Author’s response to reviews

Title: A new phototherapy regimen during winter as an add-on therapy, coupled with oral vitamin D supplementation, for the long-term control of atopic dermatitis: study protocol for a multicentre, randomized, crossover, pragmatic trial. The PRADA trial

Authors:

Catherine Droitcourt (catherine.droitcourt@chu-rennes.fr)

Sébastien Barbarot (sebastien.barbarot@chu-nantes.fr)

Annabel Maruani (annabel.maruani@univ-tours.fr)

Laure Darrieux (laure.darrieux@ch-stbrieuc.fr)

Laurent Misery (laurent.misery@chu-brest.fr)

Emilie Brenault (emilie.brenault@chu-brest.fr)

Henri Adamski (henri.adamski@chu-rennes.fr)

Cécile Chabbert (cecile.chabbert@ch-perigueux.fr)

Annie Vermersch (vermersch-a@ch-valenciennes.fr)

Marie Weiborn (weiborn.a@ch-valenciennes.fr)

Julien Seneschal (julien.seneschal@chu-bordeaux.fr)

Patrice Plantin (p.plantin@ch-cornouaille.fr)

Hervé Maillard (hmaillard@ch-lemans.fr)

Alice Phan (alice.phan01@chu-lyon.fr)

François Skowron (fskowron@ch-valence.fr)

Manuelle Viguier (mviguier@chu-reims.fr)

Delphine Staumont-Salle (delphine.salle@chru-lille.fr)

Audrey Nosbaum (audrey.nosbaum@chu-lyon.fr)

Angèle Soria (angele.soria@tnn.aphp.fr)
Dear Editor,

We are very pleased to be given the opportunity to submit a revised version of our manuscript TRLS-D-18-00412 entitled “A new phototherapy regimen during winter as an add-on therapy, coupled with oral vitamin D supplementation, for the long-term control of atopic dermatitis: study protocol for a multicentre, randomized, crossover, pragmatic trial. The PRADA trial.” by Catherine Droitcourt et al. We thank the referee for their careful reading, and their insightful comments. We are pleased to address their comments in a point-by-point response below. We did our best to take into account all the points and we believe we are now able to submit a substantially improved revised version of our manuscript.
In the manuscript, changes are in colored text. As requested, we provide both marked and unmarked version of the manuscript.

We hope that the changes we made have improved our manuscript and made it acceptable for publication in Trials.

On behalf of the authors,

Yours sincerely,

Catherine Droitcourt

Reviewer #1: In general, this paper is quite well written and the protocol generally follows SPIRIT 2013 guidelines for a clinical trial protocol. However, several issues need to be addressed.

1. There is a reference to "Figure 1" in the paper, however, no figure has been included with the manuscript. Please correct this.

Reply: Thank you for your comment. We do apologize for that; the Figure 1 has been included with the revised manuscript.

2. Table 1 mentions 18 post-screening visits - how often will visits occur? Every 6 weeks? Please clarify in the text.

Reply: Thank you for your comment. A patient will be in the trial for 2 years post-randomization. After the inclusion visit, visits will occur every three months during 2 years, not every six weeks.

Furthermore, PO-SCORAD score will be computed every four weeks; patients will fill the self-administered questionnaire (PO-SCORAD) at home through an application onto their computer or phone, or on a sheet of paper, every four weeks.

These points have been clarified and corrected in the Table 1 and in the text as follows:

“After the inclusion visit, visits will occur every three months during two years. Furthermore, each patient will fill a self-administered questionnaire including PO-SCORAD and POEM at
home through an application onto their computer or phone, or on a sheet of paper otherwise, every four weeks.” (page 14, lines 287-290)

3. The Outcomes section needs to be clarified. Will all outcomes be used to evaluate the effects of both phototherapy and vitamin D interventions? The "Primary outcome" subsection needs to be renamed "Primary outcomes" and re-worded. Only at the end of the Statistical methods section, it became clear that there is not one primary outcome constructed using a hierarchical procedure, but actually two primary outcomes that will be evaluated with a multiplicity adjustment.

Reply: We agree with reviewer's comment: indeed, there are two primary endpoints that will be tested following a hierarchical testing procedure. The "Primary outcomes" subsection is re-worded as follows:

“Primary outcomes. The primary outcomes are: (1) repeat measures of the PO-SCORAD severity score over 1 year; (2) cumulative consumption of TAT (number of tubes) during the winter. They will be tested following a hierarchical testing procedure.” (page 11, lines 216-221)

First, EASI that is an investigator-based severity score has been chosen instead of PO-SCORAD at the time of the submission of the study protocol for funding by the French Ministry of Social Affairs and Health (French National Program of Clinical Research [PHRC-N]). The EASI was selected because it was considered the best physician-assessed instrument to assess the clinical signs of AD, according to recent HOME recommendations [21]. Once the project has received funding by the French Ministry of Social Affairs and Health, PO-SCORAD was considered one of the best patient-assessed instruments to assess the symptoms of atopic dermatitis. The choice of a patient-related score as the primary outcome will allow the capture of more data points without increasing the number of visits and also respecting the pragmatic design of the study. This amendment to the study protocol has been submitted and validated by all the investigators before the start of the study and the first inclusion. We do apologize for this mistake and we provide the complete study protocol (additional file 6, see page 33 in section 4.3.1. primary endpoints) in additional file and the link to the clinicaltrials website: https://clinicaltrials.gov/ct2/show/NCT02537509?term=droitcourt&rank=1. Please note the date when modification was applied (on September 1, 2015).

This point has been modified in the revised version of the manuscript as follows:

“The primary outcomes are: (1) repeat measures of the PO-SCORAD severity score over 1 year; (2) cumulative consumption of TAT (number of tubes) during the winter. They will be tested following a hierarchical testing procedure.” (page 4, lines 75 to 78)
“The main predictor variable is phototherapy (yes or no) and the response variable is numerical (PO-SCORAD).” (page 12, lines 243 and 244)

“Of note, the minimum clinically significant difference for the PO-SCORAD score is around 9 points. (page 13, lines 255 and 256)

"The statistical issue for the primary outcomes is the analysis of repeated measures (up to 12 measures per each one-year period, PO-SCORAD, self-administered questionnaire collected through an application onto their computer or phone, at home) applied to a crossover (two one-year periods, two sequences) design." (page 16, lines 335 to 338)

“we have chosen the average and the standard deviation of the PO-SCORAD score, the area under the curve (AUC), and the mean amplitude of PO-SCORAD score ranges. A normal distribution is assumed for the primary outcomes (PO-SCORAD severity score over two one-year periods; cumulative consumption of TAT during winter) and this will be checked at this stage." (page 17, lines 355 to 359)

“The multiplicity adjustment strategy used is a hierarchical closed test procedure: the primary endpoints are ordered as follows: (E1) repeat measures of the PO-SCORAD severity score over 1 year; (E2) cumulative consumption of TAT over winter." (page 17, lines 366 to 368)

4. Please explain why EASI scores will be analyzed only over 1 year, while patients are going to be followed up for 2 years and the secondary outcomes will also be analyzed over 2 years. I suppose this makes sense for the evaluation of phototherapy, since only winter data will be used (I assume) - but what about the evaluation of the effect of vitamin D?

Reply: The trial is designed to demonstrate that phototherapy as an add-on to standard care is superior to standard care using a crossover design (two one-year periods). It is anticipated that vitamin D could act as a quantitative effect-modifier on phototherapy. Hence, main analysis on primary outcomes will be conducted on two one-year periods. The "Statistical methods" is re-worded as follows:

“Statistical methods. The statistical issue for the primary outcomes is the analysis of repeated measures (up to 12 measures per each one-year period, PO-SCORAD, self-administered questionnaire collected through an application onto their computer or phone, at home) applied to a crossover (two one-year periods, two sequences) design." (page 16, lines 335 to 338)
5. Recruitment: it is fairly concerning that patients could be screened in January yet would have to wait until September to be randomized. Some patients may drop out of the study before they are even randomized because of the long wait. What if in September it turns out that the number of patients that can be randomized is less than planned - will additional patients have to be identified and then randomized next September, delaying the trial completion? Perhaps the authors could address this concern in the Discussion. The alternative, of course, would be to allow patients to be randomized throughout the year and then address seasonal effects in the analysis.

Reply: We agree and we have discussed this point in the discussion section as follows:

“We choose a fixed and common period for inclusion to avoid confounding by seasonality. Even if patients will be able to sign their written consent at any time in the pre-screening period, they will be randomized in September and we cannot exclude that some patients may drop out of the study before the randomization. The alternative would be to randomize patients throughout the year and to take into account seasonal effects but the approach to measure it would have been complex in a study with pragmatic options and long-term follow-up.” (page 22, lines 475 to 481)

6. The Statistical methods section needs to be considerably clarified. Importantly, specific hypotheses need to be specified - without them, it is impossible for the reader to evaluate whether the chosen models are even appropriate. When repeated measures are analyzed, there are many possible hypotheses that could be of interest. For example, investigators may expect that in both placebo and vitamin D groups, measurements follow average linear trajectories over time and one hypothesis of interest may be that the trajectory in vitamin D group has negative slope, indicating that EASI scores decrease over time (I assume that smaller scores are better - this needs to be clarified in the paper), while the trajectory in the placebo group may be expected to be flat. In this situation, the appropriate hypothesis would concern difference in the trajectory slopes. Or - investigators may not be willing to assume a specific form of the trajectories (i.e. linear) and instead they may be interested in the estimated average difference at a single time point (often, last visit). Or - the investigators may be interested in comparing average differences across all time points, etc. In this study, I cannot tell from the paper which hypothesis (or hypotheses) the authors intend to evaluate.

Reply: When repeated measures are analyzed, there are indeed many possible hypotheses that could be of interest. We focused on a main effect hypothesis test for the effect of a single predictor variable (phototherapy) averaged across all other factors and sample size calculation was based accordingly. In other words, we wish to test if PO-SCORAD severity score differs among any of the follow-up times, when averaged across the two treatment groups (phototherapy or not). This hypothesis tests for the main effect of time on PO-SCORAD severity score.
The first sentence in the Statistical methods section says that 4 repeated measures will be analyzed for the primary outcome - please clarify what you mean by "4 measures", do you refer to the number of the outcomes or the number of repeated measures? In general, description of the model(s) needs to be much clearer, and I would urge the authors to consider how models of this type are usually described in the literature. Specifically, please clarify:

Reply: The primary outcome is the analysis of repeat measures of PO-SCORAD recorded every four weeks. This point has been clarified and corrected as follows:

“The statistical issue for the primary outcomes is the analysis of repeated measures (up to 12 measures per each one-year period, PO-SCORAD, self-administered questionnaire collected through an application onto their computer or phone, at home) applied to a crossover (two one-year periods, two sequences) design.” (page 16, lines 335 to 338)

1) whether a single model or separate models will be used to evaluate the effects of the two interventions, phototherapy vs. no phototherapy and vitamin D vs. placebo.

Reply: We will use a single model with an interaction term between Treatment-2 (vitamin D / placebo) and Treatment-1 (Phototherapy / No phototherapy).

2) The distribution that will be assumed for the outcomes (normal? Poisson? etc.)

Reply: Primary endpoints are continuous: PO-SCORAD severity score over two one-year periods; cumulative consumption of TAT during winter, and normal distribution is assumed. This assumption will be checked at the descriptive phase of the analysis.

3) The number of repeated measures that will be analyzed for each outcome. Currently, it appears that for the analysis of the effect of phototherapy 8 repetitions of each outcome (4 measured in the first and 4 in the second winter) will be analyzed, is this correct? If a separate model were to be used for the analysis of the effect of vitamin D vs. placebo, the authors need to clarify whether they will include in the model a) outcomes collected at all visits, or b) outcomes collected at all post-screening visits, or c) outcomes collected at visits 2-18, or d) some other combination.

Reply: We acknowledge that the text was misleading. The crossover design encompasses two one-year periods with up to 12 measurements (self-administered questionnaire). The main analysis will then use up to 12 measurements per one-year period.
4) The terms that will be included in the model(s). Currently, it appears that for the analysis of the effect of phototherapy, the model will include terms for treatment (phototherapy vs no phototherapy), time period (first/second winter), treatment order sequence (Phototherapy - No phototherapy / No Phototherapy - Phototherapy) and carry-over term, which is appropriate. Of note, if this model will only be used to analyze outcomes from the first and second winter, then a carry-over term should not be necessary - in fact, the authors made this point on p.20. Additionally, the authors note that they would consider including interaction between vitamin D/placebo and Phototherapy/No phototherapy terms to allow for possible effect modification. In that case, it would also be advisable to include a main effect for vitamin D/placebo in the model.

Reply: We agree with reviewer's comment and the revised version clearly mentions which terms will be included in the model, interaction term and main effects.

5) It appears that the modeling analysis of the effect for vitamin D/placebo have not been addressed at all, apart from possible phototherapy effect modification. Please either confirm that this was indeed the intention or include the detailed description for such model.

Reply: It is anticipated that vitamin D could act as a quantitative effect-modifier on phototherapy. The revised version clearly mentions that two main effects terms and an interaction term will be included in the model.

Please note that the items "Methods for any additional analyses (eg, subgroup and adjusted analyses)" and "Definition of analysis population relating to protocol non-adherence (eg, as randomised analysis), and any statistical methods to handle missing data (eg, multiple imputation)" from the SPIRIT checklist have not been satisfactorily addressed in the paper. Please include these items in the paper or alternatively, include a reference to a document where these details can be found.

Reply: These items are addressed in the revised version. Briefly, no subgroup analyses are planned. Adjustment will be made on variables used in the allocation process. The main analysis will be based on the adjusted model and will follow the "intention to treat" principle. Lastly, mixed model allows handling missing data in analysis of repeated measurements. The "Statistical methods" is re-worded as follows:

“Statistical methods. The statistical issue for the primary outcomes is the analysis of repeated measures (up to 12 measures per each one-year period, PO-SCORAD, self-administered questionnaire collected through an application onto their computer or phone, at home) applied to a crossover (two one-year periods, two sequences) design. When repeated measures are analyzed, there are many possible hypotheses that could be of interest. We focused on a main
effect hypothesis test for the effect of a single predictor variable (treatment: phototherapy) averaged across all other factors and sample size calculation was based accordingly. We are interested in the usual factors: Treatment (Phototherapy / No phototherapy) and Time. However, the nature of the design (crossover trial) has to be taken into account. This means that we have to consider the following factors: Patient, Period: first winter / second winter, Order: Phototherapy - No phototherapy / No Phototherapy – Phototherapy, and Carry-over. It was anticipated that vitamin D could act as a quantitative effect-modifier on phototherapy. Then, we decided to include an interaction term in the analysis. We will use a single model with the following main effect Treatment-1 (Phototherapy / No phototherapy), Treatment-2 (vitamin D / placebo), Patient, Period, Order and Carry-over term as well as an interaction term between Treatment-2 and Treatment-1. Adjustment will be made on variables used in the allocation process (regional area (north versus south), age as a continuous variable and disease severity (moderate versus severe)).

For descriptive purposes, it is instructive to carry out an analysis based on a summary of the statistics (a quantity calculated from each curve), which can reflect important aspects of the problem at hand: we have chosen the average and the standard deviation of the PO-SCORAD score, the area under the curve (AUC), and the mean amplitude of PO-SCORAD score ranges. A normal distribution is assumed for the primary outcomes (PO-SCORAD severity score over two one-year periods; cumulative consumption of TAT during winter) and this will be checked at this stage.

The main analysis will be based on the adjusted model and will follow the "intention to treat" principle: all participants, as randomized will be analyzed; outcome data obtained from all participant, regardless of protocol adherence will be used. Mixed model techniques (PROC MIXED) will be applied with a restricted maximum likelihood (REML) estimation method, because it enables the use of different covariance structures for the covariance matrix, in order to find the most suitable covariance structure for the data.

The multiplicity adjustment strategy used is a hierarchical closed test procedure: the primary endpoints are ordered as follows: (E1) repeat measures of the PO-SCORAD severity score over 1 year; (E2) cumulative consumption of TAT over winter. We will test E1 and E2 sequentially at the same two-sided level of 0.05 as follows: We will test E1 first. If it proves significant, will we test E2. A Wald test with Kenward-Roger degrees of freedom will be used.

No subgroup analyses nor interim analysis were planned.

For secondary analyses we will use up to 4 measurements per one-year period, collected at each clinical visit (every three months) for: investigator-based severity scores (EASI, SCORAD, IGA); patient-reported severity and quality of life scores (POEM, DLQI); serum Vitamin D levels (25-(OH)-vitamin D); total IgE serum levels; weeks of satisfactory control; inter-visit
cumulative consumption of TAT; synthetic patient-reported satisfaction at the end of each winter.

All analyses will use procedures available in SAS software 9.4 (SAS Institute, Carry, N.C., USA).” (Pages 16 to 18, lines 335 to 380).

7. Sample size: please include a reference for GLIMMPSE, as well as the version of this software. It appears that sample size calculations were only done for the evaluation of one intervention (I assume it’s phototherapy) - were sample size calculations performed for the evaluation of vitamin D vs. placebo? Sample size evaluation needs to match the analysis methods specified in the Statistical methods section, and I am not sure whether this is the case. A mixed model fitted using PROC MIXED with REML does not provide Hotelling-Lawley trace test, why was this particular test used for sample size evaluation? Please provide some brief justification.

Reply: The following reference for GLIMMPSE is cited in the revised version:

Kreidler SM, Muller KE, Grunwald GK, Ringham BM, Coker-Dukowitz ZT, Sakhadeo UR, Barón AE, Glueck DH. GLIMMPSE: online power computation for linear models with and without a baseline covariate. J Stat Softw. 2013 Sep;54(10). (page 28, lines 639 to 641)

The version 2.1.0 of this software was used by the time of calculation.

Sample size calculation was indeed only done for the evaluation of one intervention (phototherapy) as the main goal of the trial was to demonstrate that phototherapy as an add-on to standard care is superior to standard care.

We acknowledge that sample size calculation did not take into account interaction with vitamin D, keeping in mind that a quantitative effect-modification of vitamin D on phototherapy was anticipated, not a qualitative interaction.

We wish to test if PO-SCORAD severity score differs among any of the follow-up times, when averaged across the two treatment groups (phototherapy or not). This hypothesis tests for the main effect of time on PO-SCORAD severity score. For this kind of design with a simple between hypothesis and a complex within hypothesis, Hotelling-Lawley trace test was recommended for sample size calculation. We indeed plan to use a mixed model for our data analysis, and are going to use the Wald test with Kenward-Roger degrees of freedom, and we followed GLIMMPSE recommendation for sample size analysis using the Hotelling-Lawley test for the general linear multivariate model. Of note, Hotelling-Lawley trace test for the general linear multivariate model coincides with the Wald test for the general linear mixed model with Kenward-Roger degrees of freedom when there are no missing observations, and each independent sampling unit has the same number of observations.
The section of sample size is re-worded as follows: "We used GLIMMPSE [31] to calculate the sample size for a repeated measure analysis with a desired power of at least 80%, a Type I error set at 0.05 and equal group sizes. The main predictor variable is phototherapy (yes or no) and the response variable is numerical (PO-SCORAD). Repeated measures are described as follows: 10 to 12 measures with time as the unit, and an equal distance between repeat numerical measures (every 4 weeks); we entered means for each time according to an anticipated evolution over time, standard deviations of the response variable set at 18 and any correlation between them for each level of repeated measures. GLIMMPSE currently assumes that the standard deviation is constant across repeated measures. As we have some uncertainty about what standard deviation value to use, we used alternative values for variability and computed power for half the variance, the variance as specified, and twice the variance. This was also be done for means. GLIMMPSE automatically combines the sources of correlation into a final covariance matrix using a structured correlation based on the linear exponential auto-regressive model. The model describes a correlation that decreases monotonously with the distance between measurements; the base correlation is set at 0.6 and the decay rate at 0.05. Of note, the minimum clinically significant difference for the PO-SCORAD score is around 9 points.

We considered a main effect hypothesis test for the effect of a single predictor variable averaged across all other factors. We used the Hotelling-Lawley Trace as the recommended statistical test for this kind of design with a simple between hypothesis and a complex within hypothesis. We indeed plan to use a mixed model for our data analysis, and are going to use the Wald test with Kenward-Roger degrees of freedom. Of note, Hotelling-Lawley trace test for the general linear multivariate model coincides with the Wald test for the general linear mixed model.

On the basis of sample size simulation (Additional file 1), we chose a total sample size of 340 subjects. This calculation was performed considering a parallel-arm design; hence, using a crossover design, we thought we could halve the sample size. The assumption of no carry-over effect is admitted when planning sample size for a crossover trial. Considering an intervention only in winter, we thought carryover is minimized by design. Using an uniform and balanced crossover design the treatment difference will not be aliased with sequence or period effects. However, an interaction of the time and treatment should be taken into account as we anticipate this could happen.

Power consideration when considering repeated measures, crossover design and interaction of the time and treatment is not straightforward. Thus, we used GLIMMPSE as for a two-arm trial. We understand that halving sample size looks naive and we compared sample size calculation from (1) a two-arm parallel trial and (2) a crossover trial, but considering only the last measurement as the outcome, to check that sample size reduction was not too optimistic. Finally, anticipating that 10-15% of patients will not complete the entire study (two one-year sequences), the planned number of subjects to be enrolled is set at 200."
Provided "Additional file 1" with sample size and power calculations is certainly helpful, yet it is still confusing to me how the investigators ultimately chose their sample size. The first figure provides a scenario apparently used to evaluate a main effect hypothesis - although there is clearly time by treatment interaction present, according to the data entered, since the differences between groups at different time points are not the same. Using the provided numbers, my GLIMMPSE calculations resulted in a total of 168 subjects needed for the evaluation of the main effect hypothesis using Hotelling-Lawley trace test with 80% power. The second figure provides a different scenario, with no specified hypothesis. Assuming that the main effect hypothesis is still of interest, GLIMMPSE yielded that a total of 298 subjects are needed to achieve 80% power - note that this is more than 240 subjects selected by the authors. Finally, a table provided in the document considers different assumptions for the means and SD (quantified by scaling factors) - this table does not seem to be related to the previous two scenarios as it now states a clinically meaningful difference of 7 - so, either the outcome trajectories are not considered to be parallel in time between the two groups, or perhaps the hypothesis was switched from main effect to grand means. To summarize, the authors are to be commended for including an additional file with the details of the sample size and power calculations, but this file needs to be clarified considerably to clearly state the assumptions and the hypotheses tested. I believe it may be the easiest (and clearest) to simply include all the options used in GLIMMPSE for these calculations.

Reply: We provide in the revision version of "additional file 1" a detail of all the options used in GLIMMPSE for sample size calculation as well as a script to be uploaded to re-do the calculation. Briefly, the options selected were as follows:

1. solving for sample size
2. desired power: 0.80, 0.85, 0.90
3. model
   a. clustering: participant, no additional clustering
   b. predictor: treatment (phototherapy)
   c. covariate: none
   d. response variable: score (PO-SCORAD)
   e. repeated measure:
      i. dimension: time
      ii. type: numeric
      iii. number: 10
iv. spacing: equal

f. group sizes: equal

4. hypothesis
   
a. hypothesis type: main effect; factor of interest: between-participant factor: treatment
   
b. statistical test: Hotelling-Lawley test
   
c. type 1 error rate: 0.05

5. means
   
a. means
   
   40 39 38 37 36 35 34 33 32 32
   
   40 40 40 40 40 40 40 40 40 40

   b. scale factor for means: 0.9, 1.0, 1.1

6. variability
   
a. within-participant variability
   
   i. variability across time: base correlation: 0.06; decay rate: 0.05 (LEAR model)
   
   ii. variability across response (score): standard deviation: 18 (unstructured correlation)
   
   b. scale factors for variability: 0.5, 1, 2

7. options: don't include confidence interval for power

On p. 13, the authors state that "This calculation is performed considering a parallel-arm design; hence, using a cross-over design, only 120 patients are required." I believe this to be an incorrect statement. Sample size needed for a parallel arm trial cannot simply be halved in order to infer sample size needed for a cross-over trial since calculations for a parallel arm trial do not include assumptions about the effect of time period, treatment order sequence and carry-over effect.

Reply: The assumption of no carry-over effect is admitted when planning sample size for a crossover trial. Washout periods in the experimental design can diminish the impact of carryover effect. Considering an intervention only in winter, we thought carryover is minimized by design. Other standard constraints are also assumed: that is the interaction effects sum to zero over the
levels of time and the two levels of the main effect. It is also assumed that each set of error terms are independent and normally distributed. Using an uniform and balanced crossover design the treatment difference will not be aliased with sequence or period effects. However, an interaction of the time and treatment should be taken into account as we anticipate this could happen.

Power consideration when considering repeated measurement, crossover design and interaction of the time and treatment is not straightforward. Thus, we used GLIMMPSE as for a two-arms trial. We understand that halving sample size looks naive and we compared sample size calculation from (1) a two-arm parallel trial and (2) a crossover trial, but considering only the last measurement as the outcome, to check that sample size reduction was not too optimistic.

```
proc power;
   twosamplemeans gmeans=(40 32) stddev=18 power= 0.9 sides=2 ntotal=.; run;
proc power;
   pairedmeans test=diff meandiff=8 stddev=18 corr=.4 .2 power=.9 alpha=.05 sides=2 npaires=.; run;
```

Even with a correlation as low as 0.2, the number of pairs is half the sample size for a two-arm trial.

8. Please clarify whether or not the data monitoring technician will be independent from the sponsor and from the rest of the study investigators.

Reply: This point has been clarified as follows: "The data monitoring technician will be dependent from the sponsor and independent from the study investigators." (page 18, lines 386-387).

9. Please address the SPIRIT guideline item regarding interim analyses and stopping guidelines: "Description of any interim analyses and stopping guidelines, including who will have access to these interim results and make the final decision to terminate the trial".

Reply: These items are addressed in the revised version; specifically the "Statistical methods" section is modified (pages 17, line 372). Briefly, no subgroup analyses nor interim analysis were planned.