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 electroacupuncture for the treatment of constipation in Parkinson's disease：study protocol for a multi-centre randomised controlled trial
AUTHORS	Li, kunshan; Wang, Zhaoqin; Chen, Yiyi; Shen, Lirong; Li, Zhongqiu; Wu, Yiwen; Yuan, Canxing; Huang, Yan; Wu, Luyi; Bao, Chunhui; Zhang, wei; Xu, Shifen; Wu, Huangan

VERSION 1 - REVIEW

REVIEWER	Zhishun Liu Guang An Men Hospital, China Academy of Chinese Medicine
REVIEW RETURNED	03-Mar-2019

GENERAL COMMENTS	1.This is a interesting and clinical meaningful study topic;2.The protocol is ok generally though major revision is needed. Some questions: 1. The constipation resulted from PD , Why the diagnostic criteria of Roman IV diagnostic criteria for functional constipation will be used? 2.Why so many objectives and unclear hypothesis ?3. Whether the participants with constipation before PD will be included?4. Why selects SBMs as primary outcome , not CSBMs(completely spontanious bowel movements);5.Why only intervention group participants will be followed up? 6.Why motor function of PD will be used for outcomes?

REVIEWER	Guihua Tian Key Laboratory of Chinese Internal Medicine of Ministry of Education and Beijing,Dongzhimen Hospital, Beijing University of Chinese Medicine, Beijing, China
REVIEW RETURNED	25-Mar-2019

GENERAL COMMENTS	1.The exclusion criteria need to be further included the aspect of the past medical history and the allergic history. 2.The plan of allocation concealment needs to be described in detail. 3.Please provide the study proofs about“ the 12 week treatment period” . 4.The connection method of wire for the electric needle and the stimulating wave of EA should be further illustrated. 5.The outcome measures should contain the evaluation of quality of life and relevant laboratory indexes. 6.I suggest that the outcome evaluation time point of this protocol should be better illustrated as a table.

REVIEWER	Irina Chis Ster St George's University of London
REVIEW RETURNED	11-Apr-2019

GENERAL COMMENTS	The authors did not recognize the Poisson nature of the outcome, i.e. the number of bowel movements per week. This has implications on the sample size calculation as well as the analyses plan which are not adequately explained. The English language is very poor. Many statements are made using past tense or past perfect only to switch to the future which makes the manuscript very difficult to be read.

REVIEWER	Dr Andrew Hinde University of Southampton
REVIEW RETURNED	13-May-2019

GENERAL COMMENTS	Major points I cannot help feeling that it might be worth following the control group as well as the intervention group at weeks 16 and 24. By not following the control group up you are making quite a strong claim that there are no potentially confounding changes over time going on that might affect both groups (e.g. changes to the staff in the hospitals). The description of the statistical analysis you propose to do needs revising and amplifying. First, I cannot follow your sample size calculations on p. 17, ll. 14-22. It looks initially as if you are planning to do an independent samples t-test, but eventually I was not sure of this. There are three issues here. (1) If the mean number of spontaneous bowel movements (SBMs) in your control group is 1.23, and you expect the treatment group to experience a mean change of 2.20 (i.e. they should get an average of 3.43 SBMs), and the standard deviation of both groups is 1.93, then assuming the number of SBMs is normally distributed, you do not need anything like 52 cases in each group to detect such a difference at conventional power levels. But (2) the distribution of SBMs is not normal (with a mean of 1.23 and a standard deviation of 1.93 it cannot be, as it is presumably truncated at zero!). You could use non-parametric tests to take account of this, and you mention these later (p. 18, ll. 15-16). But if you used these, what calculations did you do in the Power Analysis and Sample Size software? I guess that for some of the tests you want to do you might need a much larger sample size of up to 52 in each group, but I cannot check this without more information about exactly what analyses you propose. I should like more details of all these calculations.

	(3) What does 'test efficiency' (l. 19) mean? I know what 'statistical efficiency' is but I think you mean something different. Do you mean 'test power'? Then on p. 17, ll. 17-19 it should not be '52 samples' and '63 samples' but 'two samples, each of size 52' and 'two samples, each of size 63'. Actually, if you have a 20 per cent drop out rate and want to end up with 52 in each group, you need 65 in each group to start with, rather than 63. So these lines should read: 'According to the Power Analysis and Sample Size software, we need 52 cases in each group. Assuming that we have a drop-out rate of 20 per cent, this means we need to recruit 65 individuals into each group'. If you really did only recruit 63 in each group, then you should amend the supposed drop out rate to 17.5 per cent. Second, on p. 18, ll. 14-21, you list a whole range of statistical tests. Are you planning to use all of them, and could you provide more details of where you will use each of them? You also mention a 'mixed effects' model. With 104 cases you are likely to struggle to estimate a random effect. Could you explain more about the model you want to estimate?Then on p. 18, ll. 16-17 what does '[t]he hypothesis test of the primary outcome needs to correct the central effect' mean? Minor points p. 8, l. 22 Lower case 's' in 'subjective'. p. 9, l. 5 'participants' not 'participates'. p. 11, l. 10 Insert 'suffering from' before 'schizophrenia'. p. 11, ll. 11-13 I could not understand the scoring system for educational levels. Do you mean that < 14 denotes illiteracy, 14-20 is primary educational level, and 21-24 is secondary school educational level and higher? p. 16, l. 7 'data collectors' plural. p. 18, ll. 3-5 I do not understand what these lines mean. p. 20, ll. 1-3 If the motor symptoms of Parkinson's Disease might reduce compliance, but those patients who do comply might experienced beneficial effects on their motor symptoms, this could lead to a bifurcation in the motor symptoms of your cases and controls. p. 20, l. 11 'negative results may be related to the small sample size'. No! The sample size will effect the reliability of the results, not their sign or size. p. 20, ll. 17-20 'The purpose of maintenance treatments is to observe whether reduction of treatment times after a certain amount of acupuncture treatment can still maintain the therapeutic effect'. I am not sure you will gain much new knowledge here. Is there much difference between simply stopping treatment immediately and running it down over 2-3 weeks like you are planning to do? It is possibly asking too much of this trial to contribute to knowledge of the impact of a reduction in treatment times. p. 20, ll. 28-30' Either 'Rome' or 'Roman' but not a mixture. Some abbreviations would benefit from definition. These are: 'UPDRS' (p. 8, l. 26; and Table 1, pp. 9-10) 'DSMB' (p. 17, l. 4) 'PASS' (p. 17, l. 20) 'NMS' (p. 19, l. 13)

REVIEWER	Peter Herbison University of Otago New Zealand
REVIEW RETURNED	16-May-2019

GENERAL COMMENTS	I found this article difficult to review in part as the English was confusing. The tenses were wrong, words misused and some sentences meaningless. From what I could understand there were several problems. 1. The only information about the generation of the randomisation sequence was that it was computer-generated. The randomisation was stratified by two factors. Was a separate random sequence generated for each strata? Was it blocked in any way? 2. The authors say they will use last observation carried forward to deal with missing values. This is known to produce biased results. Multiple imputation is the currently recommended method for dealing with missing data in the analysis. 3. A simple t-test (or equivalent non-parametric test) is proposed for examining the difference between the randomised groups. As the primary outcome (and some secondary ones) is a change from baseline this will be biased and less powerful that an ANCOVA. ANCOVA uses the final values and adjusts for the baseline values. It can also adjust for the stratification variables and any potential confounders. The t-test simply tests whether the groups are different but the ANCOVA will give an estimate (with a confidence interval) of how different they are. 4. The primary outcome is a count variable. The usual ANCOVA is on a continuous variable with some assumptions about normality. But an equivalent of an ANCOVA can be done with a count outcome using Poisson regression. Some thought should be given about using this for the analysis. 5. It was unclear who was going to get consent from participants and explain the trial to them 6. I cannot replicate the power calculation with the information given. 7. In the exclusion criteria there is mention of "primary school education level is <20 scores". I do not understand what this means. Is there any way of explaining this in a way that those not used to the Chinese education system would understand?

REVIEWER	Sally Kerry Queen Mary Univeristy of London
REVIEW RETURNED	28-May-2019

GENERAL COMMENTS	There are numerous errors in the written English; where it si clear what they mean I have ignored this (e.g. participates instead of participants) but I think in some places they have used past tense where they mean future tense which makes it very confusing to understand exactly what is being proposed. there is an overuse of abbreviations which is also confusing. The basic design is OK, randomising patients with Parkinsons disease to accupunture plus usual care or usual care alone.I think teh randomisation process is also OK alsthough teh language is unclear.

	The statistical anlaysis proposed is flawed but could be remedied. It is unclear whether there is a statistician on the team. Detailed review, points in the order they occur in the paper. The last sentence of the background , quoted below, would normally refer to the study being designed but is written in the past tense. Therefore I am a little confused over whether this study has already been done or it is a grammatical mistake. 'Therefore, this study observed the efficacy of EA in treating constipation in PD through a multi-center randomised controlled trial (RCT), in order to clarify the feasibility and advantages of acupuncture treatment on constipation in PD.' Randomisation will use 'hierarchical random method of variable region.' I am a senior statistician in a trials unit but have no idea what this means and there isn't a reference to look it up. The followup time appears to be different in the two groups. The intervention arm appears to have an additional 12 weeks. This section is not very clear but the groups should be treated the same with regard to data collection. Outcome should be collected at the same time point post randomisation in the intervention and control arms. If it is intended to observe the intervention arm for longer and not the control arm then there needs to be a good justification for this because much stronger evidence for the prolonged efficacy will come from a randomised comparison. Participants would normally be provided with an information sheet not the protocol. Exclusion criteria – not very clear mainly language rather than definitions. However exclusions for antihistamine and analgesia is rather vague and may exclude a number of patients because they take common medications for relatively minor complaints It is often difficult for statisticians to be completely blind at the analysis stage when there is compliance data. This data will be collected in the intervention arm only. It is usually possible to clean the outcome data blind to group by only looking at data collected from both groups and analysing the compliance data at a later stage. It is not clear where the intervention will be administered and whether this is acceptable to the patients. As the intervention is delivered 3 times a week this could be quite burdensome to attend hospital. Primary outcome is change from baseline at 8 weeks. Without using a regression model to adjust for baseline value this will be biased where there are differences between the groups at baseline due to regression to the mean. It is perfectly legitimate, statistically, to report change from baseline as the primary outcome but it must the analysed adjusting for baseline value. This will give the same answer as simply analysing the outcome adjusting for baseline value . It is unclear why 8 weeks has been chosen as the primary outcome point while the intervention lasts for 12 weeks, It also

	seems very odd to have a secondary analaysis using all the data points except 8 weeks. There is no clear rationale for this and it is not very clear what is being proposed and why. If each centre has only one data collector what happens if that person leaves or is sick?. I don't think it is necessary to have the same person and stating that you will means you may have to deviate from your protocol. Last observation carried forward can lead to bias and is no longer recommended. It can lead to bias and can underestimate the variability leading to confidence intervals than are narrower than they should be. Better methods are multiple imputation using the missing at random principle and also multiple imputation under reasonable assumptions about what might happen to the missing participants. This could be done as a sensitivity analysis. As the term ITT can be interpreted in a number of ways (see CONSORT) it is better to specify exactly what this means in this context. The sentence wasn't clear to me what it meant. The role of the DSMB described here appears to be one of guarantor of the study results. This is not normally the role of the DSMB, which is to protect the rights and safety of participants by monitoring the unblinded results (as presented to them by the trial team or and independent statistician) and SAEs as the trial progresses. They would not normally be able to verify the accuracy and authenticity of the results although if they are concerned they can ask to see the data. It is up to the DSMB to decide their remit and they may take on this role but it would not be usual. The sample size is correct for comparing 2.2 with 0.97 i.e. a change of 2.2 in the intervention arm and 0.97 in the control arm. 2.2-0.97 =1.23 which is the control mean. However this is not what is stated in the paragraph on sample size. The sample size should clearly state the primary outcome and difference to be detected. When it is a mean difference the actual value is not important for the calculation, Per protocol analysis will be biased as the inclusion criteria only applies to the intervention arm. Statistical analysis 'The hypothesis test of the primary outcome needs to correct the central effect.' I don't know what this means. The emphasis of the analysis should be on estimations and confidence intervals not hypothesis tests (see CONSORT). Therefore non-parametric tests and Chi-square test should not be used in isolation but in conjunction with an estimate of effect e.g. odds ratio, relative risk ect. It is good practice to adjust for stratification factors in the analysis. The statistical analysis is not very fully described and a good way of dealing with this is to write a full statistical analysis plan (SAP) prior to any analysis by group. This should include what data will be presented at baseline and make it clear this is merely descriptive. It would be a good idea to state in the protocol that a

	full SAP will be written and reviewed by an independent statistician e.g. the DSMB statistician. Data Sharing. Link does not work. There should be more explanation about what data will be shared, under what circumstances and with whom, bearing in mind the whole data set may contain fields that pose a non-negligible risk of being able to identify the participants . Flow chart is confusing and seems to imply the groups are treated differently in terms of data collection

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

1. The constipation resulted from PD, Why the diagnostic criteria of Roman IV diagnostic criteria for functional constipation will be used?

Reply: Because constipation in Parkinson's disease also belongs to the category of functional constipation, and so far we do not find a diagnosis standard about constipation in Parkinson's disease. The diagnostic criteria of constipation in Parkinson's disease mostly used Roman III diagnostic criteria for functional constipation. [Barichella M, Pacchetti C, Bolliri C, et al. Probiotics and prebiotic fiber for constipation associated with Parkinson disease: An RCT. Neurology 2016;87(12):1274; Parkinson Study Group. A randomized trial of relamorelin for constipation in Parkinson's disease (MOVE-PD): Trial results and lessons learned. Parkinsonism Relat Disord 2017;37:101-5.] In 2016, Roman III is updated to Roman IV. Therefore, our research chosen the Roman IV as diagnosis standard.

2. Why so many objectives and unclear hypothesis ?

Reply: Thank you for such a good suggestion. We combined them together, revised these in manuscript, and marked in red. Page 8.

3. Whether the participants with constipation before PD will be included?

Reply: Patients with constipation before PD are also included. First of all, constipation as a non-motor symptoms of PD may appear earlier than motor symptoms, about 50% of patients have constipation symptoms 10 to 20 years before motor symptoms occur in PD. [Chen Y, Yu M, Liu X, et al. The Clinical characteristics and peripheral T cell subsets in Parkinson 's diseases, patients with constipation. Int J Clin Exp Pathol 2015;8(3):2495-504.] In addition, according to our clinical observation, the constipation may be aggravated after the occurrence of PD motor symptoms. Therefore, the constipation that occurred before the diagnosis of PD could not be proved to be independent of PD.

4. Why selects SBMs as primary outcome, not CSBMs (completely spontanious bowel movements) ?

Reply: Currently, there is no uniform standard for evaluating the Parkinson's disease constipation. We found that SBMs was used as an outcome for constipation in PD.[A randomized trial of relamorelin for constipation in Parkinson's disease (MOVE-PD): Trial results and lessons learned. Parkinsonism Relat Disord 2017;37:101-05.] Meanwhile, SBMs was also a common outcome to evaluate functional

constipation. [Zheng H , Liu Z S , Zhang W , et al. Acupuncture for patients with chronic functional constipation: A randomized controlled trial. Neurogastroenterology & Motility 2018:e13307; Johanson JF, Morton D, Geenen J, et al. Multicenter, 4-week, double-blind, randomized, placebo-controlled trial of lubiprostone, a locally-acting type-2 chloride channel activator, in patients with chronic constipation. Am J Gastroenterol 2008;103(1):170-7; Barish CF, Drossman D, Johanson JF, et al. Efficacy and safety of lubiprostone in patients with chronic constipation. Dig Dis Sci 2010;55(4):1090-7.]Patients taking the drug for the treatment of PD motor symptoms are aggravating constipation, which makes them are different from the general functional constipation. Based on the above reasons, it can better evaluate the effect of acupuncture on constipation of Parkinson's disease.

5. Why only intervention group participants will be followed up?

Reply: The purpose of follow-up was to observe the long-term efficacy of acupuncture in constipation with PD. In addition, there were some reasons for not following up the control group. Firstly, main efficacy evaluation point is at week 8 for comparing the primary outcome between the two groups. Secondly, Follow-up in the control group would mean that the patient had to wait longer, about 6 months, and did not receive acupuncture treatment to relieve constipation.

6. Why motor function of PD will be used for outcomes?

Reply: Motor function of PD is only a secondary outcome in this study. It is used for outcomes because motor function and constipation of PD interplay. The aggravation of motor symptom leads to a decrease in patients' activities and an increase in drug dosage, both of them may aggravate constipation. On the other hand, constipation is a risk factor for the occurrence and development of PD. [Abbott RD, Petrovitch H, White LR, et al. Frequency of bowel movements and the future risk of Parkinson's disease. Neurology 2001;57(3):456-62; Gao X, Chen H, Schwarzschild MA, et al. A prospective study of bowel movement frequency and risk of Parkinson's disease. Am J Epidemiol 2011;174(5):546-51; Adams-Carr KL, Bestwick JP, Shribman S, et al. Constipation preceding Parkinson's disease: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry 2016;87(7):710-6.] The aggravation of constipation may lead to weakening gastrointestinal function and reducing the absorption of drugs, which lead to producing adverse effects on motor symptoms. If the patients' motor symptoms become worse, they may need to increase the dosage of drugs, and some drugs will aggravate the symptoms of constipation, which makes a vicious circle. To some extent, the treatment of motor symptoms also helps to improve the symptoms of constipation. Therefore, when treating constipation, we also conducted intervention on the motor symptoms and included the motor symptoms as a secondary outcome.

Reviewer: 2
1. The exclusion criteria need to be further included the aspect of the past medical history and the allergic history.

Reply: The first to fourth items of the exclusion criteria include the past medical history. And the diagnosis criteria for Idiopathic Parkinson's disease include 16 exclusion criteria used to exclude secondary Parkinson's disease and Parkinson's syndrome, such as a history of recurrent stroke and progressive deterioration of PD symptoms; a history of repeated brain injury; history of encephalitis; family history; brain tumors or traffic hydrocephalus was examined by CT; High-dose of levodopa was ineffective (eliminating absorption disorder); contact history of 1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP).

According to your advice, we added patients who were allergic to lactose solution or glycerine enema to the exclusion criteria. We revised and marked in red. Page 11.

2. The plan of allocation concealment needs to be described in detail.

Reply: Thanks for your advice. We describe the plan of allocation concealment in more detail. Page 11-12.

3. Please provide the study proofs about“ the 12 week treatment period” .

Reply: Our main treatment period is 8 weeks, and 9-12 weeks is maintenance therapy. The treatment period of high-quality RCT on acupuncture treatment for functional constipation is 8 weeks. [Liu Z, Yan S, Wu J, et al. Acupuncture for Chronic Severe Functional Constipation: A Randomized Trial. Ann Intern Med 2016;165(11):761-69.] At present, we have not found high-quality RCT on acupuncture for treating constipation in PD with. Therefore, we referred to other interventions to treat constipation in PD which treatment period was also 8 weeks. [Zangaglia R , Martignoni E , Glorioso M , et al. Macrogol for the treatment of constipation in Parkinson's disease. A randomized placebo-controlled study. Movement Disorders 2010, 22(9):1239-1244.] Based on the results of small sample preliminary pre-trial, we found that treatment for 8 weeks is effective, so we finally selected treatment for 8 weeks. Starting in week 9, we will reduce the frequency of acupuncture treatment until the 12th week. It is called maintain treatment, and is widely used in current researches related to acupuncture. [Hershman D L , Unger J M , Heather G , et al. Effect of Acupuncture vs Sham Acupuncture or Waitlist Control on Joint Pain Related to Aromatase Inhibitors Among Women With Early-Stage Breast Cancer. JAMA, 2018, 320(2):167-176; Horta D, Lira A, Sanchez-Lloansi M, et al. A Prospective Pilot Randomized Study: Electroacupuncture vs. Sham Procedure for the Treatment of Fatigue in Patients With Quiescent Inflammatory Bowel Disease. Inflamm Bowel Dis 2019.] On the one hand, it will be observed whether reducing the frequency of treatment can maintain the therapeutic effect of acupuncture treatment, and in this way the over-treatment will be prevented. On the other hand, we designed such research from the perspective of economics. If reducing the frequency of acupuncture can play a role in maintenance treatment, the economic burden of patients can be reduced in clinical application. In a word, we designed this study to provide guidance for clinical treatment of acupuncture.

4. The connection method of wire for the electric needle and the stimulating wave of EA should be further illustrated.

Reply: We have further explained the connection method of wire for the electric needle and the stimulating wave of EA in details and marked in red. Thanks a lot! Page 14.

5. The outcome measures should contain the evaluation of quality of life and relevant laboratory indexes.

Reply: In the preliminary study, we designed the PAC-QOL scale to assess the quality of life in patients, but we found that the vast majority of patients in this study have a longer course of disease, and short-term treatment is not obvious for improvement in quality of life. Meanwhile, this scale contains a lot of contents, which may increase the burden on patients to fill out. So in the end we didn't add it, but used the VAS score to evaluate the improvement of constipation.

The severity of PD is mainly evaluated by the patients' clinical manifestations. Currently, no specific laboratory indexes can be used to evaluate the severity of PD. Common laboratory indexes for evaluating constipation are colon transport test and defecography, and so on. Colon transport test and defecography are generally used as diagnostic criteria rather than clinical evaluation indicators in constipation. In this study, the purpose is to observe the clinical effect of acupuncture treatment on constipation patients with PD, so relevant laboratory indexes are not involved.

6. I suggest that the outcome evaluation time point of this protocol should be better illustrated as a table.

Reply: We have already listed the outcome evaluation time point in table 1 and Figure 2. Page 9; Page 29.

Reviewer: 3
1. The authors did not recognize the Poisson nature of the outcome, i.e. the number of bowel movements per week. This has implications on the sample size calculation as well as the analyses plan which are not adequately explained.

Reply: Thank you for such a good suggestion. We will use generalized linear models for statistical inference of primary outcome if necessary. When calculating the sample size, we still adopt the method of parameter statistics as Liu's study.[Liu Z, Yan S, Wu J, et al. Acupuncture for Chronic Severe Functional Constipation: A Randomized Trial. Ann Intern Med 2016;165(11):761-69; Liu Z, Xu H, Chen Y, et al. The efficacy and safety of electroacupuncture for women with pure stress urinary incontinence: study protocol for a multicenter randomized controlled trial. TRIALS 2013;14:315.]

2. The English language is very poor. Many statements are made using past tense or past perfect only to switch to the future which makes the manuscript very difficult to be read.
Reply: We are very sorry that the manuscript is difficult to read. The manuscript has been performed by professional editors at Editage, a division of Cactus Communications, to ensure that the language is clear.

Reviewer: 4
Major points
1. I cannot help feeling that it might be worth following the control group as well as the intervention group at weeks 16 and 24. By not following the control group up you are making quite a strong claim that there are no potentially confounding changes over time going on that might affect both groups (e.g. changes to the staff in the hospitals).

Reply: The reasons why we do not set up follow-up in control group are as follows: The purpose of follow-up was to observe the long-term efficacy of acupuncture in constipation with PD, which it can be proved by comparing the difference of outcomes between treatment and follow-ups period in the intervention group. Therefore, it will not happen that potentially confounding changes over time going on that might affect both groups. In addition, there were some reasons for not following up the control group. First, main efficacy evaluation point is at week 8 for comparing the primary outcome between the two groups. Secondly, Follow-up in the control group would mean that the patient had to wait longer, about 6 months, and did not receive acupuncture treatment to relieve constipation.

2. If the mean number of spontaneous bowel movements (SBMs) in your control group is 1.23, and you expect the treatment group to experience a mean change of 2.20 (i.e. they should get an average of 3.43 SBMs), and the standard deviation of both groups is 1.93, then assuming the number of SBMs

is normally distributed, you do not need anything like 52 cases in each group to detect such a difference at conventional power levels.

Reply: Thank you for finding our mistake. The change of SBMs in the control group is 0.97, which has been modified in page 17. Meanwhile, we will make corrections accordingly in Ethical approval and Clinical trial registration.

3. But the distribution of SBMs is not normal (with a mean of 1.23 and a standard deviation of 1.93 it cannot be, as it is presumably truncated at zero!). You could use non-parametric tests to take account of this, and you mention these later (p. 18, ll. 15-16). But if you used these, what calculations did you do in the Power Analysis and Sample Size software? I guess that for some of the tests you want to do you might need a much larger sample size of up to 52 in each group, but I cannot check this without more information about exactly what analyses you propose. I should like more details of all these calculations.

Reply: We assume that the mean of the waitlist control group is 0.97 (formerly 1.23 is a writing error) with a standard deviation of 1.93. As the disease progresses, patients' SBMs in the waitlist control group may decreased because there is no other treatment except emergency treatment for constipation in the wait time. The average SBMs of 8-9 weeks minus the average of -2-0 weeks may have a negative number. Therefore, it is presumably not truncated at zero.

The specific calculations are as follows:

PASS 15.0.5 2019/6/15 13:17:46 1

Two-Sample T-Tests Assuming Equal Variance

Numeric Results for Two-Sample T-Test Assuming Equal Variance

Alternative Hypothesis: H1: $\delta = \mu1 - \mu2 \neq 0$

Target Power	Actual Power	N1	N2	N	μ1	μ2	δ	σ	Alpha
0.90	0.90153	53	53	106	2.2	1.0	1.2	1.9	0.050

References

Julious, S. A. 2010. Sample Sizes for Clinical Trials. Chapman & Hall/CRC. Boca Raton, FL.
Chow, S. C., Shao, J., and Wang, H. 2008. Sample Size Calculations in Clinical Research (Second Edition).
Chapman & Hall/CRC. Boca Raton, FL.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd
Edition. Blackwell Science. Malden, MA.
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions

Target Power is the desired power value (or values) entered in the procedure. Power is the probability of
rejecting a false null hypothesis.
Actual Power is the power obtained in this scenario. Because N1 and N2 are discrete, this value is often
(slightly) larger than the target power.
N1 and N2 are the number of items sampled from each population.
N is the total sample size, N1 + N2.
μ1 and μ2 are the assumed population means.
δ = μ1 - μ2 is the difference between population means at which power and sample size calculations are made.
σ is the assumed population standard deviation for each of the two groups.
Alpha is the probability of rejecting a true null hypothesis.

Summary Statements

Group sample sizes of 53 and 53 achieve 90.153% power to reject the null hypothesis of equal
means when the population mean difference is μ1 - μ2 = 2.2 - 1.0 = 1.2 with a standard deviation for both groups of 1.9 and with a significance level (alpha) of 0.050 using a two-sided two-sample equal-variance t-test.

Dropout-Inflated Sample Size

Dropouts — Dropout Rate	Sample Size — N1 / D2	N2 / D	N	Dropout-Inflated Enrollment Sample Size — N1'	N2'	N'	Expected Number of — D1
20%	53 / 14	53 / 28	106	67	67	134	14

Definitions

Dropout Rate (DR) is the percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e. will be treated as "missing").

N1, N2, and N are the evaluable sample sizes at which power is computed. If N1 and N2 subjects are evaluated out of the N1' and N2' subjects that are enrolled in the study, the design will achieve the stated power.

N1', N2', and N' are the number of subjects that should be enrolled in the study in order to end up with N1, N2, and N evaluable subjects, based on the assumed dropout rate. After solving for N1 and N2, N1' and N2' are calculated by inflating N1 and N2 using the formulas N1' = N1 / (1 - DR) and N2' = N2 / (1 - DR), with N1' and N2' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., and Wang, H. (2008) pages 39-40.)

D1, D2, and D are the expected number of dropouts. D1 = N1' - N1, D2 = N2' - N2, and D = D1 + D2.

Two-Sample T-Tests Assuming Equal Variance

Procedure Input Settings

Autosaved Template File

C:\Users\king\Documents\PASS 15\Procedure Templates\Autosave\Two-Sample T-Tests Assuming Equal Variance - Autosaved 2019_6_15-13_17_46.t388

Design Tab

Solve For:	Sample Size
Alternative Hypothesis:	Two-Sided
Power:	0.90
Alpha:	0.05
Group Allocation:	Equal (N1 = N2)
Input Type:	Means
μ1:	2.20
μ2:	0.97
σ:	1.93

4. What does 'test efficiency' (l. 19) mean? I know what 'statistical efficiency' is but I think you mean something different. Do you mean 'test power'?
Reply: Thank you for finding this mistake. We are referring to test power, which is a writing error and has been modified in page 17.

5. Then on p. 17, ll. 17-19 it should not be '52 samples' and '63 samples' but 'two samples, each of size 52' and 'two samples, each of size 63'. Actually, if you have a 20 percent drop out rate and want to end up with 52 in each group, you need 65 in each group to start with, rather than 63. So these lines should read: 'According to the Power Analysis and Sample Size software, we need 52 cases in each group. Assuming that we have a drop-out rate of 20 per cent, this means we need to recruit 65 individuals into each group'. If you really did only recruit 63 in each group, then you should amend the supposed drop out rate to 17.5 per cent.
Reply: Thank you very much! There is indeed an error in calculating 20 percent drop out rate. Calculated by power analysis and sample size (PASS) software, 53 samples will be needed for a single group. Considering the 20% drop-off rate, 67 samples will be needed for a single group, and a

total of 134 subjects will be recruited. We revised and marked in red in Page 17-18. Meanwhile, we will make corrections accordingly in Ethical approval and Clinical trial registration.

6. Second, on p. 18, ll. 14-21, you list a whole range of statistical tests. Are you planning to use all of them, and could you provide more details of where you will use each of them? You also mention a 'mixed effects' model. With 104 cases you are likely to struggle to estimate a random effect. Could you explain more about the model you want to estimate?Then on p. 18, ll. 16-17 what does '[t]he hypothesis test of the primary outcome needs to correct the central effect' mean?

Reply: Thank you for such a good suggestion. We have invited an independent statistician from the department of biostatistics of Fudan University to revise and review a full statistical analysis plan. We don't use 'mixed effects' model for analysis any more. The details are shown in statistical analysis section of the manuscript in page 18-19.

Minor points
1. p. 8, l. 22 Lower case 's' in 'subjective'.

Reply: This content was deleted for the recommendations of other reviewers.

2. p. 9, l. 5 'participants' not 'participates'.

Reply: Thanks for your advice. We have made corresponding modifications in page 8.

3. p. 11, l. 10 Insert 'suffering from' before 'schizophrenia'.

Reply: Thanks for your advice. We have made corresponding modifications in page 11.

4. p. 11, ll. 11-13 I could not understand the scoring system for educational levels. Do you mean that < 14 denotes illiteracy, 14-20 is primary educational level, and 21-24 is secondary school educational level and higher?

Reply: We regret that the sentences is difficult to understand. Actually we mean that if the patient is illiteracy, the score of Mini-mental State Examination (MMSE) < 14 scores indicates cognitive impairment; if the patient only receives primary education, the score of MMSE < 20 scores indicates cognitive impairment; if the patient's education levels higher than primary education, MMSE < 24 scores indicate cognitive impairment. So we have re-detailed the relevant content in the exclusion criteria to expresse more clearly and marked in red. Page 11.

5. p. 16, l. 7 'data collectors' plural.

Reply: Thanks for your advice. We have made corresponding modifications in page 16.

6. p. 18, ll. 3-5 I do not understand what these lines mean.

Reply: We are very sorry that the sentences is difficult to understand. The number of acupuncture treatments has a great influence on the treatment effect. Therefore, in order to better explain the

results of the protocol, we require the number of electroacupuncture treatments to reach at least 80% of the total electroacupuncture sessions. However, since the waitlist control group is similar to the blank control group, the second criterion is only for the requirements of the intervention group. We have re-detailed the relevant content in the criteria to expressed more clearly and marked in red. Page 18.

7. p. 20, ll. 1-3 If the motor symptoms of Parkinson's Disease might reduce compliance, but those patients who do comply might experienced beneficial effects on their motor symptoms, this could lead to a bifurcation in the motor symptoms of your cases and controls.

Reply: The reason we are intervention in motor symptoms is that constipation with PD patients is different from simple functional constipation. Motor symptoms and constipation of PD interaction. The aggravation of motor symptom leads to a decrease in patients' activities and an increase in drug dosage, both of which may aggravate constipation. On the other hand, constipation is a risk factor for the occurrence and development of PD. [Abbott RD, Petrovitch H, White LR, et al. Frequency of bowel movements and the future risk of Parkinson's disease. Neurology 2001;57(3):456-62; Gao X, Chen H, Schwarzschild MA, et al. A prospective study of bowel movement frequency and risk of Parkinson's disease. Am J Epidemiol 2011;174(5):546-51; Adams-Carr KL, Bestwick JP, Shribman S, et al. Constipation preceding Parkinson's disease: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry 2016;87(7):710-6.] The aggravation of constipation may lead to weakening gastrointestinal function and reducing drug absorption, which lead to producing adverse effects on motor symptoms. The aggravation of constipation may lead to weakening gastrointestinal function and reducing the absorption of drugs, which lead to producing adverse effects on motor symptoms. If the patients' motor symptoms become worse, they may need to increase the dosage of drugs, and some drugs will aggravate the symptoms of constipation, which makes a vicious circle. To some extent, the treatment of motor symptoms also helps to improve the symptoms of constipation.

8. p. 20, l. 11 'negative results may be related to the small sample size'. No! The sample size will effect the reliability of the results, not their sign or size.

Reply: Thanks to your advice, the relevant content has been deleted.

9. p. 20, ll. 17-20 'The purpose of maintenance treatments is to observe whether reduction of treatment times after a certain amount of acupuncture treatment can still maintain the therapeutic effect'. I am not sure you will gain much new knowledge here. Is there much difference between simply stopping treatment immediately and running it down over 2-3 weeks like you are planning to do? It is possibly asking too much of this trial to contribute to knowledge of the impact of a reduction in treatment times.

Reply: Maintain treatment is widely used in current researches related to acupuncture. [Hershman D L , Unger J M , Heather G , et al. Effect of Acupuncture vs Sham Acupuncture or Waitlist Control on Joint Pain Related to Aromatase Inhibitors Among Women With Early-Stage Breast Cancer[J]. JAMA, 2018, 320(2):167-176; Horta D, Lira A, Sanchez-Lloansi M, et al. A Prospective Pilot Randomized Study: Electroacupuncture vs. Sham Procedure for the Treatment of Fatigue in Patients With Quiescent Inflammatory Bowel Disease. Inflamm Bowel Dis 2019.] On the one hand, it will be observed whether reducing the frequency of treatment can maintain the therapeutic effect of acupuncture treatment, and in this way the over-treatment will be prevented. On the other hand, we designed such research from the perspective of economics. If reducing the frequency of acupuncture can play a role in maintenance treatment, the economic burden of patients can be reduced in clinical

application. In a word, we designed this study to provide guidance for clinical treatment of acupuncture.

10. p. 20, ll. 28-30' Either 'Rome' or 'Roman' but not a mixture.

Reply: Thank you for such a good suggestion. In order to better describe the discussion section, we have modified the content of the discussion section and removed this content.

11. Some abbreviations would benefit from definition. These are: 'UPDRS' (p. 8, l. 26; and Table 1, pp. 9-10) 'DSMB' (p. 17, l. 4) 'PASS' (p. 17, l. 20) 'NMS' (p. 19, l. 13)

Reply: Thank you for such a good suggestion. We have added the full name in the way you described and marked in red. Page 15, 9, 17, 17, 20.

Reviewer: 5
1. The only information about the generation of the randomisation sequence was that it was computer-generated. The randomisation was stratified by two factors. Was a separate random sequence generated for each strata? Was it blocked in any way?

Reply: Each strata is generated from a separate random sequence. To prevent the researcher from guessing the grouping situation of the next patient, we conducted the variable block random method, (block size is 4 or 6). We adjusted accordingly, the details were shown in sample size section of the manuscript. Page 11.

2. The authors say they will use last observation carried forward to deal with missing values. This is known to produce biased results. Multiple imputation is the currently recommended method for dealing with missing data in the analysis.

Reply: Thanks a lot for your advice. We changed it into “Missing data will assumed to be under the missing-at-random assumption and will be imputed using multiple imputation” in page 17.

3. A simple t-test (or equivalent non-parametric test) is proposed for examining the difference between the randomised groups. As the primary outcome (and some secondary ones) is a change from baseline this will be biased and less powerful that an ANCOVA. ANCOVA uses the final values and adjusts for the baseline values. It can also adjust for the stratification variables and any potential confounders. The t-test simply tests whether the groups are different but the ANCOVA will give an estimate (with a confidence interval) of how different they are.

Reply: We have invited an independent statistician from the department of biostatistics of Fudan University to revise and review a full statistical analysis plan. We will assess primary outcome by a random effect generalized linear model with possible covariates such as group, baseline, study site, visit, rescue medicine, other defecation aids, etc.

4. The primary outcome is a count variable. The usual ANCOVA is on a continuous variable with some assumptions about normality. But an equivalent of an ANCOVA can be done with a count outcome using Poisson regression. Some thought should be given about using this for the analysis.

Reply: Thank you for such a good suggestion. Our primary outcome is the change from baseline in mean spontaneous bowel movements (SBMs) per week in the weeks 8-9. That means the total number of SBMs over two weeks needs to be divided by 2 to get the average number of defecations per week. Thus, it is a continuous variable. We have revised our statistical analysis plan, and we will assess primary outcome by a random effect generalized linear model with possible covariates such as group, baseline, study site, visit, rescue medicine, other defecation aids, etc.

5. It was unclear who was going to get consent from participants and explain the trial to them

Reply: We set up an independent researcher to get the participants' consent and explain the trial to them. The additional sentences are added in manuscript and marked in red. Page 10.

6. I cannot replicate the power calculation with the information given.

Reply: Thanks to your correction, we do have some errors in the calculation of sample size. We adjusted accordingly, the details were shown in sample size section of the manuscript. Page 17-18. Meanwhile, we will make corrections accordingly in Ethical approval and Clinical trial registration.

7. In the exclusion criteria there is mention of ‘primary school education level is <20 scores’. I do not understand what this means. Is there any way of explaining this in a way that those not used to the Chinese education system would understand?

Reply: We are very sorry that the sentence is difficult to understand. We have re-detailed the relevant content in the exclusion criteria and marked in red. Page 11.

Reviewer: 6
1. There are numerous errors in the written English; where it si clear what they mean I have ignored this (e.g. participates instead of participants) but I think in some places they have used past tense where they mean future tense which makes it very confusing to understand exactly what is being proposed. there is an overuse of abbreviations which is also confusing.

Reply: We reduced the use of abbreviations, and wrote full names before abbreviations in the first sentence that appears again within 4 paragraphs.

2. The statistical anlaysis proposed is flawed but could be remedied. It is unclear whether there is a statistician on the team.
Reply: We invited an independent statistician from the department of biostatistics of Fudan University to revise and review a full statistical analysis plan. The details are shown in statistical analysis section of the manuscript in page 18-19.

3. The last sentence of the background , quoted below, would normally refer to the study being designed but is written in the past tense. Therefore I am a little confused over whether this study has already been done or it is a grammatical mistake. ‘Therefore, this study observed the efficacy of EA in

treating constipation in PD through a multi-center randomised controlled trial (RCT), in order to clarify the feasibility and advantages of acupuncture treatment on constipation in PD.'

Reply: Thanks for your advice. It is a grammatical error, and 'observed' has been changed into 'will investigate' and marked in red. Page 8.

4. Randomisation will use 'hierarchical random method of variable region.' I am a senior statistician in a trials unit but have no idea what this means and there isn't a reference to look it up.

Reply: Thank you for finding our mistake. We would like to describe the stratified random method, and variable block random method is designed for better allocation concealment, that is to say for preventing researchers who performs a random assignment from guessing the next grouping of patients. We have adjusted accordingly, the details are shown in randomization and allocation concealment section of the manuscript. Page 11-12.

5. The followup time appears to be different in the two groups. The intervention arm appears to have an additional 12 weeks. This section is not very clear but the groups should be treated the same with regard to data collection. Outcome should be collected at the same time point post randomisation in the intervention and control arms. If it is intended to observe the intervention arm for longer and not the control arm then there needs to be a good justification for this because much stronger evidence for the prolonged efficacy will come from a randomised comparison.

Reply: The purpose of follow-up was to observe the long-term efficacy of acupuncture in constipation with PD, which can be achieved by comparing the difference of outcomes between treatment and follow-ups period in the intervention group. In addition, there were some reasons for not following up the control group. First, main efficacy evaluation point is at week 8 for comparing the primary outcome between the two groups. Secondly, Follow-up in the control group would mean that the patient had to wait longer, about 6 months, and did not receive acupuncture treatment to relieve constipation.

6. Participants would normally be provided with an information sheet not the protocol.

Reply: Thank you for such a good suggestion. We have adjusted accordingly and marked in red. Page 10.

7. Exclusion criteria – not very clear mainly language rather than definitions. However exclusions for antihistamine and analgesia is rather vague and may exclude a number of patients because they take common medications for relatively minor complaints
Reply: Thank you for such a good suggestion. We have delete the exclusion criteria "patients taking antidepressants, analgesics, or antihistamines that affect PD symptoms within 2 weeks before treatments" and adjusted accordingly in Page 10-11. Meanwhile, we will make corrections accordingly in Ethical approval and Clinical trial registration.

8. It is often difficult for statisticians to be completely blind at the analysis stage when there is compliance data. This data will be collected in the intervention arm only. It is usually possible to clean the outcome data blind to group by only looking at data collected from both groups and analysing the compliance data at a later stage.

Reply: Thank you for such a good suggestion. Statisticians don't need to complete statistical analysis in a blind state. An independent group of data managers will be responsible for data collection and cleaning. Then locked dataset will be transferred to the statisticians and help them to carry out statistical procedures in accordance with pre-prepared statistical analysis plan. We adjusted accordingly in page 6, 8, 12.

9. It is not clear where the intervention will be administered and whether this is acceptable to the patients. As the intervention is delivered 3 times a week this could be quite burdensome to attend hospital.

Reply: We are going to conduct the study in the outpatient departments of four hospitals. The therapeutic method and period in this protocol are adjusted based on the preliminary trial, which have been proved to be acceptable to patients. Since the painful caused by this disease has brought a deep impact on patients, we find patients have a strong desire to receive treatment and can cooperate with treatment. In addition, in order to reduce the burden of patients coming to the hospital for treatment, we have set up four centers in different areas of Shanghai, so that patients can come to the hospital for treatment more convenient.

10. Primary outcome is change from baseline at 8 weeks. Without using a regression model to adjust for baseline value this will be biased where there are differences between the groups at baseline due to regression to the mean. It is perfectly legitimate, statistically, to report change from baseline as the primary outcome but it must the analysed adjusting for baseline value. This will give the same answer as simply analysing the outcome adjusting for baseline value .

Reply: Thanks a lot for your advice. We will assess primary outcome by a random effect generalized linear model with baseline as a possible covariate.

11. It is unclear why 8 weeks has been chosen as the primary outcome point while the intervention lasts for 12 weeks, It also seems very odd to have a secondary analaysis using all the data points except 8 weeks. There is no clear rationale for this and it is not very clear what is being proposed and why.

Reply: Our main treatment period is 8 weeks, and 9-12 weeks is maintenance therapy. The treatment period of high-quality RCT on acupuncture treatment for functional constipation is 8 weeks. [Liu Z, Yan S, Wu J, et al. Acupuncture for Chronic Severe Functional Constipation: A Randomized Trial. Ann Intern Med 2016;165(11):761-69.] At present, we have not found high-quality RCT on acupuncture for treating constipation in PD with. Therefore, we referred to other interventions to treat constipation in PD which treatment period was also 8 weeks. [Zangaglia R , Martignoni E , Glorioso M , et al. Macrogol for the treatment of constipation in Parkinson's disease. A randomized placebo-controlled study. Movement Disorders 2010, 22(9):1239-1244.] Based on the results of small sample preliminary pre-trial, we found that treatment for 8 weeks is effective, so we finally selected treatment for 8 weeks. Starting in week 9, we will reduce the frequency of acupuncture treatment until the 12th week. It is called maintain treatment, and is widely used in current researches related to acupuncture. [Hershman D L , Unger J M , Heather G , et al. Effect of Acupuncture vs Sham Acupuncture or Waitlist Control on Joint Pain Related to Aromatase Inhibitors Among Women With Early-Stage Breast Cancer. JAMA, 2018, 320(2):167-176; Horta D, Lira A, Sanchez-Lloansi M, et al. A Prospective Pilot Randomized Study: Electroacupuncture vs. Sham Procedure for the Treatment of Fatigue in Patients With Quiescent Inflammatory Bowel Disease. Inflamm Bowel Dis 2019.] On the

one hand, it will be observed whether reducing the frequency of treatment can maintain the therapeutic effect of acupuncture treatment, and in this way the over-treatment will be prevented. On the other hand, we designed such research from the perspective of economics. If reducing the frequency of acupuncture can play a role in maintenance treatment, the economic burden of patients can be reduced in clinical application. In a word, we designed this study to provide guidance for clinical treatment of acupuncture.

We originally want to describe the change from baseline in mean spontaneous bowel movements (SBMs) per week in the weeks 8-9 is the primary outcome, the change from baseline in mean SBMs per week in the other time points is not the primary outcome. But it may be ambiguous to say 'Secondary outcome: the change from baseline in mean SBMs (excluding the weeks 8-9)', so the related sentence has been deleted and the corresponding descriptions of SBMs have been added to the primary outcomes in Page 15.

12. If each centre has only one data collector what happens if that person leaves or is sick?. I don't think it is necessary to have the same person and stating that you will means you may have to deviate from your protocol.

Reply: Because the evaluation of UPDRS in outcomes needs to be trained, and even after the training, the scores of it by different evaluators may be different. So patient's evaluation is best collected by the same data collector. When selecting the personnel, we also consider that it is better not to change the personnel during the study. Therefore, the personnel who may not leave before the completion of the study are selected. In order to prevent from the above situation happening, we trained two researchers at the same time. One researcher is the main one and another researcher is the substitute. In the event of a special situation, the substitute researcher will evaluate. During the pre-test, it also happened that on the day of the evaluation the evaluator could not make the evaluation due to special circumstances. We allowed the data collector to make evaluation within ±3 days.

13. Last observation carried forward can lead to bias and is no longer recommended. It can lead to bias and can underestimate the variability leading to confidence intervals than are narrower than they should be. Better methods are multiple imputation using the missing at random principle and also multiple imputation under reasonable assumptions about what might happen to the missing participants. This could be done as a sensitivity analysis.

Reply: Thanks a lot for your advice. Missing data will assumed to be under the missing-at-random assumption and will be imputed using multiple imputation.

14. As the term ITT can be interpreted in a number of ways (see CONSORT) it is better to specify exactly what this means in this context. The sentence wasn't clear to me what it meant.

Reply: In the sentence of "Full analysis set(FAS or Modified ITT dataset)" we mean intention-to-treat dataset. Another sentence that mentions the ITT, is removed according to the Reviewer's suggestion.

15. The role of the DSMB described here appears to be one of guarantor of the study results. This is not normally the role of the DSMB, which is to protect the rights and safety of participants by monitoring the unblinded results (as presented to them by the trial team or and independent statistician) and SAEs as the trial progresses. They would not normally be able to verify the accuracy and authenticity of the results although if they are concerned they can ask to see the data. It is up to the DSMB to decide their remit and they may take on this role but it would not be usual.

Reply: Thank you for your correction. Our understanding of the responsibilities of the DSMB is inadequate, and incorrect content has been corrected. “the accuracy and authenticity of the results” has been delete.

16. The sample size is correct for comparing 2.2 with 0.97 i.e. a change of 2.2 in the intervention arm and 0.97 in the control arm. 2.2-0.97 =1.23 which is the control mean. However this is not what is stated in the paragraph on sample size. The sample size should clearly state the primary outcome and difference to be detected. When it is a mean difference the actual value is not important for the calculation.
Reply: Thanks to your correction, we do have some errors in description and writing. We have adjusted accordingly, the details are shown in sample size section of the manuscript. Page 17-18. Meanwhile, we will make corrections accordingly in Ethical approval and Clinical trial registration.

17. Per protocol analysis will be biased as the inclusion criteria only applies to the intervention arm.

Reply: We define the Per-protocol set like this because that the number of acupuncture treatments has a great influence on the treatment effect. Therefore, in order to better explain the results of the protocol, we require the number of electroacupuncture treatments to reach at least 80% of the total electroacupuncture sessions. However, the waitlist control group is similar to the blank control group, it is only for the requirements of the intervention group.

18. ‘The hypothesis test of the primary outcome needs to correct the central effect.’ I don’t know what this means.

Reply: We have invited an independent statistician from the department of biostatistics of Fudan University to revise and review a full statistical analysis plan.

Generalized linear model is used as the method of statistical analysis. It may be different that baseline characteristics of the patients and the level of treatment in different , so we consider the central effect(i.e. study site) by generalized linear model.

19. The emphasis of the analysis should be on estimations and confidence intervals not hypothesis tests (see CONSORT). Therefore non-parametric tests and Chi-square test should not be used in isolation but in conjunction with an estimate of effect e.g. odds ratio, relative risk ect.

Reply: Thank you for such a good suggestion. Confidence interval of primary outcome and secondary outcomes will also be provided in our analysis except P values according to CONSORT. It’s a routine way of presenting the results in most statistical analysis, so we don’t particularly emphasize it in our protocol. Moreover, when considering the unbalances of baseline information and center effect, generalized linear equation will be applied to obtain the significance and estimate the 95% CI of the effect.

20. The statistical analysis is not very fully described and a good way of dealing with this is to write a full statistical analysis plan (SAP) prior to any analysis by group. This should include what data will be presented at baseline and make it clear this is merely descriptive. It would be a good idea to state in the protocol that a full SAP will be written and reviewed by an independent statistician e.g. the DSMB statistician.

Reply: Thank you for such a good suggestion. We have invited an independent statistician from the department of biostatistics of Fudan University to revise and review a full statistical analysis plan. The details are shown in statistical analysis section of the manuscript in page 18-19.

21. Data Sharing. Link does not work. There should be more explanation about what data will be shared, under what circumstances and with whom, bearing in mind the whole data set may contain fields that pose a non-negligible risk of being able to identify the participants .

Reply: We are deeply sorry that the connection does not work. The website of the management platform has been slightly changed and we have updated it in page 23. We plan to share the results of the study within six months of publication of the research results. We will input information of CRF into the platform, but the platform only discloses the patient number, gender, time and place of inclusion and the information of the inputting personnel. The basic information of patients is not open to the public.

22. Flow chart is confusing and seems to imply the groups are treated differently in terms of data collection
Reply: There is a difference in the data collection between the two groups. Both groups conduct data collection at week 4, week 8 and week 12. On this basis, the intervention group also conduct data collection at the follow-up period, while the control group did not conduct efficacy evaluation at the follow-up period. For the specific reasons, please refer to the reply of question 5.

VERSION 2 – REVIEW

REVIEWER	Zhishun Liu Guang An Men Hospital, China Academy of Chinese Medical Science China
REVIEW RETURNED	05-Jul-2019

GENERAL COMMENTS	The revisied manuscript is much better than before. All my questions have been addressed and explained properly.

REVIEWER	Guihua Tian Dongzhimen Hospital, Beijing University of Chinese Medicine, China.
REVIEW RETURNED	04-Jul-2019

GENERAL COMMENTS	This protocol is established to evaluate the efficacy and safety of electroacupuncture (EA) for treating constipation in PD. The method and statistical analysis were well-conducted to investigate if EA can play a role in maintenance therapy. However, some questions need to be answered. 1. The title of this protocol used ‘acupuncture’ while the major intervention in the manuscript was ‘Electroacupuncture’. I suggest that the title should be changed. 2. The sample size of this protocol needs to be recalculated. There is a big difference between the study[25] and this protocol.

	3. Whether there was evidence to support the selection of acupoints for this protocol. 4. Whether there was evidence to support the selection of time point for this protocol. 5. Was the treatment in the controlled group generally effective and had the clinical support? How to avoid the treatment in the controlled group becomes a confounding factor to this protocol?

REVIEWER	Irina Chisster Dr
REVIEW RETURNED	17-Jul-2019

GENERAL COMMENTS	I still hold the opinion that the manuscript is poorly written in general and lacks statistical clarity in particular. I suggest the authors to include a professional statistician on board before re-submission. The assumptions on which the sample size calculations are not mentioned (normality of the outcome perhaps?). I still hold the view that the main outcome is of Poisson nature and hence the sample size needs calculated appropriately - the minimum necessary numbers would be less than they calculated. There is nothing wrong if they can recruit more than necessary of course, but the statistical planning needs accuracy. I would be less concerned about a detailed statistical analysis plan. But as it stands in this revised version of the manuscript, it does not match the sample size calculation based on the main outcome which is the change in SBMs from baseline at weeks 8-9. The study is not adequately powered for mixed effects and the generalized linear model is mentioned without its context. I cannot comment on the clinical importance of such experiment in this population but I simply cannot see how the authors would conduct it without a statistician involved in this phase of planning and design. I also attach an example of sensitivity analysis to the sample size executed with the PASS software under Poisson distributional assumption of the outcome. This is just an example with some hypothetical values closed to those of the authors' - they should be responsible for their final input settings. Tests for the Difference Between Two Poisson Rates Numeric Results for Testing the Difference Between Two Poisson Rates Alternative Hypothesis: Two-Sided Group 1: Control Test Statistic: Large-Sample Group 2: Treatment Power N1 N2 N Grp 1 Event Rate λ1 Grp 2 Event Rate λ2 Diff λ2-λ1 Ratio λ2/λ1 Alpha 0.91256 22 22 44 0.50 1.50 1.00 3.0000 0.050 0.91518 16 16 32 0.50 1.75 1.25 3.5000 0.050 0.90764 12 12 24 0.50 2.00 1.50 4.0000 0.050 0.90423 8 8 16 0.50 2.50 2.00 5.0000 0.050 0.90546 6 6 12 0.50 3.00 2.50 6.0000 0.050 0.90423 32 32 64 1.00 2.00 1.00 2.0000 0.050 0.90187 22 22 44 1.00 2.25 1.25 2.2500 0.050 0.91083 17 17 34 1.00 2.50 1.50 2.5000 0.050 0.91256 11 11 22 1.00 3.00 2.00 3.0000 0.050 0.91518 8 8 16 1.00 3.50 2.50 3.5000 0.050

0.90248	53	53	106	2.00	3.00	1.00	1.5000	0.050
0.90546	36	36	72	2.00	3.25	1.25	1.6250	0.050
0.90344	26	26	52	2.00	3.50	1.50	1.7500	0.050
0.90423	16	16	32	2.00	4.00	2.00	2.0000	0.050
0.90187	11	11	22	2.00	4.50	2.50	2.2500	0.050

References

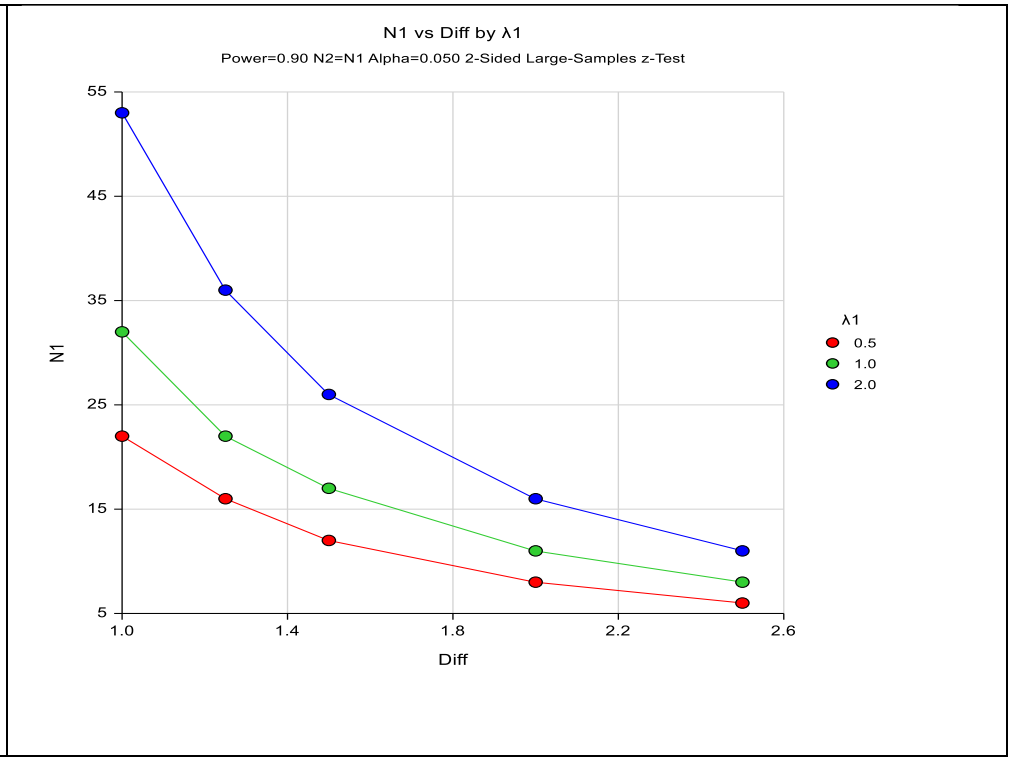
Mathews, Paul. 2010. Sample Size Calculations: Practical Methods for Engineers and Scientists. Mathews Malnar
and Bailey. Fairport Harbor, OH. www.mmbstatical.com
Smith, P.G. and Morrow, R.H. 1996. Field Trials of Health Intervention in Developing Countries: A Toolbox.
Macmillan Education. Oxford, England.
Campbell, M.J. and Walters, S.J. 2014. How to Design, Analyse and Report Cluster Randomised Trials in Medicine
and Health Related Research. John Wiley. New York.

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
N1 and N2 are the number of subjects in groups 1 and 2, respectively.
N is the total sample size. N = N1 + N2.
λ1 is the mean event (or incidence) rate of group 1, the control, reference, or baseline group.
λ2 is the mean event (or incidence) rate of group 2, the treatment group.
λ2-λ1 is the difference between the two event rates.
λ2/λ1 is the ratio of the two event rates.
Alpha is the probability of rejecting the null hypothesis when it is true. It should be small.

Summary Statements

A total sample size of 44 subjects with 22 in group 1 and 22 in group 2, each measured for a
fixed duration, achieves 91.256% power to detect a difference of 1.00 between the group 2
(treatment) event rate of 1.50 and the group 1 (control) event rate of 0.50 using a two-sided,
large-samples z-test of the Poisson event-rate difference at a significance level of 0.050.



REVIEWER	Andrew Hinde University of Southampton
REVIEW RETURNED	22-Jul-2019

GENERAL COMMENTS	Because of pressure of other work and unexpected family commitments I have only been able to give this revision a quick look. I am afraid that this is all I have time to do now and for the foreseeable future. Based on my quick look, I can say that you have addressed most of the points I raised in my previous report but not all. The main issue outstanding is that I remain to be convinced about the justification for observing the intervention group for 12 weeks more than the control group. You would get a comparison of the long-run impact of acupuncture by observing the control group as well. If you only observe the intervention group you will still be able to say something about how the outcomes change in the 12 weeks after the intervention is finished, but you will not be able to refute the suggestion that some general environmental changes have caused this. The statistical analysis you are planning to do is somewhat clearer. My understanding now is that your main outcome is the change in the mean spontaneous bowel movements (SBMs) between a period immediately prior to the intervention and weeks 8-9 of the the intervention. This can be analysed using a two independent samples t-test. I am glad that you are no longer planning to use mixed effects models, as you will not gain much insight from them with a sample as small as the one you have.

REVIEWER	Peter Herbison University of Otago New Zealand
REVIEW RETURNED	15-Jul-2019

GENERAL COMMENTS	This is now much better. The standard of English could be better, but it is not misleading. In the abstract they say one of the outcome is a VAS but they do not say what the VAS measures.

REVIEWER	Sally Kerry Queen Mary University of London UK
REVIEW RETURNED	14-Jul-2019

GENERAL COMMENTS	General comments The English is much improved but still needs attention. Similarly the statistical methods are improved but there doesn't appear to be a statistician on the team or a clinical trails unit. As with most pragmatic trials there are a number of pitfalls for the unwary and I would strongle recommend a methodologist on the team Abstract 1. 'Efficacy will be assessed in both groups'- efficacy is generally taken to mean the effect of a drug/treatment. Therefore, it doesn't make sense to talk of efficacy in both groups. Efficacy is best measured by comparing the 2 groups- I would not call the follow-up evaluation measure of efficacy – you could call it 'maintenance'. 2. 'In order to optimize the authenticity and generalizability of the research results, a multi-centre trial will be utilised for this study, which will include four Tertiary A Hospitals located in Shanghai, China' I don't know what the authors mean by 'authenticity' and it doesn't seem very generalizable using tertiary centres only. Many PD patients with constipation would not be treated in tertiary centres and would not want to attend hospital twice a week. 3. It should be 'stratified randomisation' not 'stratified random sampling' 4. Table 1 is very complicated and not very informative. It could be simplified. It isn't clear why AEs and divided by intervention and control but other measures are not. 5. Secondary outcomes include 'straining score from baseline per week based on the stool diary'. If measured at baseline then it is not an outcome 6. Adverse events- there needs to be a differentiation between adverse events that are directly attributable to the intervention which may need to be addressed in deciding whether to continue treatment and those of a more general nature. Collecting adverse events data to investigate whether the intervention leads to more adverse events should be done in both groups in the same way. In interventions where the intervention group has more health

	professional contact more adverse events maybe reported but only due to the greater exposure to the health professional. E.g. feeling unwell may lead to a session being cancelled in the intervention arm and be reported. A similar episode may be overlooked in the control arm. 7. Sample size statement is unclear. The authors seem to have assumed that the control participants will improve by 0.97 and so the study will detect an improvement of 2.0 in the intervention arm. It might be simpler to say that 'a previous study has shown an increase in 0.97 in the control and 2.20 in the intervention arms. In this study in order to detect a difference of 1.23 between intervention and control groups 52 participants in the anlaysis are required in each arm' It is the difference that is important not the two values. It could be 0 and 1.23 or 1 and 2.23. 8. Missing data – need to give more information on imputation process. 9. Full analysis set 'All patients will be randomised to receive at least one treatment and to have a therapeutic evaluation after the treatments.' It is not clear what is meant here. Some patients are randomised to control hence not to any treatment. I think you mean that all patients will be included regardless of the number of treatments they receive and multiple imputation will be used to impute missing data values. 10. Per protocol analysis without adjustment for confounding is potemtially biased. See N Engl J Med 2017; 377:1391-1398 DOI: 10.1056/NEJMsm1605385. It is unclear why multiple imputation has not been used here. 11. Safety assessment – this not clearly described as a measure applicable to both groups and it isn't clear what is being proposed here. 12. There is an over reliance on significance test in the analysis. Significance tests at baseline are not generally recommended as the null hypothesis should be true unless there is something wrong with the randomisation (see CONSORT). For analysis of efficacy an estimate of effect with confidence interval should be the main focus. 13. Baseline value should be included in the model to avoid regression to the mean and potential for biased results. Other covariates shodul also be chosen in advance and the protocol should state when this will be done. Using baseline data to decide covariates for the analysis may introduce bias. 14. It is a good idea to write a full statistical analysis plan and to have this reviewed by the statistician on your DMEC before unblinding of the data 15. DMEC – when will the interim analysis be carried out and for what purpose? 16. Data Sharing statement is unclear what is meant by 'basic information' – does this include age and gender? You may need to

	assess risk of re-identification of subjects with respect to demographic data and create categories of age to reduce re-identification. I clicked on the link given and it appears to be in Chinese- will the data be available to a wider international audience. It might be prudent to give a less open access or to put in some qualifier –e.g. subject to data protection legislation. In the UK the law governing data and data sharing has changed over the years and it will be several years before the data would become available 17. Flow chart What does heirachical random mean? The flow chart does not follow the normal CONSORT format and is very confusing.

VERSION 2 – AUTHOR RESPONSE

The responses to the reviewers' comments are as follows:

Reviewer: 1

The revisied manuscript is much better than before. All my questions have been addressed and explained properly.

Reply: Thanks for your encouraging comment.

Reviewer: 2

1. The title of this protocol used 'acupuncture' while the major intervention in the manuscript was 'Electroacupuncture'. I suggest that the title should be changed.

Reply: Thanks for your advice. We changed "acupuncture" to "electroacupuncture" in the title and marked it in red.

2. The sample size of this protocol needs to be recalculated. There is a big difference between the study[25] and this protocol.

Reply: At present, we have not found high-quality RCT on acupuncture for treating constipation in PD. However, the article[25] about acupuncture treatment for constipation we used to calculate the sample size of this protocol is high-quality RCT. In this study the main purpose was to observe the efficacy of acupuncture in the treatment of constipation in patients with Parkinson's disease. The primary outcomes for constipation are spontaneous bowel movements, effective rate, complete bowel movements, and stool frequency [The American Journal of Chinese Medicine 2013;41(4):717–742; NPJ Parkinsons Dis 2018;16(4):6.]. We selected the references based on our primary outcome (spontaneous bowel movements). Thus, we finally chose this study as reference. This reference is already quite in line with our research objectives.

3. Whether there was evidence to support the selection of acupoints for this protocol.

Reply: The acupoints for constipation are Tianshu (ST25), Fujie (SP14), Quchi (LI11), Shangjuxu (ST37), Sanyinjiao (SP6) and Zhaohai (KI6) in this protocol [Ann Intern Med 2016;165(11):761-69; Neurogastroenterol Motil 2018;30(7):e13307; Acupuncture and moxibustion. Beijing:science press 2017:319 (Book in Chinese); Acupuncture and moxibustion. Beijing:People's Medical Publishing House 2017:285(Book in Chinese).].

On the other hand, the rest acupoints are for PD motor symptoms. The positions of scalp acupoints Connect Qianding (GV21) to Xuanlu (GB5), Connect Qianshencong (EX-HN1) to Xuanli (GB6) are similar to chorea-tremor control area, and motor area of JIAOs scalp acupuncture points, respectively. The scalp acupoints and Quchi (LI11), Hegu (LI4), Yanglingquan (GB34), Sanyinjiao (SP6), Taichong (LR3) are the main acupoints commonly used in the treatment of PD [Shanghai Journal of Acupuncture and Moxibustion 2015;34(1):70-72.(Article in Chinese)].

All acupoints we used for constipation and PD motor symptoms were based on repeated discussion by experts in acupuncture, such as Xu bin, Xu shifen, Lao Lixing, Wu Huangan, Zhang Ren and Li Jing.

4. Whether there was evidence to support the selection of time point for this protocol.

Reply: Our main treatment period is 8 weeks. For treatment of constipation in PD patients, we have not found high-quality RCT for acupuncture. But there was a high-quality RCT of acupuncture for functional constipation with a treatment period of 8 weeks [Ann Intern Med 2016;165(11):761-69.], and isosmotic macrogol electrolyte solution was also used to treat constipation of PD for 8 weeks [Movement Disorders 2010; 22(9):1239-1244.]. Based on the results of a small sample preliminary pre-trial acupuncture treatment of constipation in PD patients, acupuncture treatment for 8 weeks can effectively relieve constipation, so we finally selected treatment for 8 weeks.

The treatment period 9-12 weeks is maintenance therapy. From week 9 to 12 week, we will reduce the frequency of acupuncture treatment, which is widely used in current researches on acupuncture [JAMA 2018, 320(2):167-176; Inflamm Bowel Dis 2019. doi: 10.1093/ibd/izz091].

The follow-up points are week 16 and week 24, which are to observe the short-term and long-term effects of acupuncture. During the follow-up period, the 4th week and 12th week after treatment is generally selected in the article about acupuncture [JAMA Intern Med 2019. doi: 10.1001/jamainternmed.2019.2407; Ann Intern Med 2016;165(11):761-69.]. In our protocol, week 16 and week 24 are equivalent to 4th week and 12th week after treatment.

We also evaluated the efficacy at week 4 during treatment, to observe the changes of acupuncture treatment over time.

5. Was the treatment in the controlled group generally effective and had the clinical support? How to avoid the treatment in the controlled group becomes a confounding factor to this protocol?

Reply: In the waitlist control group, emergency medication for constipation and basic drugs for PD motor symptom are used. If the patients have no bowel movements for three or more consecutive days, the patients will take lactulose solution orally or glycerine enema based on the urge to defecate or not. Both medicines are classic and are commonly used in many clinical studies. [J Gastroenterol 2019;54(6):530-40; Journal of complementary & integrative medicine 2015;12(4):325-31; Cochrane Database Syst Rev 2017;1:Cd011128.]. In our pilot trial, we found in the waitlist control group these drugs could alleviate the symptoms of constipation, and without these drugs the patients would feel uncomfortable. After discussion with gastroenterologists, the treatment of this trial was determined

after full consideration of the efficacy, safety and adverse effects of the drug. Prior to the enrollment of this trial, the treatment of motor symptoms in PD patients was determined by a neurologist at the Tertiary A Hospitals.

Based on the waitlist control group, the intervention group increased with electroacupuncture (EA). Our hypothesis is that EA + UC will increase the frequency of defecation compared with usual care (UC) alone. The waitlist control group has been established to prevent basic treatment from becoming a confounding factor [JAMA Intern Med 2019. doi:10.1001/jamainternmed.2019. 2407; JAMA Intern Med. 2019. doi: 10.1001/ jamainternmed. 2019.2407.]. The severity of constipation affects the doses of emergency medication for constipation. In order to avoid the different doses of constipation medication between the two groups as a confounding factor, we used a stratified random sampling method to control the baseline.

Reviewer: 3

1. I still hold the opinion that the manuscript is poorly written in general and lacks statistical clarity in particular. I suggest the authors to include a professional statistician on board before re-submission.

Reply: Thanks for your suggestions. To improve the language, the manuscript has been polished again by professionals from Editage. We have also invited Zhang Wei,

an independent statistician from the Department of Biostatistics of Fudan University, to participate in our team. The statistical plan and design were changed, revised statistics are marked in red. Page 15, 16,17,18.

2. The assumptions on which the sample size calculations are not mentioned (normality of the outcome perhaps?). I still hold the view that the main outcome is of Poisson nature and hence the sample size needs calculated appropriately - the minimum necessary numbers would be less than they calculated. There is nothing wrong if they can recruit more than necessary of course, but the statistical planning needs accuracy. I would be less concerned about a detailed statistical analysis plan. But as it stands in this revised version of the manuscript, it does not match the sample size calculation based on the main outcome which is the change in SBMs from baseline at weeks 8-9. The study is not adequately powered for mixed effects and the generalized linear model is mentioned without its context.

Reply: Thanks for your suggestions. We used the results of the pilot trial to estimate the sample size (Please see online supplementary file 2-pilot trial in the attachment for details). We understand you focus on Poisson nature of the primary outcome, so we specially validated its normality in the pilot trial using Shapiro-Wilk test. We did not use the pilot trial results in the first place due to insufficient cases completed at that time. The “sample size” section is changed and marked in red. Page 16.

The primary outcome should be the change in mean weekly spontaneous bowel movements (SBMs) from baseline to weeks 8-9. The number of SBMs during weeks 8-9 and weeks -2-0 will be divided by two respectively to get the mean weekly SBMs. Then, the mean weekly SBMs at weeks -2-0 will be subtracted from the mean weekly SBMs at weeks 8-9. Thus, it is a continuous variable. In statistical analysis plan, we plan to assess the primary outcome by a generalized linear model, including group, baseline SBMs and study site as fixed effects and age, visit number, rescue medicine and other defecation aids as possible covariates. At the same time, we also referred to the statistical strategies in a similar article [Ann Intern Med 2016;165(11):761-69.].

3. I cannot comment on the clinical importance of such experiment in this population but I simply cannot see how the authors would conduct it without a statistician involved in this phase of planning and design.

I also attach an example of sensitivity analysis to the sample size executed with the PASS software under Poisson distributional assumption of the outcome. This is just an example with some hypothetical values closed to those of the authors' - they should be responsible for their final input settings.

Reply: Thanks for your suggestions. According to your advice we have invited Zhang Wei, an independent statistician from the Department of Biostatistics of Fudan University, to participate in our team. He is responsible for the statistical plan and design. We have also updated the input settings for the estimation of sample size according to the results of the pilot trial(Please see online supplementary file 2-pilot trial in the attachment for details). The revised sections are marked in red in Page 16.

Reviewer: 4

1. Based on my quick look, I can say that you have addressed most of the points I raised in my previous report but not all. The main issue outstanding is that I remain to be convinced about the justification for observing the intervention group for 12 weeks more than the control group. You would get a comparison of the long-run impact of acupuncture by observing the control group as well. If you only observe the intervention group you will still be able to say something about how the outcomes change in the 12 weeks after the intervention is finished, but you will not be able to refute the suggestion that some general environmental changes have caused this.

Reply: Thanks for your comment. Because this is a major adjustment to the research protocol, our team decided to adopt your proposal after discussion. The ethical and clinical registration will also be altered accordingly. The related contents have been adjusted and marked in red.

2. The statistical analysis you are planning to do is somewhat clearer. My understanding now is that your main outcome is the change in the mean spontaneous bowel movements (SBMs) between a period immediately prior to the intervention and weeks 8-9 of the the intervention. This can be analysed using a two independent samples t-test. I am glad that you are no longer planning to use mixed effects models, as you will not gain much insight from them with a sample as small as the one you have.

Reply: Two independent samples t-test is our basic statistical analysis method. But if necessary, we will also use generalized linear model to correct other covariates. In addition, we also referred to the statistical strategies in a similar article. [Ann Intern Med 2016;165(11):761-69.]

Reviewer: 5

1. This is now much better. The standard of English could be better, but it is not misleading. In the abstract they say one of the outcome is a VAS but they do not say what the VAS measures.

Reply: Thanks for your suggestion. We have made corresponding modifications and marked it in red. Page 5.

Reviewer: 6

General comments

The English is much improved but still needs attention. Similarly the statistical methods are improved but there doesn't appear to be a statistician on the team or a clinical trails unit. As with most pragmatic trials there are a number of pitfalls for the unwary and I would strongly recommend a methodologist on the team.

Reply: Thanks for your suggestions. In order to improve the language, the manuscript has been polished again by professionals from Editage. We have also invited Zhang Wei, an independent statistician from the Department of Biostatistics of Fudan University, to participate in our team. The revised statistics section is marked in red. Page 15, 16,17,18.

Shifen Xu, one of our corresponding authors, contributed to the design of this study. She is an methodologist with rich experience in designing clinical trials.[BMJ OPEN 2019;9(4):e021484; TRIALS 2019;20(1):308; TRIALS 2019;20(1):117; J Pain Res 2018;11:1489-96; TRIALS 2018;19(1):52; Sleep Med 2017;37:193-200.] She reviewed the protocol carefully.

1. 'Efficacy will be assessed in both groups'- efficacy is generally taken to mean the effect of a drug/treatment. Therefore, it doesn't make sense to talk of efficacy in both groups. Efficacy is best measured by comparing the 2 groups- I would not call the follow-up evaluation measure of efficacy – you could call it 'maintenance'.

Reply: Thank you for such a good suggestion. We have changed "Efficacy" to "Outcomes" which is marked in red. Page 5.

2. 'In order to optimize the authenticity and generalizability of the research results, a multi-centre trial will be utilised for this study, which will include four Tertiary A Hospitals located in Shanghai, China' I don't know what the authors mean by 'authenticity' and it doesn't seem very generalizable using tertiary centres only. Many PD patients with constipation would not be treated in tertiary centres and would not want to attend hospital twice a week.

Reply: Thanks for your advice. The description was not so accurate and has been revised as: " In order to optimise the credibility of the research results, a multi-centre trial will be utilised for this study. " (Page 6) Actually, we meant to emphasize multi-centre rather than Tertiary A Hospitals. We will conduct the trial at multiple centers to demonstrate that the treatment is reproducible.

Indeed, many constipation patients may not be treated in Tertiary A Hospitals, but most PD patients are treated in these hospitals. This research focused on PD patients with constipation. In our pilot trial, We found more patients were recruited in Tertiary A Hospitals. If the treatment is effective, it can be promoted at all levels of hospitals.

Studies show that acupuncture has a time-effect feature. Patients who don't receive enough dosage of acupuncture won't be able to experience the best benefits of the treatment, because interval between two acupuncture treatments has a great influence on the effect. [Guo Y, Fang J. Experimental acupuncture science. Beijing:China Press of Traditional Chinese Medicine 2012.239-256. (Book in Chinese)].

3. It should be 'stratified randomisation' not 'stratified random sampling'

Reply: Thank you for finding this mistake. We have adjusted accordingly and marked it in red. Page 8, 11.

4. Table 1 is very complicated and not very informative. It could be simplified. It isn't clear why AEs and divided by intervention and control but other measures are not.

Reply: Thank you for such a good suggestion. We have adjusted it according to your suggestion and marked it in red. Page 9.

5. Secondary outcomes include 'straining score from baseline per week based on the stool diary'. If measured at baseline then it is not an outcome

Reply: This sentence was to express "the change from baseline in mean weekly straining score based on the stool diary". We might be improper in the expression, and have made corresponding modifications which are marked in red. Page 13-14.

6. Adverse events- there needs to be a differentiation between adverse events that are directly attributable to the intervention which may need to be addressed in deciding whether to continue treatment and those of a more general nature. Collecting adverse events data to investigate whether the intervention leads to more adverse events should be done in both groups in the same way. In interventions where the intervention group has more health professional contact more adverse events maybe reported but only due to the greater exposure to the health professional. E.g. feeling unwell may lead to a session being cancelled in the intervention arm and be reported. A similar episode may be overlooked in the control arm.

Reply: Thanks for your suggestions. We define the general AEs or the other AEs in "Safety Assessment" section. To prevent the waitlist control group from evaluating incorrect AEs due to fewer visits to the hospital, we asked the two groups of patients to call acupuncturists to report any discomfort during the trial. We made adjustments in Table 1 and the "Safety Assessment" section, marked in red. Page 9, 14.

7. Sample size statement is unclear. The authors seem to have assumed that the control participants will improve by 0.97 and so the study will detect an improvement of 2.0 in the intervention arm.

It might be simpler to say that 'a previous study has shown an increase in 0.97 in the control and 2.20 in the intervention arms. In this study in order to detect a difference of 1.23 between intervention and control groups 52 participants in the anlaysis are required in each arm'

It is the difference that is important not the two values. It could be 0 and 1.23 or 1 and 2.23.

Reply: Thanks for your suggestions. We understand your objection to the estimation of the sample size, so we use the results of the pilot trial to estimate the sample size. We validate its normality in pilot trial using Shapiro-Wilk test. The pilot trial results were not used in the first place was due to insufficient cases completed at that time. The details are in the "sample size" section (Page 16) and "online supplementary file 2-pilot trial" in the attachment.

8. Missing data – need to give more information on imputation process.

Reply: Thanks for your suggestion. We have adjusted it accordingly in the “Missing data processing” section. Page 15.

9. Full analysis set ‘All patients will be randomised to receive at least one treatment and to have a therapeutic evaluation after the treatments.’

It is not clear what is meant here. Some patients are randomised to control hence not to any treatment. I think you mean that all patients will be included regardless of the number of treatments they receive and multiple imputation will be used to impute missing data values.

Reply: Thanks for your suggestion. According to your advice, we have made corresponding modifications and marked it in red. Page 16.

10. Per protocol analysis without adjustment for confounding is potentially biased. See N Engl J Med 2017; 377:1391-1398 DOI: 10.1056/NEJMsm1605385. It is unclear why multiple imputation has not been used here.

Reply: Thanks for your suggestions. We will assess the primary outcome by generalized linear model, including group, baseline SBMs and study site as fixed effects and age, visit number, rescue medicine and other defecation as possible covariates, simultaneously in full analysis and per protocol analysis. Multiple imputation will be used to impute missing data in full analysis set and per-protocol set.

11. Safety assessment – this not clearly described as a measure applicable to both groups and it isn’t clear what is being proposed here.

Reply: Thanks for your suggestion. We adjusted accordingly in Table 1 and “Safety assessment” section, and marked it in red. Page 9, 14.

12. There is an over reliance on significance test in the analysis. Significance tests at baseline are not generally recommended as the null hypothesis should be true unless there is something wrong with the randomisation (see CONSORT).

For analysis of efficacy an estimate of effect with confidence interval should be the main focus.

Reply: Thanks for your suggestions. We looked at CONSORT and did find it not recommended for significance tests at baseline. We have deleted the relevant content. For analysis of efficacy an estimate of effect with confidence interval will be used in our trial. We added the sentence “The result (or results) will be presented by P value and 95% confidence interval. ” in the “statistical analysis” part of the primary outcomes and secondary outcomes, and marked it in red. Page 17, 18.

13. Baseline value should be included in the model to avoid regression to the mean and potential for biased results. Other covariates should also be chosen in advance and the protocol should state when this will be done. Using baseline data to decide covariates for the analysis may introduce bias.

Reply: Thanks for your suggestions. Based on our knowledge of constipation in Parkinson's disease and experience with trial design, we redefined group, baseline SBMs and study site as fixed effects and age, visit number, rescue medicine and other defecation as possible covariates in advance(Page 17). Meanwhile, we referred to the statistical strategies in a similar article. [Ann Intern Med 2016;165(11):761-69.].

14. It is a good idea to write a full statistical analysis plan and to have this reviewed by the statistician on your DMEC before unblinding of the data

Reply: Thanks for your encouraging comment.

15. DMEC – when will the interim analysis be carried out and for what purpose?

Reply: Thanks for your question. Having fully discussed the advantages and disadvantages of conducting the interim analysis we have chosen not to do it. So we deleted the sentence of “The DSMB will have access to the interim results. ”.

16. Data Sharing statement is unclear what is meant by ‘basic information’ – does this include age and gender? You may need to assess risk of re-identification of subjects with respect to demographic data and create categories of age to reduce re-identification. I clicked on the link given and it appears to be in Chinese- will the data be available to a wider international audience.

It might be prudent to give a less open access or to put in some qualifier –e.g. subject to data protection legislation. In the UK the law governing data and data sharing has changed over the years and it will be several years before the data would become available

You may need to assess risk of re-identification of subjects with respect to demographic data and create categories of age to reduce re-identification.

Reply: The basic information includes the number of participants, gender, inclusion time, inclusion location, and entry personnel, excluding age. When we register the trial, it requires that raw data must be shared. The Research manager is the most commonly used data management platform in China. The Redcap electronic data capture (EDC) system will be used for data entry and data management. According to the data dictionary we could convert the results into English. If someone want to get the data, we could provide relevant English version of the data after publishing relevant results of the trial.

17. Flow chart

What does heirachical random mean?

The flow chart does not follow the normal CONSORT format and is very confusing.

Reply: Thanks for your suggestion. The meaning of “hierarchical random” is stratified randomization, it is a language mistake and we have corrected it. Meanwhile, the flow chart was corrected as the normal CONSORT format. Page 25.

VERSION 3 - REVIEW

REVIEWER	Guihua Tian dongzhimen hospital, Beijing University of Chinese Medicion, China
REVIEW RETURNED	24-Oct-2019

GENERAL COMMENTS	This research has certain significance

REVIEWER	Irina Chis Ster St George's University of London
REVIEW RETURNED	14-Sep-2019

GENERAL COMMENTS	The manuscript has considerably improved and that include the English language. I welcome the inclusion of a professional statistician on board. I also welcome the publication of the pilot data and clarifying their input parameters to power the trial and the choice of their outcome - this is important. I wish them good luck with data collection.