify what level of probability for cost-effectiveness is commonly used, e.g. 95% CI, p<0.05 etc. This should also be stated in the first paragraph of the discussion. - p12, row 45: The hypothesis that it may be clinically relevant although not statistically significant seems bold and needs to be clarified or taken away. -p12, row 52: If the difference is not statistically significant, I find it hard to state that the costs were lower. -p13, row 8: To discuss costs for secondary care separately seems problematic since it's a part of the health economic analysis and not stated as an outcome. Risk of type I error due to multiple testing.

REVIEWER	Gregg Nelson Tom Baker Cancer Centre Canada
REVIEW RETURNED	04-Aug-2017

GENERAL COMMENTS	Well designed study and rigorous cost effectiveness analysis. Significant contribution to the literature. Well done.

REVIEWER	David Story University of Melbourne, Australia No Competing Interest
REVIEW RETURNED	21-Aug-2017

GENERAL COMMENTS	The Authors have conducted a cost-effectiveness study as part of a stepped wedge implementation study of an e-health intervention to enhance return to work after gynaecological surgery. The Authors have made my job much easier by attaching both the companion clinical outcome paper and the methods paper: thank you! I note that I am a clinician researcher in perioperative medicine rather than a health economist. I think this is an important e-Health study and it is pleasing to see both clinical (in the companion manuscript) and economic analyses. 1. The analysis boils down to the Cost Effectiveness Acceptability curves. I think the Authors describe these results well but I would consider adding the alternate wording (perhaps in the Discussion) to say that the results also mean there is a 79% chance that there is no added cost / day saved in RTW and a 97% chance the added cost is less than 76 Euros per day saved in RTW. 2. I think the conclusion of the abstract should be more in line with the text, that the program is cost-effective if 75 Euros is a

	reasonable price to pay per day of earlier return to work. 3. For Figure 3 I would keep the cost-effectiveness planes and not use the CEACs. 4. The Authors seem to have been caught up by statistical complexity. RTW using means was 49.6 vs 56.2 days; mean difference -4.1 days (CI -10.8 to 2.6) but was not normally distributed and in the companion paper was median 49 vs median 62 days. The hazard analysis was hampered by differencing patterns before and after 85 days. I think it is reasonable to include the means and medians for RTW in this paper. 5. How does that differing outcome (85 days) affect these cost-effectiveness results? I know I'm not reviewing that manuscript but my view is the results are more complex than the Authors seem to suggest. The intervention seemed to work for about 80% of patients and not work for the other 20%, i.e. 1 in 5 patients. I note the Authors wonder is this is due to statistical anomalies but it also may be true. 6. Is there any evidence that the 85+ group differed from the <85 group? Should patients be screened? 7. Overall, the group seems to be reasonably young, well educated, computer using women who speak reasonable Dutch, and salaryed. I wonder if clinical and cost effectiveness may fall away with a group with less literacy (education or Dutch as a second language) and less computer literacy. I think the Authors should discuss that point in a discussion of generalisability. 8. Do the Authors think there are any Dutch cultural factors (such as attitudes to doctors and hospitals) that affected the results and may limit generalisability? I suspect comparison with other European countries might be easiest. 9. I have mentioned statistical review but really mean formal helath economics rview.

REVIEWER	Basile Pache Lausanne University Hospital CHUV, Switzerland
REVIEW RETURNED	21-Aug-2017

GENERAL COMMENTS	Bouwsma et al performed an economical analysis (cost-effectiveness and cost-utility) of their internet-based care programme running from 2011 and 2014 in nine centers in Holland. They found out that the intervention is cost-effective for duration until return to work but had no impact on societal costs. The study is nicely designed and open the way to other studies for promoting eHealth post-operative programs Minor remarks Page 4 Line 38 : Please add reference of the article of clinical effectiveness Page 8 line 10 : please add reference of the formula of presenteeism as it is a major calculation basis. Page 8 line 21 : As the study runned from 2011 to 2014, why isn't the productivity indexed to the year ? (as economical crisis stroked around 2011) Page 9 Line 41 : Is there a reason with numbers of patient defers with page 8 line 28 ? (433 patients vs 454) Page 23 ; has the follow been assessed in demographic characteristics? Ethical origins, stratification of income before

	surgery and type of work sector (primary , secondary or tertiary sector of economy) ? Please have the manuscript read by English native for little improvement. Question to the authors : Were the mains reasons for delayed RTW assessed? Was it as well assessed from the patient point of view ?

VERSION 1 – AUTHOR RESPONSE

Response to Reviewer 1

Comment1: The article is well-written and easy to read. However, my main concern is that I find it hard to justify a health economic analysis. Could you justify why it should be done? Health economy is based on the clinical outcome measures, mainly RTW. Given that the primary outcome measure is not significant a health economic analysis seems redundant. In this manuscript the primary and secondary effect measures are discussed once again although they are submitted for publication in a separate manuscript. Is this a manuscript on exclusively the health economy? In that case I suggest that the reports on other effect measures such as RTW should be excluded and only reported in the manuscript for that purpose.

Response: Economic evaluations are increasingly being used for decision making and are an important aspect of health technology assessments internationally.(1) Therefore, the decision to perform an economic evaluation to assess the cost-effectiveness of our developed web-based care programme was made prior to the start of the trial and has been included in our research protocol. To prevent publication bias, it is important that the results of our study are reported in accordance with our research protocol, irrespective of the (clinical) effect. Performing (and publishing) an economic evaluation merely in cases of proven effectiveness would lead to an overestimation of the economic benefits of health interventions.

In addition, a programme can be considered as cost-effective – even if it is not performing better than the alternative strategy – for example, when it turns out that the intervention is less costly and policy makers are willing to accept a lower effectiveness.(2)

In an economic evaluation, costs are related to health benefits. Therefore, the suggestion to exclude reporting on the effect measures such as RTW in the current manuscript seems infeasible.

Comment2: I find the discussion too elaborate and positive regarding the results. Results are presented as differences although they are almost exclusively not statistically significant. I think that these should be changed from e.g. "...total societal costs in the intervention group were €647 lower compared to the usual care group..." to "there was no statistically significant difference..." with the numbers within parenthesis.
To hypothesise that there would be a difference if it would have been possible to use median instead of mean seems optimistic and would have to be justified more than it is in the manuscript.

Response: We regret that the reviewer feels we interpreted the data too positive. We rephrased several sentences in the discussion section, in order to realize a more neutral and transparent presentation of the data (also see suggestion 6 below). However, we disagree with the reviewer that our data show no additional advantage of the intervention over usual care. Our results show that for the primary outcome duration until full resumption of work, the probability that the care programme is

cost-effective as compared to usual care is 0.97 at a willingness to pay of €76 per day earlier RTW. Taking into account that the average costs per sick leave day are €230, we conclude that the intervention is cost-effective.

The reviewer also stated that the discussion section is too elaborate. We have the opinion that all items discussed contribute to the understanding and interpretation of the data for the readers. We leave it up to the opinion of the editor if we should shorten the discussion, however, in that case we would like to receive some more specific instructions on the specific parts that could be deleted or shortened.

Suggestion1: Use the term intervention instead of "care programme" through all of the manuscript when used in the same sentence as usual care, e.g. in the abstract under "Intervention".

Response: We have changed this when possible.

Suggestion2: p 5, row 31: Is it possible to conceal allocation until inclusion with unblinded researchers and clinicians?

Response: Recruitment was performed by an independent research nurse using waiting lists from each hospital and not by their own health care providers. During counselling the research nurse explained to the patient that the chances of receiving the intervention or usual care were equal. Only after signing informed consent, patients were informed about the allocation of the hospital they had their surgery in.

We agree that there is a risk of recruitment bias due to the cluster design, however, we believe this was minimized as the proportion of patients included during the control phase, was broadly similar to the proportion of inclusions during the intervention phase, across all participating hospitals.

We did not add this argument to the manuscript as it is discussed in the effectiveness paper.

Suggestion3: p5 row 58, wrong order of references (11 before 10)

Response: The reference numbers are listed in sequence of appearance. Reference number 9 is already used on page 4 first paragraph, and references 10 and 11 on page 4 in the second paragraph. Because the references are used more than once, the order on page 5 seems wrong, however, in fact is correct.

Suggestion4: Participants: Why did the study not reach full inclusion (433 out of planned 454)?

Response: A stepped wedge design is a unidirectional crossover design in which the clusters cross over from the control group to the intervention group. This means that at the end of the study, all clusters provide care according to the intervention. In our study the number of patients in the usual care group was smaller than the number of patients in the intervention group (206 versus 227). Although lengthening the inclusion phase would lead to reaching the number of patients calculated in the power-analysis, we decided against this, because this would have led to a greater misbalance between the number of patients in the control and intervention group.

We will add this note to the effectiveness manuscript, as this paper handles the stepped wedge design in more detail.

Suggestion5: p 9 row 53: It seems unlogical to assume missing at random when the intervention is computer-based and dropouts not using computer as much as participants with complete data. See also p 14, row 35.

Response: It is important to note that participants with uncomplete cost data were not necessarily patients that dropped out of the study. Data on service use was collected monthly and by missing one data entry point the health cost data of that particular participant was considered as incomplete (this was the case in 130/433 patients). In total, 32 patients dropped out of the study for several reasons (see effectiveness paper).

Reasons for missing data are commonly classified as: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).(3) In case of missing at random, any systematic difference between the missing values and the observed values can be explained by differences in observed data.(3) As Internet-use is an observed variable, it is plausible to consider data missing at random, making multiple imputation the appropriate method to deal with the missing data.

Suggestion6: I suggest that you change from mentioning the exact number to "did not differ statistically significant" or similar term on p 10, row 13, row 18, and row 31.

Response: We decided to use our original phrasing to indicate the differences in the total societal costs and the difference in duration until RTW between both groups, as these mean differences were used to calculate the ICER (incremental cost-effectiveness ratio). We did rephrase the sentences of the remaining outcomes.

Suggestion7: p 10, row 41: I find this result hard to understand: If there is no statistically significant difference in RTW between the two groups, how can it make any difference how much the society is willing to pay for each day? It is not intuitively easy to understand to me.

Response: We acknowledge that cost-effectiveness data are sometimes hard to interpret for readers who are not frequently exposed to health economic studies. It is important to realize that economic evaluations involve the joint distribution of costs and consequences and can demonstrate cost-effectiveness when neither cost nor effect differences are individually significant.(4)

An ICER indicates the additional investments needed for the intervention to gain one extra unit of effect compared with usual care. It is calculated by dividing the mean difference in cost between study arms by the mean difference in effect. When plotted on a cost-effectiveness plane (CE-plane), it is possible to illustrate if the new intervention is more or less effective than the alternative, and simultaneously if it is more or less costly than the alternative:

Northeast quadrant: the intervention is more effective and more costly than the alternative.
Southeast quadrant: the intervention is more effective and less costly than the alternative.
Southwest quadrant: the intervention is less effective and less costly than the alternative.
Northwest quadrant: the intervention is less effective and more costly than the alternative.

When an intervention is more effective and more costly (NE-Q) or when it is less effective and less costly (SW-Q), the decision whether or not to adopt it, depends on the so-called willingness-to-pay (the maximum amount of money decision-makers are willing to pay for an additional unit of effect). The CE-plane can be divided into a cost-effective halve and a non-cost-effective halve by plotting a hypothesized willingness-to-pay. The steepness of the slope of this line, depends on the amount of money society is willing to pay more for an additional unit of effect. And therefore, if society is willing

to pay a higher amount of money, the chances increase that the ICER falls into the cost-effective halve of the CE-plane.(2, 4)

For further reading please see the papers by van Dongen et.al.(2) and Petrou et.al.(4)

Suggestion8: p11, row 38: Given that health economy is underreported as stated in the article, I find it justified to clarify what level of probability for cost-effectiveness is commonly used, e.g. 95% CI, $p<0.05$ etc. This should also be stated in the first paragraph of the discussion.

Response: Statistical uncertainty is combined with decision uncertainty in the cost-effectiveness acceptability curves, which show the probability that the intervention is cost-effective as compared to usual care for different values of the ceiling ratio. Unfortunately, there are no generally accepted cut-off points for either the probability at which an intervention is considered cost-effective or the ceiling ratio. Therefore, the final decision is up to decision-makers.(5) However, in this study the probability of cost-effectiveness is over 0.95 at a ceiling ratio of €76 per earlier day of RTW. Since the willingness-to-pay is lower than the average costs per sick leave day (€230) the intervention can be considered cost-effective.

Suggestion9: p12, row 45: The hypothesis that it may be clinically relevant although not statistically significant seems bold and needs to be clarified or taken away.

Response: We have changed this paragraph and removed the sentence that bothered the reviewer.

Suggestion10: p12, row 52: If the difference is not statistically significant, I find it hard to state that the costs were lower.

Response: In our study the societal costs were indeed lower in the intervention group compared to the control group: €12266 versus € 12795 euros. In order to investigate if this difference is based on coincidence or can be accounted for by an effect of the intervention, statistical analysis were performed. We argue that it is possible that the lack of statistical significance is not caused by a lack in effect by the intervention, but could also be caused by a lack of power of the statistical test, because of a relative small sample size.

Suggestion11: p13, row 8: To discuss costs for secondary care separately seems problematic since it's a part of the health economic analysis and not stated as an outcome. Risk of type I error due to multiple testing.

Response: We have rephrased the sentence, not emphasizing on the statistical difference anymore. However, we would like to keep this paragraph as this finding led to a hypothesis that we would like to investigate in the near future.

Response to Reviewer 2

Comment: Well designed study and rigorous cost effectiveness analysis. Significant contribution to the literature. Well done.

Response: We thank reviewer 2 for his compliments on the design of the study. We are pleased he has no additional comments.

Response to Reviewer 3

Comment: The Authors have conducted a cost-effectiveness study as part of a stepped wedge implementation study of an e-health intervention to enhance return to work after gynaecological surgery. The Authors have made my job much easier by attaching both the companion clinical outcome paper and the methods paper: thank you! I note that I am a clinician researcher in perioperative medicine rather than a health economist.

Response: We thank reviewer 3 for his kind words and are happy he can appreciate both the clinical effectiveness as and the cost-effectiveness manuscript.

Comment 1. The analysis boils down to the Cost Effectiveness Acceptability curves. I think the Authors describe these results well but I would consider adding the alternate wording (perhaps in the Discussion) to say that the results also mean there is a 79% chance that there is no added cost / day saved in RTW and a 97% chance the added cost is less than 76 Euros per day saved in RTW.

Response: We thank the reviewer for this suggestion, however, we decided not to use this alternative wording, as reviewer 1 stated that our discussion was presented as too positive.

Comment 2. I think the conclusion of the abstract should be more in line with the text, that the program is cost-effective if 75 Euros is a reasonable price to pay per day of earlier return to work.

Response: We have changed the abstract accordingly.

Comment 3. For Figure 3 I would keep the cost-effectiveness planes and not use the CEACs.

Response: The cost-effectiveness acceptability curves form an important part of our analyses, as those demonstrate the maximum probability of cost- effectiveness and the accompanying willingness to pay. Therefore, in our opinion, we think the CEA curves belong in the main manuscript.

Comment 4. The Authors seem to have been caught up by statistical complexity. RTW using means was 49.6 vs 56.2 days; mean difference -4.1 days (CI -10.8 to 2.6) but was not normally distributed and in the companion paper was median 49 vs median 62 days. The hazard analysis was hampered by differencing patterns before and after 85 days. I think it is reasonable to include the means and medians for RTW in this paper.

Response: Due to the characteristics of the statistical analyses performed, different parameters were used in the effectiveness paper (median number of days until RTW) and the current cost-effectiveness paper (mean days until RTW) to measure the difference is RTW between study arms. We have now clarified this better under the section ‘Interpretation of the findings’. We decided not to quantify the medians from the effectiveness paper in the current manuscript, as we are afraid this could confuse the reader even more.

Comment 5. How does that differing outcome (85 days) affect these cost-effectiveness results? I know I’m not reviewing that manuscript but my view is the results are more complex than the Authors seem to suggest. The intervention seemed to work for about 80% of patients and not work for the

other 20%, i.e. 1 in 5 patients. I note the Authors wonder is this is due to statistical anomalies but it also may be true.

Response: In the economic analyses, we did not use survival analyses and hazard ratios. Therefore, we did not investigate the effect of the division between fast and slow recoverees on the cost-effectiveness results. Although this is an interesting research question, we think any post-hoc analyses on this subject fall outside the scope of the current manuscript.

Comment 6. Is there any evidence that the 85+ group differed from the <85 group? Should patients be screened?

Response: Unfortunately, with the currently available data this research question cannot be answered. Our future work will focus on identifying factors that can predict compliance to our intervention program and potential factors that can predict delayed convalescence. We agree with the reviewer that it is of great importance to determine the population that might benefit the most from our care programme, and develop alternative interventions for the remaining patients.

Comment 7. Overall, the group seems to be reasonably young, well educated, computer using women who speak reasonable Dutch, and salaried. I wonder if clinical and cost effectiveness may fall away with a group with less literacy (education or Dutch as a second language) and less computer literacy. I think the Authors should discuss that point in a discussion of generalisability.

Response: We have added an additional paragraph about the generalisability to the section 'Strengths and weaknesses of the study' discussing this point.

Comment 8. Do the Authors think there are any Dutch cultural factors (such as attitudes to doctors and hospitals) that affected the results and may limit generalisability? I suspect comparison with other European countries might be easiest.

Response: Cultural aspects could hamper generalisability, and is now discussed under the section 'Strengths and weaknesses of the study'. However, the differences in the organisation of healthcare and social security systems across countries might even cause a bigger barrier towards the transferability of the results. We discussed this already in the section 'Policy implications and recommendations'.

Comment 9. I have mentioned statistical review but really mean formal health economics review.

Response: We leave this decision up to the discretion of the editor.

Response to Reviewer 4

Bouwsma et al performed an economical analysis (cost-effectiveness and cost-utility) of their internet-based care programme running from 2011 and 2014 in nine centers in Holland.

They found out that the intervention is cost-effective for duration until return to work but had no impact on societal costs.
The study is nicely designed and open the way to other studies for promoting eHealth post-operative programs.

Response: We thank reviewer 4 for showing his appreciation of our work in the postoperative field.

Comment 1. Page 4 Line 38 : Please add reference of the article of clinical effectiveness

Response: This manuscript on clinical effectiveness is submitted to BMJ Open as well. In case of acceptance of both manuscripts, a reference will be inserted here.

Comment 2. Page 8 line 10 : please add reference of the formula of presenteeism as it is a major calculation basis.

Response: Reference 20 and 21 have been added.

Comment3. Page 8 line 21 : As the study ran from 2011 to 2014, why isn't the productivity indexed to the year? (as economical crisis stroked around 2011)

Response: All costs including absenteeism costs and presenteeism costs were indexed to the year 2014. We are not familiar with the method of using different index years.

Comment 4. Page 9 Line 41: Is there a reason with numbers of patient defers with page 8 line 28 ? (433 patients vs 454)

Response: A stepped wedge design is a unidirectional crossover design in which the clusters cross over from the control group to the intervention group. This means that at the end of the study, all clusters provide care according to the intervention. In our study the number of patients in the usual care group was smaller than the number of patients in the intervention group (206 versus 227). Although lengthening the inclusion phase would lead to reaching the number of patients calculated in the power-analysis, we decided against this, because this would have led to a greater misbalance between the number of patients in the control and intervention group.

We will add this note to the effectiveness manuscript, as this paper handles the stepped wedge design in more detail.

Comment 5. Page 23 ; has the follow been assessed in demographic characteristics? Ethical origins, stratification of income before surgery and type of work sector (primary , secondary or tertiary sector of economy)?

Response: Unfortunately, we did not collect data about ethical origin (other than Dutch versus non-Dutch) or income. The type of work sector was assessed and did not differ between both study groups.

Comment 6. Please have the manuscript read by English native for little improvement.

Response: We leave this decision up to the discretion of the editor.

Comment 7. Were the mains reasons for delayed RTW assessed? Was it as well assessed from the patient point of view ?

Response: The specific reason for delayed RTW was not assessed. Our future work will focus on identifying factors that can predict compliance to our intervention program and factors that can predict delayed convalescence.

References used in this response letter

1. Drummond MF, Schwartz JS, Jonsson B, Luce BR, Neumann PJ, Siebert U, et al. Key principles for the improved conduct of health technology assessments for resource allocation decisions. Int J Technol Assess Health Care. 2008;24(3):244-58; discussion 362-8.
2. van Dongen JM, van Wier MF, Tompa E, Bongers PM, van der Beek AJ, van Tulder MW, et al. Trial-based economic evaluations in occupational health: principles, methods, and recommendations. J Occup Environ Med. 2014;56(6):563-72.
3. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
4. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. BMJ. 2011;342:d1548.
5. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves--facts, fallacies and frequently asked questions. Health Econ. 2004;13(5):405-15.

VERSION 2 – REVIEW

REVIEWER	Aron Onerup Department of Surgery, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg
REVIEW RETURNED	06-Oct-2017

GENERAL COMMENTS	The article is well-written and easy to read. I appreciate some of the minor changes since first submission and think that the discussion is more well-balanced and in line with the results. As a novel to health economic studies I still find it hard to intuitively understand how an intervention that costs more than control (since control is doing everything that the intervention group receives except thet intervention that costs 80 Euros per participant) that is analyzed with regard to return to work where no statistically significant effect was found can be cost effective. My understanding is that this is due to the fact that total costs are lower in the intervention group due to e.g. lower costs for absenteeism from work, and that two combined non-significant effects result in a statistically significant effect. In my understanding there would be a risk of calculating the effect on return to work twice (both as decreased cost and as faster return to work). However, since health economy is not my field of work I accept that there is an effect that is well described and well argued for. I have no further suggestions.

REVIEWER	David Story University of Melbourne Australia No Competing Interest
REVIEW RETURNED	27-Sep-2017

GENERAL COMMENTS	The Authors have done a good job with this revision

REVIEWER	Nicolas Demartines Departement Surgery University Hospital CHUV, Lausanne Switzerland
REVIEW RETURNED	06-Oct-2017

GENERAL COMMENTS	MAJOR : 1. Statistical results should be analysed by a competent bio-statistician. 2. Concerning the reference n°12 ; not sure to understand the correlation between the article "Episodes of low back pain: a proposal for uniform definitions to be used in research" and the definition of Return to Work in GYN field. Is it a model we can transpose ? I can't find in the literature other mention of such a re-use of this definition. Other comment : Otherwise, the author has embedded the study limitation, which is sufficient now.

VERSION 2 – AUTHOR RESPONSE

Response to Reviewer 1

Comment: The article is well-written and easy to read. I appreciate some of the minor changes since first submission and think that the discussion is more well-balanced and in line with the results.
As a novel to health economic studies I still find it hard to intuitively understand how an intervention that costs more than control (since control is doing everything that the intervention group receives except thet intervention that costs 80 Euros per participant) that is analyzed with regard to return to work where no statistically significant effect was found can be cost effective. My understanding is that this is due to the fact that total costs are lower in the intervention group due to e.g. lower costs for absenteeism from work, and that two combined non-significant effects result in a statistically significant effect. In my understanding there would be a risk of calculating the effect on return to work twice (both as decreased cost and as faster return to work). However, since health economy is not my field of work I accept that there is an effect that is well described and well argued for. I have no further suggestions.

Response: Economic evaluations involve the joint distribution of costs and consequences and can demonstrate cost-effectiveness when neither cost nor effect differences are individually significant.1, 2

Double-counting was avoided by excluding the productivity costs due to sick leave in the ICER for duration until RTW.

Response to Reviewer 4

Comment1: Statistical results should be analysed by a competent bio-statistician.

Response: The analysis of the cost-effectiveness data was supervised by Dr. Judith E. Bosmans who is an associate professor Methodology in Cost-Effectiveness Research. She has ample expertise in the design, conduct and reporting of cost-effectiveness studies as reflected in the large number of publications she (co-)authored in this field (>70 publications). Moreover, she recently received a large personal grant to further develop methods for cost-effectiveness studies.

Comment2: Concerning the reference n°12 ; not sure to understand the correlation between the article "Episodes of low back pain: a proposal for uniform definitions to be used in research" and the definition of Return to Work in GYN field. Is it a model we can transpose ? I can't find in the literature other mention of such a re-use of this definition.

Response: The primary outcome of our study was duration until sustainable RTW, defined as the resumption of own work or other work with equal earnings, for at least 4 weeks without (partial or full) recurrence of sick leave. This definition was adopted as interventions aimed at expediting return to work of sick-listed employees should also aim at reducing recurrence of sickness absence in order to sustain employees at work after initial RTW. In theory, an intervention can lead to faster RTW compared to the alternative, but could simultaneously lead to higher rates of recurrences of sick-leave, leading to no additional value.

Initially, this definition was used to evaluate RTW-interventions in patients sick-listed due to low back pain and musculoskeletal disorders.3, 4 However, since then, the same definition has also been used in publications describing interventions aiming at RTW in patients sick-listed due to mental disorders5, gynaecological disease6 and in cancer survivors7.

We referred to the article by prof. de Vet (#12) as we wanted to emphasize the importance of uniform definitions to describe an outcome-parameter such as ‘duration until RTW’ in research. We have now removed this reference as we understand this might lead to confusion.

References used in this response letter

1. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. BMJ. 2011;342:d1548.
2. van Dongen JM, van Wier MF, Tompa E, Bongers PM, van der Beek AJ, van Tulder MW, et al. Trial-based economic evaluations in occupational health: principles, methods, and recommendations. J Occup Environ Med. 2014;56(6):563-72.
3. Lambeek LC, van Mechelen W, Knol DL, Loisel P, Anema JR. Randomised controlled trial of integrated care to reduce disability from chronic low back pain in working and private life. BMJ. 2010;340:c1035.
4. Vermeulen SJ, Anema JR, Schellart AJ, Knol DL, van Mechelen W, van der Beek AJ. A participatory return-to-work intervention for temporary agency workers and unemployed workers sick-listed due to musculoskeletal disorders: results of a randomized controlled trial. Journal of occupational rehabilitation. 2011;21(3):313-24.
5. Lammerts L, Vermeulen SJ, Schaafsma FG, van Mechelen W, Anema JR. Return to work of workers without a permanent employment contract, sick-listed due to a common mental disorder: design of a randomised controlled trial. BMC Public Health. 2014;14:594.
6. Vonk Noordegraaf A, Anema JR, van Mechelen W, Knol DL, van Baal WM, van Kesteren PJ, et al. A personalised eHealth programme reduces the duration until return to work after gynaecological surgery: results of a multicentre randomised trial. BJOG. 2014;121(9):1127-35; discussion 36.
7. van Egmond MP, Duijts SF, Jonker MA, van der Beek AJ, Anema JR. Effectiveness of a tailored return to work program for cancer survivors with job loss: results of a randomized controlled trial. Acta Oncol. 2016;55(9-10):1210-9.