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)	Efficacy of a standardised acupuncture approach for women with bothersome menopausal symptoms: a pragmatic randomised study in primary care (the ACOM study)
AUTHORS	Lund, Kamma; Siersma, Volkert; Brodersen, John; Waldorff, Frans

VERSION 1 – REVIEW

REVIEWER	Prof. Masakazu Terauchi Department of Women's Health, Tokyo Medical and Dental University, Tokyo, Japan
REVIEW RETURNED	09-May-2018

GENERAL COMMENTS	This is a report of the efficacy of a standardized brief “Western-style” acupuncture for women with moderate-to-severe menopausal hot flushes. Although the paper is of clinical relevance, the reviewer has a couple of concerns. #1 The biggest flaw of the study is in its design. It is perplexing that the authors limited their analysis to the first six weeks of the study even though they seem to have a good chance to compare the two groups in a crossover fashion (Appendix 2). Considering the lack of an adequate placebo comparator as the authors describe, a crossover analysis would have been a perfect fit. #2 The questionnaire used in the study is not displayed in the paper. Appendix 1 is not the questionnaire, but the scoring system.

REVIEWER	Boudewijn Kollen University Medical Center Groningen, Netherlands
REVIEW RETURNED	18-Jun-2018

GENERAL COMMENTS	The manuscript, entitled “Efficacy of a standardized acupuncture approach for woman with bothersome menopausal symptoms: A randomized study in primary care (the ACOM study)” by Kamma Sundgaard Lund, Volkert Dirk Siersma, John Brodersen, Frans Boch Waldorff reports on an investigation on the efficacy of a brief acupuncture approach using scales of the MenoScores as outcomes in 70 women. I have only a few comments to make that basically serve to provide a better understanding of the choices you made in this study. 1. There is no information given in this manuscript whether patients used any self-chosen medication or were subjected to an alternative treatment to ameliorate the symptoms during the control period. Was the patient instructed to keep a diary for this

	information? This information will facilitate the interpretation of the results and its functional implications. 2. Your study sample does not appear to be a random sample of patients with menopausal symptoms. Please, clarify in your paper the generalizability of your findings from this more or less selective population. 3. All treatments were given by GPs that were also certified acupuncturists. This may have introduced some bias as these "believers" are likely to favour the intervention treatment and as a result (subconsciously) put more effort and conviction in this treatment possibly intensifying any placebo effect. 4. Linear mixed models were used to analyze study results. It is not clear whether assumptions were met and which variables were used to control for the clustering of information. The limited sample size must have affected the distributions of your basically ordinal scaled data if I understand correctly. Shouldn't you report medians instead of means in Appendix 4? You used continuous scaled information for your sample size calculation. Please, justify. Did you transform data in order to comply with the assumptions in the analyses? 5. Why did you decide not to analyze the longitudinal results all together? In my view that information is more informative than that of only change scores between baseline and 6 weeks. Moreover, the ICH/GCP guidelines stipulate that a correction for baseline is necessary when analyzing change scores as they state that … the use of change from baseline without adjusting for baseline does not generally constitute an appropriate covariate adjustment …. Please, justify your choice for not adding the baseline variable of outcome as a covariate. 6. You report statistically significant results but are these results also clinically relevant? Please, substantiate your claim. 7. I commend you for implementing a correction for multiplicity even though strictly speaking this correction is not called for when you only have one primary endpoint.

REVIEWER	Robin Prescott University of Edinburgh, Medical Statistics Unit
REVIEW RETURNED	08-Aug-2018

GENERAL COMMENTS	I am reviewing this paper as a statistician. This is a well-designed study and much of presentation is good. I have no concerns about the validity of the conclusions from this study but I am not sure that the most efficient analysis has always taken place. My lack of certainty is because the descriptions of the analyses are sometimes lacking in sufficient detail to be sure of precisely what was done. Firstly though, I would like to comment about the title of the paper. This is a pragmatic randomised controlled trial and it would be helpful to potential readers to be aware of that from the title. Within the section on randomisation, there needs to be a little bit of detail about the stratification by age. What age categories were used in the stratification? From Table 2 I would guess that the categories are 40-55 and 56-65 but this should be stated explicitly. The randomisation itself looks sound. Apart from the very minor comment above, I found all of the sections up to 'Statistical methods' were clear. The primary outcome was clearly defined, as were the secondary outcomes.

	However, the description of the linear mixed models was insufficiently detailed. The description indicates that subjects were fitted as a random effect with the randomisation factors of age and level of symptoms appropriately included as covariates. It is perhaps implicit that age was fitted using the categories utilised in the randomisation, but perhaps that could be made explicit. There is, however, no indication that the baseline level of the outcome variable was included as a covariate, as would be necessary for a fully efficient analysis. It is not even clear whether a separate model was used for each of the two follow-up times or whether there was a unified analysis with terms being fitted for treatment, visit and visit by treatment interactions and possibly visit by baseline interactions to allow for a different baseline effect at each visit. The advantage of the unified model is that the data from the subject who was present for a three month assessment but not the six month assessment would still contribute to the estimate of treatment effects at six months, thereby gaining a little efficiency. I am not suggesting that the unified analysis is essential but what is essential is clarity in the description of the analysis. If it has not already been included, I would consider it wise to include the baseline levels as a covariate in any analysis. The 'Statistical methods' goes on to describe the testing of covariates at baseline. This is something that should not be done as we know that the Null hypothesis of both groups coming from the same population is certain to be true because of the randomisation. You will see in section 15 of the CONSORT guidelines that this is something that is mentioned explicitly. The tabulation is recommended but the significance tests are not. "Statistical methods' should also describe the method the authors have used to allow for multiple testing of the secondary outcomes. Table 3 has a footnote that states "significant at a 0.0069 level to control for the false discovery rate at 5%". It is not obvious to me how this has been arrived at. Again note that tests of significance should not be applied to the baseline differences in Table 3. In the section on 'Harms', it is not clear which of the four subjects described is the individual who appears in Figure 1 as a withdrawal and is described as finding the treatment unpleasant. A heading such as 'Adverse Effects' or 'Adverse Events' would be more usual than 'Harms'. The legends for Figures 2 and 3 should explain the meaning of the error bars. In Appendix 4, I am unconvinced that the significance tests are justified. Those at baseline should not appear, as mentioned previously, while at subsequent times there has already been a more relevant assessment of the significance of changes from baseline. I like the idea of including this table but the tests of significance are not helpful. Many of the results are presented with excessive precision. As the sample sizes are under 100, a decimal place in a percentage contains no useful information and is determined by the denominator. Thus percentages should all be shown as integers. The outcome variables are all recorded as integers and so a second decimal place in means and SDs has no relevance. The outcome variables should be rounded to one decimal place. For

	reporting of p-values, I can appreciate why some need 4 decimal places but most do not. Two significant digits are the most that are required and I know that many of my colleagues would prefer to see only one significant digit. The Abstract reports the numbers randomised to each group and someone just reading the Abstract would assume that these are the numbers for the results presented subsequently. Please ensure that the reader is aware of the numbers on which the analysis is based. There is nothing in the results that makes me suspicious of any problems with the residuals but it is best practice to check the distribution of the residuals and, more importantly, to check for any influential points. The authors may well have made such checks but the methods used and the findings should be mentioned. The magnitude of effect of acupuncture in this study is impressive and the authors correctly discuss that the placebo effect may be important. I think it is also worth discussing whether that placebo effect might be greater in a situation where the participants are all volunteers, presumably with a prior expectation of acupuncture being beneficial. The design is unusual in that the controls know they will receive the acupuncture after a delay of six weeks. It is conceivable that this would lead to an expectation of no improvement or deterioration in the intervening six weeks which might influence their scores adversely. I note that the only variable to show any appreciable degree of improvement in the control group is hot flushes, where a regression to the mean effect could be expected because of the admission criteria for the study.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1
Reviewer Name: Prof. Masakazu Terauchi
Institution and Country: Department of Women's Health, Tokyo Medical and Dental University, Tokyo, Japan
Please state any competing interests or state ‘None declared’: I received an unrestricted research grant from Kikkoman Corporation.

Please leave your comments for the authors below
This is a report of the efficacy of a standardized brief “Western-style” acupuncture for women with moderate-to-severe menopausal hot flushes. Although the paper is of clinical relevance, the reviewer has a couple of concerns.

#1
The biggest flaw of the study is in its design. It is perplexing that the authors limited their analysis to the first six weeks of the study even though they seem to have a good chance to compare the two groups in a crossover fashion (Appendix 2). Considering the lack of an adequate placebo comparator as the authors describe, a crossover analysis would have been a perfect fit.

Response 1: The aim of our study was to investigate the impact of acupuncture versus no treatment which is here done in a randomized design where effect is to be determined before the crossover. After study week six the intervention group might still experience effect of the acupuncture treatment and is therefore not an optimal "control" neither for the control group nor for itself. The analysis is in accordance with our protocol article and stated in the trial register. We acknowledge the idea with the cross over design, and have planned this as an ad hoc article when the results from this primary effect article has been published. The purpose of offering delayed treatment to the control group was to enhance the adherence and later on, in another paper, to investigate potential sustained effect of the intervention.

#2
The questionnaire used in the study is not displayed in the paper. Appendix 1 is not the questionnaire, but the scoring system.

Response 2: You are right. We have now added the reference with the questionnaire used in the study.

Reviewer: 2
Reviewer Name: Boudewijn Kollen
Institution and Country: University Medical Center Groningen, Netherlands
Please state any competing interests or state 'None declared': none

Please leave your comments for the authors below
The manuscript, entitled "Efficacy of a standardized acupuncture approach for woman with bothersome menopausal symptoms: A randomized study in primary care (the ACOM study)" by Kamma Sundgaard Lund, Volkert Dirk Siersma, John Brodersen, Frans Boch Waldorff reports on an investigation on the efficacy of a brief acupuncture approach using scales of the MenoScores as outcomes in 70 women. I have only a few comments to make that basically serve to provide a better understanding of the choices you made in this study.
1. There is no information given in this manuscript whether patients used any self-chosen medication or were subjected to an alternative treatment to ameliorate the symptoms during the control period. Was the patient instructed to keep a diary for this information? This information will facilitate the interpretation of the results and its functional implications.

Response 1: Thank you for noticing this. The participants were *not* allowed to use any additional treatment (self-chosen medication or alternative treatment) for menopausal symptoms or medication that might alleviate menopausal symptoms beginning four weeks prior to enrolment until study week 11. This was stated under "participants/exclusion criteria" page 6 and mentioned in the discussion under "strengths and weaknesses" in the first section on page 16. However, we now see the need to be clearer and this is therefore also stated under "intervention" page 6.

2. Your study sample does not appear to be a random sample of patients with menopausal symptoms. Please, clarify in your paper the generalizability of your findings from this more or less selective population.

Response 2: The majority of women experience menopausal symptoms in the menopause, however, only 10-20% report these symptoms to be intolerable. Hence, not all women with menopausal symptoms need or request treatment, and we believe this acupuncture intervention is most relevant to

women who experience moderate-to-severe bothersome menopausal symptom. We wanted to reduce the influence of co-factors such as illness, medications or co-interventions, co-interventions due to cancer etc. which otherwise could affect the outcome. These co-factors are also relevant but requires additional studies and resources. Hence, we decided that the participants should reflect women attending their general practitioner (GP) requesting treatment for menopausal symptoms. We believe our sample to be representative of such women. The intervention was pragmatic and designed to be suitable for primary care practice, to be manageable by both participants and GPs but could easily be transferred to most clinical settings leading to high generalizability of the findings.

3. All treatments were given by GPs that were also certified acupuncturists. This may have introduced some bias as these “believers” are likely to favour the intervention treatment and as a result (subconsciously) put more effort and conviction in this treatment possibly intensifying any placebo effect.

Response 3: All participants were offered treatment, the control group just with a six week delay. Hence, the GPs did not favour one group (intervention) over the other. Furthermore, the data was based on patient-reported outcomes (self-reported by the participants independently of the GPs). However, a discussion of expectations and beliefs in relation to the placebo effect is now incorporated in the discussion on page 16.

4. Linear mixed models were used to analyze study results. It is not clear whether assumptions were met and which variables were used to control for the clustering of information. The limited sample size must have affected the distributions of your basically ordinal scaled data if I understand correctly. Shouldn’t you report medians instead of means in Appendix 4? You used continuous scaled information for your sample size calculation. Please, justify. Did you transform data in order to comply with the assumptions in the analyses?

Response 4: We adjust for the two variables that were used in the stratification of the randomisation as to optimize efficiency in our analyses. Furthermore, a woman random effect was used to account for the inherent clustering of multiple (longitudinal) outcome measurement time points for each woman.

We have used item response theory Rasch models to psychometrically validate our scales ensuring evidence for unidimensionality, additivity, invariance (or specific objectivity as Georg Rasch called it) and sufficiency (see our recently published paper about the development and validation of the MSQ, Lund et al 2018). If sufficiency is provided then full information of the measurement from the sum-score of the scale is given. Therefore, a score on one of scales of 4 is more than a score of 3, etc. This justifies the use of the scores of the scales as continuously valued outcome variables in our analyses.

A check of the assumptions of the linear regression analyses (see answers to reviewer 3), aided by the Central Limit Theorem, justifies the comparison of the untransformed mean difference in scores and the use of the asymptotic Wald t-tests. Since we compare means in our primary analyses, the descriptive table in Appendix 4 also reports means. Notably for scales with relatively few possible scorings, the median would be rather unsuited for descriptive analysis

5. Why did you decide not to analyze the longitudinal results all together? In my view that information is more informative than that of only change scores between baseline and 6 weeks. Moreover, the ICH/GCP guidelines stipulate that a correction for baseline is necessary when analyzing change scores as they state that … the use of change from baseline without adjusting for

baseline does not generally constitute an appropriate covariate adjustment …. Please, justify your choice for not adding the baseline variable of outcome as a covariate.

Response 5: For why we analysed the data only up to week 6 (when the crossover took place), please see our response 1 to reviewer 1. The longitudinal data for the three time points - baseline, 3 weeks and 6 weeks - was indeed analyzed jointly in a linear mixed model. The model was parameterized such that the parameter estimates at 3 weeks and 6 weeks are interpreted as the mean difference of the outcome between the randomization groups, beyond the mean difference already present at baseline; this is the same as the mean difference in the difference from baseline corrected for the baseline value of the outcome. See also our answers to reviewer 3.

6. You report statistically significant results but are these results also clinically relevant? Please, substantiate your claim.

Response 6: Yes, the significant results we also deem clinically relevant. The power calculation was based on data from the MSQ validation study considering relevant clinical effects (1). As mentioned above the intervention demonstrated a reduction in scale scores corresponding to a reduction from “a lot” to “quite a bit”. This means going from severe to moderate bothersome symptoms and this we view as clinically relevant.

7. I commend you for implementing a correction for multiplicity even though strictly speaking this correction is not called for when you only have one primary endpoint.

Response: Yes, thank you.

Reviewer: 3
Reviewer Name: Robin Prescott
Institution and Country: Emeritus Professor of Health Technology Assessment, Centre for Population Health Sciences, Usher Institute, University of Edinburgh, UK
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
I am reviewing this paper as a statistician. This is a well-designed study and much of presentation is good. I have no concerns about the validity of the conclusions from this study but I am not sure that the most efficient analysis has always taken place. My lack of certainty is because the descriptions of the analyses are sometimes lacking in sufficient detail to be sure of precisely what was done.

Firstly though, I would like to comment about the title of the paper. This is a pragmatic randomised controlled trial and it would be helpful to potential readers to be aware of that from the title.

Response 1: Thank you for reminding us about this. We have added this to the title.

Within the section on randomisation, there needs to be a little bit of detail about the stratification by age. What age categories were used in the stratification? From Table 2 I would guess that the categories are 40-55 and 56-65 but this should be stated explicitly. The randomisation itself looks sound.

Response 2: Thank you for noticing this. This is now stated explicitly in the randomization section page 9.

Apart from the very minor comment above, I found all of the sections up to 'Statistical methods' were clear. The primary outcome was clearly defined, as were the secondary outcomes. However, the description of the linear mixed models was insufficiently detailed. The description indicates that subjects were fitted as a random effect with the randomisation factors of age and level of symptoms appropriately included as covariates. It is perhaps implicit that age was fitted using the categories utilised in the randomisation, but perhaps that could be made explicit. There is, however, no indication that the baseline level of the outcome variable was included as a covariate, as would be necessary for a fully efficient analysis. It is not even clear whether a separate model was used for each of the two follow-up times or whether there was a unified analysis with terms being fitted for treatment, visit and visit by treatment interactions and possibly visit by baseline interactions to allow for a different baseline effect at each visit. The advantage of the unified model is that the data from the subject who was present for a three month assessment but not the six month assessment would still contribute to the estimate of treatment effects at six months, thereby gaining a little efficiency. I am not suggesting that the unified analysis is essential but what is essential is clarity in the description of the analysis. If it has not already been included, I would consider it wise to include the baseline levels as a covariate in any analysis.

Response 3: We actually did a unified analysis of the up to three data points available for each woman within the framework of a linear mixed model. This model then was parameterized so that the parameter estimates for week 3 and week 6 are to be interpreted as the mean difference in outcome between the randomization groups beyond the difference already present at baseline (the difference stated for week 0 is of course the estimate of the raw mean difference in outcome between randomization groups). This is similar to a regression on the differences from baseline including the baseline value of the outcome as a covariate (but slightly more efficient as the reviewer remarks).

At least, that was what we had planned (and was also written in a footnote below Table 3). However, it turned out that we – by mistake – did not report the mean differences corrected for the baseline difference in the table. We now report the correct estimates (that are not very different, since there were no differences at baseline to speak of).

We have rewritten the statistical analysis (page 9):

For each of the primary and secondary outcomes, the up to three assessments for each woman were modelled with a linear mixed model with a level for each time point for each randomisation group; the inherent correlation between observations on the same woman was accounted for by the inclusion of a subject-random effect. The effect of the intervention was estimated at week 3 and week 6 by the mean difference of the outcome beyond the difference already present at baseline and assessed by the appropriate Wald test in the model. The model additionally included as covariates the dichotomisations used in the stratification of the randomisation: age and level of symptoms. Four or more treatments were considered adequate adherence. The statistical significance was assessed controlling for the false discovery rate at 5% with the method of Benjamini and Hochberg (4). SAS v9.4 was used for the analyses.

The 'Statistical methods' goes on to describe the testing of covariates at baseline. This is something that should not be done as we know that the Null hypothesis of both groups coming from the same population is certain to be true because of the randomisation. You will see in section 15 of the CONSORT guidelines that this is something that is mentioned explicitly. The tabulation is recommended but the significance tests are not.

Response 4: Based on this comment we have removed the sentence regarding tests of covariates in the statistical methods section and removed the p-values in Table 2 and the p-values in week 0 in Table 3.

“Statistical methods’ should also describe the method the authors have used to allow for multiple testing of the secondary outcomes. Table 3 has a footnote that states “significant at a 0.0069 level to control for the false discovery rate at 5%”. It is not obvious to me how this has been arrived at. Again note that tests of significance should not be applied to the baseline differences in Table 3.

Response 5: We controlled for the False Discovery Rate at 5% with the method of Benjamini and Hochberg (4).

In the section on ‘Harms’, it is not clear which of the four subjects described is the individual who appears in Figure 1 as a withdrawal and is described as finding the treatment unpleasant. A heading such as ‘Adverse Effects’ or ‘Adverse Events’ would be more usual than ‘Harms’.

Response 6: We have changed the heading from “Harms” to “Harms and adverse events”. The participant mentioned in Figure 1, as withdrawal because she found the treatment unpleasant, was not mentioned under adverse events, as we did not interpreted this as an unexpected adverse event (adverse effect). However, acupuncture is in some cases experienced as a bit unpleasant (needles are inserted into the body) and an unpleasant experience could be interpreted as an adverse effect. We are sorry about this confusion and have now incorporated this consideration in the “Harms and adverse events” section, page 15.

The legends for Figures 2 and 3 should explain the meaning of the error bars.

Response 7: The error bars in Figures 2 and 3 denote the 95% confidence interval of the estimate of the outcome means for each randomization group for each time point. This will be stated in the figure legend.

In Appendix 4, I am unconvinced that the significance tests are justified. Those at baseline should not appear, as mentioned previously, while at subsequent times there has already been a more relevant assessment of the significance of changes from baseline. I like the idea of including this table but the tests of significance are not helpful.

Response 8: We have removed the p-values from the table in Appendix 4.

Many of the results are presented with excessive precision. As the sample sizes are under 100, a decimal place in a percentage contains no useful information and is determined by the denominator. Thus percentages should all be shown as integers. The outcome variables are all recorded as integers and so a second decimal place in means and SDs has no relevance. The outcome variables should be rounded to one decimal place. For reporting of p-values, I can appreciate why some need 4 decimal places but most do not. Two significant digits are the most that are required and I know that many of my colleagues would prefer to see only one significant digit.

Response 9: We have corrected this where seems relevant.

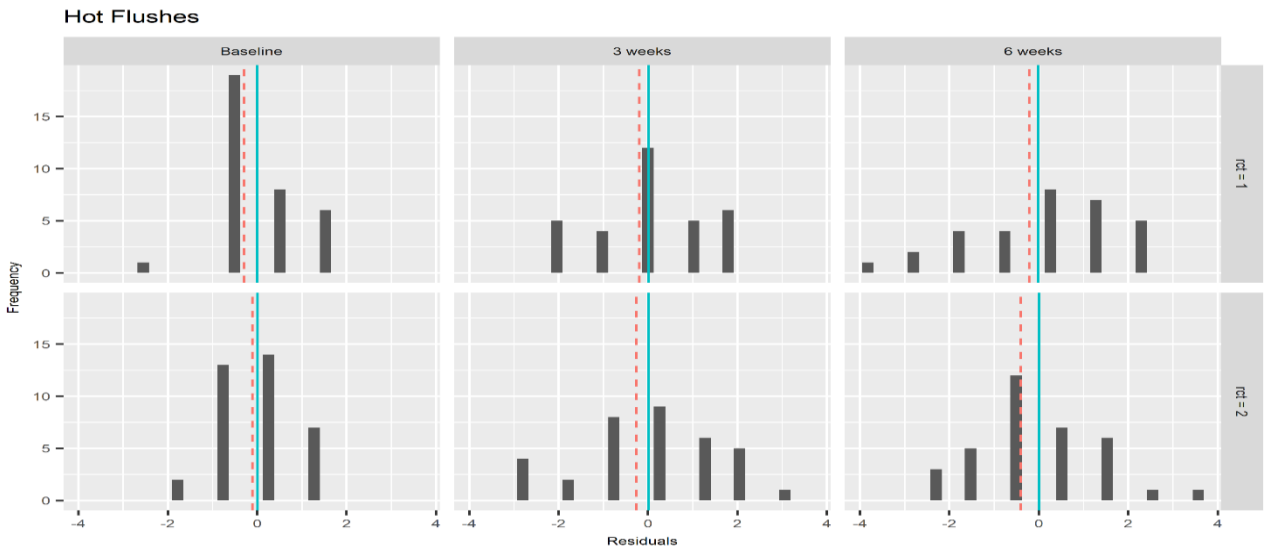
The Abstract reports the numbers randomised to each group and someone just reading the Abstract would assume that these are the numbers for the results presented subsequently. Please ensure that the reader is aware of the numbers on which the analysis is based.

Response 10: The analyses are intention-to-treat so that all randomized women are included in the analyses. Four participants dropped out before the 6-week follow-up, but they contribute to the data for the baseline and (for one of the four) the 3-week follow-up, and, because of the mixed model approach, also give some efficiency to the 6-week comparison (see reviewer 3, comment/response 3). Moreover, the mixed model approach constitutes a first line defence against bias due to differential attrition. The four participants who dropped out are fully accounted for in the flowchart in Figure 1 and are now mentioned in the abstract.

There is nothing in the results that makes me suspicious of any problems with the residuals but it is best practice to check the distribution of the residuals and, more importantly, to check for any influential points. The authors may well have made such checks but the methods used and the findings should be mentioned.

Response 11:

We have checked our (linear mixed model for longitudinal data) analyses for the assumptions on the residuals of these models. While some of the residuals deviate from the normal distribution, our sample size is not so small that we cannot invoke the Central Limit Theorem so as to claim that the parameter estimates of the linear mixed model are approximately normal distributed. Hence, the most important potential problem for the residuals is the homogeneity of variance. In the below histograms the residuals are contrasted between randomization groups for each of the time points for the primary outcome HF. Here we do not see clear differences in variance between the randomization groups, nor do we see markedly different shapes of the distributions. For the other outcomes (histograms not shown) there was also no sign of variance or distribution heterogeneity. Furthermore, the mean (dashed salmon-coloured line) and median (full turquois line) are seen very similar which can be taken as a first-line check for a normal distribution. This was the case for most of the other outcome scales as well.



We have checked our (linear mixed model for longitudinal data) analyses for influential participants using Cook's D (see below table); none of the participants proved very influential (D>1).

	Cook's D				
	Min	Q1	Median	Q3	Max
Hot Fluses	0.0001	0.0063	0.0127	0.0197	0.0549
Day and Night Sweats	0.0001	0.0080	0.0141	0.0187	0.0496
General Sweating	0.0000	0.0055	0.0087	0.0160	0.0993
Menopausal-specific Sleeping Problems	0.0012	0.0066	0.0120	0.0188	0.0543
Emotional Symptoms	0.0009	0.0048	0.0074	0.0172	0.1274
Memory Changes	0.0005	0.0049	0.0085	0.0206	0.0616
Physical Symptoms	0.0002	0.0059	0.0094	0.0161	0.1224
Urinary and Vaginal Symptoms	0.0004	0.0045	0.0102	0.0181	0.1130
Abdominal Symptoms	0.0003	0.0072	0.0108	0.0151	0.0876
Skin and Hair Symptoms	0.0010	0.0053	0.0089	0.0170	0.1011
Sexual Symptoms	0.0002	0.0066	0.0120	0.0247	0.0648
Tiredness	0.0002	0.0062	0.0110	0.0198	0.0693

Inspection of the residuals of the models did not reveal serious variance heterogeneity. Inspection of Cook's D did not reveal subjects that were particularly influential to the results." We have now mentioned this in the results section at the bottom of page 10.

The magnitude of effect of acupuncture in this study is impressive and the authors correctly discuss that the placebo effect may be important. I think it is also worth discussing whether that placebo effect might be greater in a situation where the participants are all volunteers, presumably with a prior expectation of acupuncture being beneficial. The design is unusual in that the controls know they will receive the acupuncture after a delay of six weeks. It is conceivable that this would lead to an expectation of no improvement or deterioration in the intervening six weeks which might influence their scores adversely. I note that the only variable to show any appreciable degree of improvement in the control group is hot flushes, where a regression to the mean effect could be expected because of the admission criteria for the study.

Response 12: This is a good point and we have incorporated this in our discussion of the limitations page 16.

1. Lund KS, Siersma VD, Christensen KB, Waldorff FB, Brodersen J. Measuring bothersome menopausal symptoms: development and validation of the MenoScores questionnaire. Health and quality of life outcomes. 2018;16(1):97.
2. Dodin S, Blanchet C, Marc I, Ernst E, Wu T, Vaillancourt C, et al. Acupuncture for menopausal hot flushes. The Cochrane database of systematic reviews. 2013(7):Cd007410.
3. Lund KS, Brodersen J, Siersma V, Waldorff FB. The efficacy of acupuncture on menopausal symptoms (ACOM study): protocol for a randomised study. Danish medical journal. 2017;64(3).
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995:289-300.

VERSION 2 – REVIEW

REVIEWER	Prof. Masakazu Terauchi Department of Women's Health, Tokyo Medical and Dental University, Tokyo, Japan
REVIEW RETURNED	10-Oct-2018

GENERAL COMMENTS	The reviewer's questions were appropriately answered.

REVIEWER	Boudewijn Kollen University Medical Center Groningen, University of Groningen, The Netherlands
REVIEW RETURNED	17-Oct-2018

GENERAL COMMENTS	I suggest you expand the statistical paragraph in your paper with information that was used to address the reviewers' comments. Much of that information is necessary to better understand the applied analyses. Furthermore, I remain unconvinced that this rather small (in my view more or less selected) population of women with intolerable menopausal symptoms is fully representative of women with similar symptoms elsewhere. Your assumption that it is needs to be clarified in the discussion section of your paper. Thank you.

REVIEWER	Robin Prescott University of Edinburgh, Scotland
REVIEW RETURNED	03-Oct-2018

GENERAL COMMENTS	I thank the authors for their detailed responses to my comments and for the corresponding changes to the paper. These changes address all of my original observations and I have no further comments. I congratulate the authors on a very nice paper.

VERSION 2 – AUTHOR RESPONSE

Thank you for the relevant and constructive comments. We have now corrected the manuscript according to the suggested minor revisions. We hope this is accepteble.