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)	Effects of supervised exercise training during pregnancy on psychological well-being among overweight and obese women: Secondary analyses of the ETIP-trial, a randomized controlled trial.
AUTHORS	Garnæs, Kirsti; Helvik, AS; Stafne, Signe; Mørkved, Siv; Salvesen, Kjell; Salvesen, Øyvind; Moholdt, Trine

VERSION 1 – REVIEW

REVIEWER	Associate professor PhD, Hanne Kristine Hegaard Obstetric Clinic, Rigshospitalet Copenhagen Denmark
REVIEW RETURNED	12-Jan-2019

GENERAL COMMENTS	Effects of supervised exercise training during pregnancy on psychological well-being among overweight and obese women. Secondary analyses of the ETIP-trial, a randomized controlled trial. The authors conducted a trial to assess the effect of supervised exercise training during pregnancy on gestational weight gain as primary outcome in obese pregnant women. This paper clearly describes that this is a pre-specified secondary analysis of the Exercise Training in Pregnancy (ETIP) trial and refers to the Study Protocol as well as the main RCT study and two published papers with secondary analyses. Study protocol:Moholdt TT, Salvesen K, Ingul CB, Vik T, Oken E, Mørkved S. Exercise Training in Pregnancy for obese women (ETIP): study protocol for a randomised controlled trial. Trials. 2011 Jun 17;12:154. doi: 10.1186/1745-6215-12-154. PubMed PMID: 21682869; PubMed Central PMCID: PMC3148988. Main RCT Garnæs KK, Mørkved S, Salvesen Ø, Moholdt T. Exercise Training and Weight Gain in Obese Pregnant Women: A Randomized Controlled Trial (ETIP Trial). PLoS Med. 2016 Jul 26;13(7):e1002079. doi: 10.1371/journal.pmed.1002079. eCollection 2016 Jul. PubMed PMID: 27459375; PubMed Central PMCID: PMC4961392.

	Secondary analysis Garnæs KK, Mørkved S, Salvesen KÅ, Salvesen Ø, Moholdt T. Exercise training during pregnancy reduces circulating insulin levels in overweight/obese women postpartum: secondary analysis of a randomised controlled trial (the ETIP trial). BMC Pregnancy Childbirth. 2018 Jan 8;18(1):18. doi: 10.1186/s12884-017-1653-5. PubMed PMID: 29310617; PubMed Central PMCID: PMC5759335. Secondary analysis Garnæs KK, Nyrnes SA, Salvesen KÅ, Salvesen Ø, Mørkved S, Moholdt T. Effect of supervised exercise training during pregnancy on neonatal and maternal outcomes among overweight and obese women. Secondary analyses of the ETIP trial: A randomised controlled trial. PLoS One. 2017 Mar 21;12(3):e0173937. doi: 10.1371/journal.pone.0173937. eCollection 2017. PubMed PMID: 28323893; PubMed Central PMCID: PMC5360254. Introduction: The scientific background and explanation of the rationale are well described. Methods: Most of the elements in the methods, i.e. trial design, participants, interventions, outcomes, sample size, and randomization (sequence generation, allocation concealment mechanism and implementation), are well described and in line with the Methods section in the protocol study and the main RCT study. However some elements are not described sufficiently. The authors describe in line 36 page 6: “we asked the women to exercise at home for 50 minutes twice weekly”. In the main RCT it was stated: “In addition, the women were asked to follow a 50-min home exercise program at least once weekly (35 min of endurance training and 15 min of strength exercises) and to do daily pelvic floor muscle exercises”.: I recommend that the exercise regimen be described more consistently, as the authors refer to the main RCT study and the study protocol Outcomes: PGWBI The description of the PGWBI questionnaire is very sparse and in my opinion the readers need more information. Who developed the questionnaire? Is the questionnaire generic or not? Has it been applied and validated in studies among pregnant women before? Had the psychometric properties been investigated and how was it translated into Norwegian? In the description of PGWBI the authors refer only to Wenger NK, who uses several methods to assess quality of life, including PGWBI. Wenger NK described the 6 subscales (or dimensions) as bodily distress, life satisfaction, sense of vitality, cheerful vs

	distressed, relaxed vs anxious, and self-control. This differs from the description in this paper. I do not think that Wenger NK developed the PGWBI questionnaire, please refer to the author, who developed PGWBI. I suggest that the authors revise their description of PGWBI. Please note the reference number 29 Gustaffson et al. (BJOG 2016). Gustaffson et al. also used the PGWBI questionnaire in their study. Perhaps it also relevant to refer to this paper. EPDS The authors refer to an article by Cox et al. (reference number 20). Cox et al. developed the 10-item Edinburgh Postnatal Depression Scale. Is the questionnaire generic or not? Has it been applied in studies among pregnant women in Norway before and has it been validated among Norwegian women? Had the psychometric properties been investigated and how was it translated into Norwegian? In page 7 line 21 it is stated: “The EPDS contains 10 questions assessing how the woman is coping with life changes related to pregnancy and childbirth”. In the paper by Cox et al. page 786 they wrote: “We would like to know how you are feeling. Please UNDERLINE the answer which comes closest to how you have felt in the past 7 days.” Please re-read the article by Cox et al. and consider if it is more correct to describe that the 10 questions assess feeling and not coping with life changes. In line 33, page 7 the authors write: “In addition, if the participant scored 1, 2 or 3 on question number 10, she was classified as “Suicidal risk”. Please explain where I can read this in the article by Cox et al. I suggest that authors also include an outcome related to the proportion of women with risk of depression, employing one of the cut-offs recommended by Cox et al. Other studies use both a total EPDS score and a cut-off score (please see Dodd JM, Newman A, Moran LJ, Deussen AR, Grivell RM, Yelland.LN et al. The effect of antenatal dietary and lifestyle advice for women who are overweight or obese on emotional well-being: the LIMIT randomized trial. Acta ObstetGynecol Scand 2016; 95:309–318). Page 7, line 37. The authors refer to perceived health status” and in Table 1 to “self-perceived health general status”. Please observe consistency and provide further information to facilitate the reader’s understanding of where this question came from The question about self- perceived health is well-known but has no reference. Please refer to the author, who developed the question. Randomization and blinding: Page 7, line 55. “Randomization was performed before baseline assessments using a computer random number generator.” However, it seems that randomization was done after baseline assessment. Please see the main RCT in PLoS One: “After

	baseline assessments, the participants were randomly allocated 1:1 to the intervention or the control group. Allocation was done using a computer random number generator developed and administrated at the Unit for Applied Clinical Research, NTNU". Please provide further information so that readers can understand when and how randomization took place. Table 1, page 10. The analysis for self-perceived general health status is not described. It seems unclear why parity and education are analyzed by Pearson Chi-Square Test when there is no or only one person in some categories. Please clarify. Page 13, line 5. "We found no statistically significant difference in postnatal depression three months postpartum." It is more correct to write that there was no difference in total EPSD score, since a mean score is not an expression of depression. Please include a cut-off value of EPDS as an outcome and provide information about this in the results section. Discussion: Main findings Please be careful about using the term "postnatal depression" Using the term symptoms of depressions seems more correct. Strengths and limitations Please discuss why you do not use multiple imputation. Please discuss in detail why you recommended pregnant women to participate in an exercise program three times weekly and at least one time a week at home.

REVIEWER	Dr Sara Holton Deakin University, Australia
REVIEW RETURNED	05-Mar-2019

GENERAL COMMENTS	The aim of this paper is to determine whether supervised exercise training during pregnancy has an effect on psychological well-being among women who have a high pre-pregnancy body mass index. Although this is an important paper I think it requires minor revisions before it is acceptable for publication in BMJ Open. General Please have the paper reviewed for English expression by a native English speaker. Please use person-first language eg women who are overweight or obese, women with a high BMI etc. Please define psychological wellbeing - for the purposes of your paper does this mean symptoms of anxiety and depression? Abstract Please state how many women were in the intervention (exercise) group and how many were in the control group. Introduction First paragraph: the order of the points in this paragraph could be improved. I would suggest first identifying the prevalence etc of mental health problems in pregnant women who have a high BMI

	and then discussing the other implications/risks of a high BMI during pregnancy. Or split the paragraph into two separate smaller paragraphs: one about mental health and high BMI during pregnancy and the other about the risks of high BMI for mothers and their babies. Methods BMI: how was pre-pregnancy BMI determined? Self-report? Please state/cite which weight classification system was used to categorise BMI. PGWBI: Please state if this measure has been validated for use among pregnant women. What is the timeframe for the PGWBI? ie does it assess wellbeing over the last week, month etc? Is it self-report? EPDS: The EPDS has a scoring guide and some items are reversed scored. Have you followed this guide? It doesn't seem like it from what has been written. Also need to state that EPDS is self-report measure and it doesn't measure 'coping' but rather screens women for symptoms of emotional distress during pregnancy and the postnatal period. Please ensure that it is only stated that the EPDS measures SYMPTOMS of depression/anxiety. Sociodemographic characteristics: how and when were women's sociodemographic details collected? I assume that only the pregnancy questionnaires were completed during the glucose tolerance test. When/how were the postnatal questionnaires completed? Discussion Please state why you think there was such low adherence to the exercise protocol. Conclusion Please justify why you think we need more research in this area when you and others have found there is no effect.

REVIEWER	Rahim Moineddin University of Toronto, Canada
REVIEW RETURNED	02-May-2019

GENERAL COMMENTS	Comments on Effects of supervised exercise training during pregnancy on psychological well-being among overweight and obese women. Secondary analyses of the ETIP-trial, a randomized controlled trial. This is a two arm small RCT (46 vs 45) negative study. For a negative study post hoc power analysis is required. For example in Table 2 the 95% confidence intervals for difference between PGWBI and Anxiety are fairly wide therefore a post hoc power analysis will be a safe guard against strong conclusive tone of the paper. Using Marginal Structure Models (inverse-probability-of-treatment weighting) is a more powerful method for taking into account the effects of adherence to treatment on the outcome. See for example Marginal structural models in clinical research: when and how to use them? By Tyler Williamson and Pietro Ravani

	Nephrology Dialysis Transplantation, Volume 32, Issue suppl_2, April 2017, Pages ii84–ii90, https://doi.org/10.1093/ndt/gfw341

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

8. The authors describe in line 36 page 6: “we asked the women to exercise at home for 50 minutes twice weekly”. In the main RCT it was stated: “In addition, the women were asked to follow a 50-min home exercise program at least once weekly (35 min of endurance training and 15 min of strength exercises) and to do daily pelvic floor muscle exercises”.: I recommend that the exercise regimen be described more consistently, as the authors refer to the main RCT study and the study protocol.

We apologize for this miswriting in our manuscript, this sentence is changed to “at least once weekly” in accordance to the main RCT, in the Method section line 197.

Outcomes:
PGWBI

9. The description of the PGWBI questionnaire is very sparse and in my opinion the readers need more information. Who developed the questionnaire? Is the questionnaire generic or not? Has it been applied and validated in studies among pregnant women before? Had the psychometric properties been investigated and how was it translated into Norwegian?

Thank you for giving us the opportunity to improve our description of the PGWBI questionnaire. We agree upon our description being too sparse. As to our knowledge the PGWBI questionnaire has not been validated among pregnant women, however it has been used by Gustaffson and colleagues (BJOG 2016). We have now added a more thorough description of the PGWBI questionnaire with new references in the Method section, Outcomes, line 208-244: Psychological well-being was assessed by the “Psychological General Well-Being Index” (PGWBI) questionnaire (PGWBI © 1984 Harold J. Dupuy, Mapi Research Trust)5,6 at baseline (gestational week 12-18), in late pregnancy (gestational week 34-37) and three months postpartum. The Psychological General Well-being (PGWB) scale measures the last week self-perceived psychological health and general well-being, and intends to assess health related quality of life or, said otherwise, to reflect a sense of well-being or distress that includes positive as well as negative intrapersonal affective or emotional states.35 It consists of 22 items with a six-point self-response scale that ranges from zero (= most negative option) to five (= most positive option) and includes six non overlapping dimensions: anxiety (five items), depressed mood (three items), positive well-being (four items), self-control (three items), general health (three items), and vitality (four items).35, 38 Each dimension is summed and the total (maximum = 110) forms the overall PGWB index. The anxiety dimension assessed whether the subjects were bothered by nervousness, were generally tense, anxious, worried or upset, and/or under stress strain or pressure. Depressed-mood measured if the participants were depressed, hopeless, or downhearted and ‘blue’. Positive well-being indicated the general spirit, cheerfulness, or happiness and satisfaction with personal life. The self-control dimension intended to measure whether the subjects felt emotionally stable, in firm control, or afraid of losing control. General health assessed if the subjects were bothered with pain, disorder, or illness and whether they were healthy enough ‘to do things.’ Finally, vitality contained items that assessed the participants’ energy, whether they felt active, vigorous, or sluggish, tired and worn out.35 The PGWBI questionnaire is a generic questionnaire frequent used in clinical trials across many conditions, and translated to several

languages.7-10 The PGWBI has been found suitable for subjects 14-90 years and is a highly preferred self-administered inventory.35 The internal consistency reliability is high with Cronbach´s alpha correlations between 0.90 and 0.94.35 Similar correlations (Cronbach´s alpha > 0.90) have been shown for tests done in Sweden,43 with culture and language similar to the Norwegian,8 and recently used by Gustafsson and colleges in a clinical trial among Norwegian pregnant women.11 The present Norwegian version of the questionnaire was translated by a standard forward-backward method at St. Olavs Hospital, the university Hospital, Trondheim, Norway, in February 2002.12

10. In the description of PGWBI the authors refer only to Wenger NK, who used several methods to assess quality of life, including PGWBI. Wenger NK described the 6 subscales (or dimensions) as bodily distress, life satisfaction, sense of vitality, cheerful vs distressed, relaxed vs anxious, and self-control. This differs from the description in this paper. I do not think that Wenger NK developed the PGWBI questionnaire, please refer to the author, who developed PGWBI. I suggest that the authors revise their description of PGWBI.

We have now reviewed our use of the reference of Wenger, changed our references and rewritten the description of PGWBI. Please see our answer in point 9 above.

EPDS

The authors refer to an article by Cox et al. (reference number 20). Cox et al. developed the 10-item Edinburgh Postnatal Depression Scale.

11. Is the questionnaire generic or not? Has it been applied in studies among pregnant women in Norway before and has it been validated among Norwegian women? Had the psychometric properties been investigated and how was it translated into Norwegian?

The EPDS questionnaire is a non-generic questionnaire designed to measure depressive symptoms among women in the postpartum period. The EPDS has been translated to Norwegian, and has been validated in a Norwegian population. The EPDS psychometric properties has been found to be good in a postpartum population, and have been used in both clinical and epidemiological studies.13 We have improved our description of the EPDS questionnaire in our manuscript, in the Method section, Outcomes, line 245-277: To measure the prevalence of symptoms of postnatal depression, the participants also completed the “Edinburgh Postnatal Depression Scale” (EPDS) questionnaire (Cox, Holden and Sagosky, 1987).13 The questionnaire is a non-generic self-rating scale, which measures the presence of depressive symptoms during in the postpartum period, indicating how the mother has felt during the last week.13,14 The EPDS questionnaire contains 10 questions. All questions contain four response alternatives were the women are asked to “please underline the answer which comes closest to how you have felt in the past 7 days”.13 And line 220-223: The EPDS questionnaire is developed and commonly used for measurement of depressive symptoms in the postpartum period, but is also used and validated for the pregnancy period.13 The questionnaire is translated to Norwegian and found valid to detect postpartum depression in a Norwegian population.15,16

12. In page 7 line 21 it is stated: “The EPDS contains 10 questions assessing how the woman is coping with life changes related to pregnancy and childbirth”. In the paper by Cox et al. page 786 they wrote: “We would like to know how you are feeling. Please UNDERLINE the answer which comes closest to how you have felt in the past 7 days.” Please re-read the article by Cox et al. and consider if it is more correct to describe that the 10 questions assess feeling and not coping with life changes.

We agree upon this being a more correct description, and have rewritten two sentences in the Method section, Outcomes line 247-248: which measures the presence of depressive symptoms during in the

postpartum period, indicating how the mother has felt during the last week.13,14 and line 249-250: all questions contain four response alternatives were the women are asked to “please underline the answer witch comes closest to how you have felt in the past 7 days”.13

13. In line 33, page 7 the authors write: “In addition, if the participant scored 1, 2 or 3 on question number 10, she was classified as “Suicidal risk”. Please explain where I can read this in the article by Cox et al.

We have now deleted the part about suicidal risk from the manuscript after revisiting the original paper by Cox et al. We have also re-analyzed the data according to the originally proposed cut-off values, presented in Methods section, line 253-254, and in Results section line 399-402.

14. I suggest that authors also include an outcome related to the proportion of women with risk of depression, employing one of the cut-offs recommended by Cox et al. Other studies use both a total EPDS score and a cut-off score (please see Dodd JM, Newman A, Moran LJ, Deussen AR, Grivell RM, Yelland.LN et al. The effect of antenatal dietary and lifestyle advice for women who are overweight or obese on emotional well-being: the LIMIT randomized trial. Acta ObstetGynecol Scand 2016; 95:309–318).

We thank you for making us aware of these other studies and have changed the analyses and manuscript accordingly in Results, line 399-402: No women in either the exercise group or the control group reported of a total EPSD score of 13 or more, representing indication of major depression. Two women (7.1%) in the exercise group, and three women (10.3%) in the control group reported at total EPSD score between 10 and 12, representing indication of a minor depression ($p = 0.97$).

15. Page 7, line 37. The authors refer to perceived health status” and in Table 1 to “self-perceived health general status”. Please observe consistency and provide further information to facilitate the reader’s understanding of where this question came from The question about self- perceived health is well-known but has no reference. Please refer to the author, who developed the question.

Thank you for making us aware of inconsistency in wording in the text vs wording Table 1. “Self-perceived general health status” is the correct term. This is now corrected in Methods, Outcomes, line 278, corresponding to the expression used in Table 1, page 13. The question regarding self-perceived general health status, is taken from the SF Short Form Health Survey, consisting of 12 questions regarding quality of health. This text is added to the Method section, Outcomes, line 280-281: This question is taken from the SF 36 Short Form Health Survey. This survey is translated to Norwegian and tested for reliability and validity in a Norwegian population.

Randomization and blinding:

16. Page 7, line 55. “Randomization was performed before baseline assessments using a computer random number generator.” However, it seems that randomization was done after baseline assessment. Please see the main RCT in PLoS One: “After baseline assessments, the participants were randomly allocated 1:1 to the intervention or the control group. Allocation was done using a computer random number generator developed and administrated at the Unit for Applied Clinical Research, NTNU”.
Please provide further information so that readers can understand when and how randomization took place.

Thank you very much for making us aware of this incorrect statement in our manuscript. The randomization was done per procedure AFTER the baseline assessments, as stated in our published protocol, the main RCT article published in Plos Medicine1 as well as in our two previous secondary analyses publications from the trial. This is now corrected from “before” to “after” in Method section, Randomization and blinding, line 293.

17. Table 1, page 10. The analysis for self-perceived general health status is not described. It seems unclear why parity and education are analyzed by Pearson Chi-Square Test when there is no or only one person in some categories. Please clarify.

We are sorry for this lack of description. Self-perceived general health status was analyzed by Fisher’s exact test. This was the case also for parity and education because of 2 x 2 tables and a limited number of participants in each category. This has now been corrected in the statistical methods, line 299-300: Comparisons between groups at baseline was analysed by independent samples t-tests and Fisher’s exact tests, and in Table 1. Difference between groups at late pregnancy was analyzed by general linear model analysis of covariance, this is now clarified in the Method section, Statistical methods, line 304-306: … and between-groups difference in “Self-perceived general health status” late pregnancy by general linear model analysis of covariance.

18. Page 13, line 5. “We found no statistically significant difference in postnatal depression three months postpartum.” It is more correct to write that there was no difference in total EPSD score, since a mean score is not an expression of depression. Please include a cut-off value of EPDS as an outcome and provide information about this in the results section.

We agree upon this and have changed the sentence “no statistically significant change in postnatal depression” to “no statistically significant difference in total EPSD score” in Results section, line 395. The cut-off values is now changed in accordance to Cox, to between 10 and 12, and ≥ 13, presented in the Method section line 253-254, and data presented in the Results section line 399-402. Please see our respond to point 14.

Discussion:

Main findings

19. Please be careful about using the term “postnatal depression” Using the term symptoms of depressions seems more correct.

We agree that your suggestion is a better term and have corrected “postnatal depression” to “symptoms of postnatal depression” in Discussion section line 429, and Conclusion section line 520.

Strengths and limitations

20. Please discuss why you do not use multiple imputation.

We discussed using multiple imputation, but decided not to. We have made an imputation model, but in these analyses, we got very large confidence intervals for the estimates. We therefore think that the imputations are not able to correctly predict the missing values and chose to not impute. Even if we got lower p-values when we imputed the missing values, we think this model will increase the risk of type 1 error in our findings. We are happy to include this discussion in the manuscript, if wanted, but have not done so in our first revision.

21. Please discuss in detail why you recommended pregnant women to participate in an exercise program three times weekly and at least one time a week at home.

We are unsure whether you ask us to include a general discussion of the health benefits of physical activity/exercise training in pregnancy. We have, so far, not included such a general discussion. Our intervention program is based on international and national recommendations for physical activity in pregnancy; that pregnant women, also previously inactive and obese women, should perform regular physical activity, and exercise at moderate intensity 20-30 minutes daily, on the most/all days of the week.17 The ACOG guidelines also recommends a combination of moderate endurance exercise and light strength training. Physical activity and exercise training are found to be safe for both the mother and child,17-19 even for women with risk factors as GDM, chronic hypertension, and overweight/obesity.20 It is important to notice that the ETIP trial intervention program, not is stated as our recommendation, but are based on current recommendations for preventing pregnancy related complications. The aim of the ETIP trial was to investigate whether offering pregnant women an exercise program following these recommendations, could improve health outcomes.

Reviewer: 2

22. Please use person-first language eg women who are overweight or obese, women with a high BMI etc.

We have now read through our manuscript and replaced most sentences to person-first language. The wording is corrected in Abstract section line 30 and 62, Introduction section line 90-91, 105, 150, Discussion section line 429, 475-476, 512, 520.

23. Please define psychological wellbeing - for the purposes of your paper does this mean symptoms of anxiety and depression?

We have revised our definition of psychological well-being in our background, line 91-93,:
“Psychological well-being is for the purpose of this paper defined as “people’s cognitive and affective evaluations of their lives; happiness, absence of negative emotions (e.g. depression, anxiety), satisfaction with life, and positive functioning”.21

Abstract

24. Please state how many women were in the intervention (exercise) group and how many were in the control group.

Thank you for notifying us about this lack of information. This information is now provided in Abstract section line 37-38: Ninety-one women (age 31.2±4.1 years, BMI 34.5±4.2 kg/m2), 46 in the exercise group and 45 in the control group, were included in the trial.

Introduction

25. First paragraph: the order of the points in this paragraph could be improved. I would suggest first identifying the prevalence etc of mental health problems in pregnant women who have a high BMI and then discussing the other implications/risks of a high BMI during pregnancy.
Or split the paragraph into two separate smaller paragraphs: one about mental health and high BMI during pregnancy and the other about the risks of high BMI for mothers and their babies.

Thank you for making us improve our manuscript. The order of the points in the first paragraph of the introduction is now changed, lines 89-99.

Methods

26. BMI: how was pre-pregnancy BMI determined? Self-report?

Pre-pregnancy BMI was based on self-reported information. This information is added in Method section, Participants line 176-177: Pre-pregnancy BMI was based on self-reported information.

27. Please state/cite which weight classification system was used to categorise BMI.

Thank you for making us add this precision to our manuscript. We used the WHO classification system. This is now added to the Method section, Participants, line 175-176: Categorisation of overweight and obesity related to BMI, was based on the World Health Organization (WHO) classification system.22

28. PGWBI: Please state if this measure has been validated for use among pregnant women. What is the timeframe for the PGWBI? ie does it assess wellbeing over the last week, month etc? Is it self-report?

We have improved our description of the PGWBI questionnaire and added information regarding validation in line 208-244: The PGWBI questionnaire is a generic questionnaire frequent used in clinical trials across many conditions, and translated to several languages.7-10 The PGWBI has been found suitable for subjects 14-90 years and is a highly preferred self-administered inventory.35 The internal consistency reliability is high with Cronbach´s alpha correlations between 0.90 and 0.94.35 Similar correlations (Cronbach´s alpha > 0.90) have been shown for tests done in Sweden,43 with culture and language similar to the Norwegian,8 and recently used by Gustafsson and colleges in a clinical trial among Norwegian pregnant women.11 The time-frame for PGWBI is “the last week”, which is added to the Method section, Outcomes line 209: The Psychological General Well-being (PGWB) scale measures the last week self-perceived psychological health and general well-being, …

29. EPDS: The EPDS has a scoring guide and some items are reversed scored. Have you followed this guide? It doesn't seem like it from what has been written.
Also need to state that EPDS is self-report measure and it doesn't measure 'coping' but rather screens women for symptoms of emotional distress during pregnancy and the postnatal period. Please ensure that it is only stated that the EPDS measures SYMPTOMS of depression/anxiety.

We have followed the EPDS scoring system and are aware that some of the items are reversed scored. We have not used the term “reversed scored” in our description of the scoring system, however used the expression: We estimated total score of the ten items with use of a scoring system 0-3, with “0” representing the most negative option, and “3” the most positive option, which we think explains the scoring system properly. We agree upon stating that the EPDS measures symptoms of depression and anxiety and have made this more clear in our Method section, Outcomes line 247-249: The questionnaire is a non-generic self-rating scale, which measures the presence of depressive symptoms during in the postpartum period, indicating how the mother has felt during the last week.13,14

30. Sociodemographic characteristics: how and when were women's sociodemographic details collected?

We agree that this needed a clarification. Data on sociodemographic characteristics were collected by self-reported questionnaires together with the other questionnaires during the glucose intolerance test at the baseline assessments. This information is added to the Method section line 203-207: Sociodemographic data was collected by self-reported questionnaires at baseline assessments. Information regarding the participants psychological well-being and risk of postnatal depression was assessed by self-reported questionnaires, completed at the hospital while they underwent a 2 h oral

glucose tolerance test at baseline, late pregnancy and three months postpartum, with trial researchers available to clarify questions if needed.

31. I assume that only the pregnancy questionnaires were completed during the glucose tolerance test. When/how were the postnatal questionnaires completed?

All the questionnaires were completed during the glucose tolerance test at the hospital, and the glucose tolerance test was also performed three months postpartum. This is now more clearly stated in Method section line 203-207, as stated in above comment 30.

32. Discussion
Please state why you think there was such low adherence to the exercise protocol.

We think the low adherence to the exercise protocol have several explanations. For many women, the first trimester of the pregnancy is characterized by discomfort and a high risk of abortion, and previously inactive women may be anxious to increase their physical activity. The pregnancy can also be a time of emotional changes that may impair their motivation for lifestyle changes.23 We also think that limited previous experience with exercise training combined with difficulties in prioritizing time for exercise might be an important factor reducing the adherence. The intervention protocol might have been too comprehensive for the most unexperienced women, however the exercise program was individually adjusted continually during the intervention period. We have added this sentence in the Discussion section line 443-446: The low adherence to the exercise protocol may be explained by various types of pregnancy-related discomfort, by the women being anxious to exercise, by having trouble with prioritizing time for exercise, and by having lack of motivation for lifestyle changes.

33. Conclusion
Please justify why you think we need more research in this area when you and others have found there is no effect.

Most trials assessing the effect of maternal exercise on mental well-being are limited by sample size and adherence to exercise protocol, and the true effect of intervention is difficult to detect. Therefore, we think that more statistically well-powered trials are needed in this field before we can state no effect of regular exercise on maternal mental health. We have added and changed our last sentence in the Conclusion section line 522-524: We need high sample-size trials with sufficient adherence to intervention protocols to be able to investigate the true effect of exercise during pregnancy on maternal well-being, and to examine factors associated with motivation for exercise during pregnancy.

Reviewer: 3

34. This is a two arm small RCT (46 vs 45) negative study. For a negative study post hoc power analysis is required. For example in Table 2 the 95% confidence intervals for difference between PGWBI and Anxiety are fairly wide therefore a post hoc power analysis will be a safe guard against strong conclusive tone of the paper.

Probably, our study is not powered enough to find the true effect of being offered exercise during pregnancy on mental well-being. However, we have chosen not to perform a post hoc power analyses for these secondary analyses. We are of that opinion that power-analysis are most suited when planning trials, not in analyzing the data. We agree with the reviewer regarding the wide confidence interval, making it difficult to estimate an effect. However, we think the nature of the outcomes PGWBI and Anxiety and the characteristics of the current study-population, makes wide confidence intervals expected, even with a significantly higher sample size. Looking at the observed differences in PGWBI score between the groups, it is not likely these differences represents clinical important differences in

mental well-being. If the reviewers still want us to perform a post-hoc power analysis, we will of course provide our manuscript with such information. We have modified a somewhat conclusive tone in our manuscript, by rewriting some sentences: Abstract, line 60-61: We found no statistically significant effect of supervised exercise training during pregnancy on psychological well-being in late pregnancy or postpartum. Discussion section, line 435-436: Only 50% of the women in the exercise group followed the exercise-protocol, and we included less participants than estimated in the trial protocol. This may have reduced the possibility of finding a true effect of the intervention. Conclusion section, 514-516: Low adherence to the exercise protocol may have reduced the chance of finding an effect of regular maternal exercise on mental health.

35. Using Marginal Structure Models (inverse-probability-of-treatment weighting) is a more powerful method for taking into account the effects of adherence to treatment on the outcome. See for example Marginal structural models in clinical research: when and how to use them? By Tyler Williamson and Pietro Ravani Nephrology Dialysis Transplantation, Volume 32, Issue suppl_2, April 2017, Pages ii84–ii90, https://doi.org/10.1093/ndt/gfw341

Thank you for informing us about this analyze model, and for making us reconsider our statistical methods. We have read through the article by Williamson and Ravani, and clearly see the advantages the Marginal Structure Models would have provided our analyses. As on your request we have discussed MSM with our trial statistician (now included as a co-author of this manuscript). We are unsure of what effect MSM analyses could have on our trial analyzing principle; intention to treat. We also think there are a high risk of bias regarding cause and effect relation between adherence – and effect of intervention, in addition some maternal complications may occur, regardless of adherence to exercise protocol, which may significantly affect exercise adherence, but also trial outcomes. We have not used the MSM analyses in our previous papers in the ETIP study, and would like to not perform these analyses for the current manuscript. However, we will review our statement, and will be able to provide new analyses if needed.

References

1. Garnaes KK, Morkved S, Salvesen O, Moholdt T. Exercise Training and Weight Gain in Obese Pregnant Women: A Randomized Controlled Trial (ETIP Trial). PLoS Med. 2016;13(7):e1002079.
2. Moholdt TT, Salvesen K, Ingul CB, Vik T, Oken E, Morkved S. Exercise Training in Pregnancy for obese women (ETIP): study protocol for a randomised controlled trial. Trials. 2011;12:154.
3. Christiansen T, Paulsen SK, Bruun JM, Pedersen SB, Richelsen B. Exercise training versus diet-induced weight-loss on metabolic risk factors and inflammatory markers in obese subjects: a 12-week randomized intervention study. Am J Physiol Endocrinol Metab. 2010;298(4):E824-831.
4. Wolff S, Legarth J, Vangsgaard K, Toubro S, Astrup A. A randomized trial of the effects of dietary counseling on gestational weight gain and glucose metabolism in obese pregnant women. Int J Obes (Lond). 2008;32(3):495-501.
5. Wenger NK, Mattson ME, Furberg CD, Elinson J. Assessment of quality of life in clinical trials of cardiovascular therapies. Am J Cardiol. 1984;54(7):908-913.
6. HJ. D. The Psychological general Well-Being (PGWB) Index. In: Assessment of Quality of Life in clinical trials of cardiovascular therapies. In: Wenger NK MM, Furberg CD, Elinson J. Le Jacq ed. Assessment of Quality of Life in clinical trials of cardiovascular therapies. .1984:170-183.
7. Naughton MJ, Wiklund I. A critical review of dimension-specific measures of health-related quality of life in cross-cultural research. Qual Life Res. 1993;2(6):397-432.

8. Lundgren-Nilsson A, Jonsdottir IH, Ahlborg G, Jr., Tennant A. Construct validity of the Psychological General Well Being Index (PGWBI) in a sample of patients undergoing treatment for stress-related exhaustion: a Rasch analysis. Health and quality of life outcomes. 2013;11:2.
9. Gaston JE, Vogl L. Psychometric properties of the general well-being index. Qual Life Res. 2005;14(1):71-75.
10. Wiklund I, Karlberg J. Evaluation of quality of life in clinical trials. Selecting quality-of-life measures. Control Clin Trials. 1991;12(4 Suppl):204s-216s.
11. Gustafsson MK, Stafne SN, Romundstad PR, Morkved S, Salvesen K, Helvik AS. The effects of an exercise programme during pregnancy on health-related quality of life in pregnant women: a Norwegian randomised controlled trial. BJOG. 2016;123(7):1152-1160.
12. Helvik AS, Jacobsen G, Hallberg LR. Psychological well-being of adults with acquired hearing impairment. Disabil Rehabil. 2006;28(9):535-545.
13. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry. 1987;150:782-786.
14. Cox J. Use and misuse of the Edinburgh Postnatal Depression Scale (EPDS): a ten point 'survival analysis'. Archives of women's mental health. 2017;20(6):789-790.
15. Eberhard-Gran M, Eskild A, Tambs K, Schei B, Opjordsmoen S. The Edinburgh Postnatal Depression Scale: validation in a Norwegian community sample. Nordic journal of psychiatry. 2001;55(2):113-117.
16. Berle JO, Aarre TF, Mykletun A, Dahl AA, Holsten F. Screening for postnatal depression. Validation of the Norwegian version of the Edinburgh Postnatal Depression Scale, and assessment of risk factors for postnatal depression. J Affect Disord. 2003;76(1-3):151-156.
17. ACOG Committee Opinion No. 650: Physical Activity and Exercise During Pregnancy and the Postpartum Period. Obstet Gynecol. 2015;126(6):e135-142.
18. Hellenes OM, Vik T, Lohaugen GC, et al. Regular moderate exercise during pregnancy does not have an adverse effect on the neurodevelopment of the child. Acta Paediatr. 2015;104(3):285-291.
19. Di Mascio D, Magro-Malosso ER, Saccone G, Marhefka GD, Berghella V. Exercise during pregnancy in normal-weight women and risk of preterm birth: a systematic review and meta-analysis of randomized controlled trials. Am J Obstet Gynecol. 2016;215(5):561-571.
20. Wiebe HW, Boule NG, Chari R, Davenport MH. The effect of supervised prenatal exercise on fetal growth: a meta-analysis. Obstet Gynecol. 2015;125(5):1185-1194.
21. Diener E. Subjective well-being. The science of happiness and a proposal for a national index. Am Psychol. 2000;55(1):34-43.
22. World Health Organization W. Obesity and overweight. 2015.
23. Reading AE, Cox DN, Sledmere CM, Campbell S. Psychological changes over the course of pregnancy: a study of attitudes toward the fetus/neonate. Health Psychol. 1984;3(3):211-221.

VERSION 2 – REVIEW

REVIEWER	Dr Sara Holton Deakin University Australia
REVIEW RETURNED	11-Jul-2019

GENERAL COMMENTS	I believe the authors have adequately responded to the reviewers' suggestions. However, I still have concerns about the English expression throughout the manuscript and think that it requires review by a native English speaker to ensure that expression and grammar etc are appropriate. Once this has been done I think the manuscript is suitable for publication.

REVIEWER	Rahim Moineddin University of Toronto Canada
REVIEW RETURNED	08-Jul-2019

GENERAL COMMENTS	Authors addressed my comments.