PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Triple-arm trial of pH (Tri-pH) effect on live birth after ICSI in Egyptian IVF facilities: Protocol of a randomised controlled trial
AUTHORS	Fawzy, Mohamed; Emad, Mai; Wilkinson, Jack; Mansour, Ragaa; Mahran, Ali; Fetih, Ahmed; Abdelrahman, Mohamed; AbdelGhafar, Hazem

VERSION 1 - REVIEW

REVIEWER	Ernest HY Ng Department of O & G, LKS Faculty of Medicine, HKSAR
REVIEW RETURNED	14-Sep-2019

GENERAL COMMENTS	The aim of this protocol paper was to examine if there is an effect on live birth rate using three levels of pH of culture media for IVF. Strength of the paper 1. There is no randomized study comparing different culture media pH on the pregnancy outcomes especially the live birth rate. 2. It is double blind study as the patients and the physicians would not be aware of the randomized arms. 3. There is clear description of the sample size calculation and randomization plan in the text. Major concerns 1. Introduction • The authors should state the adverse effects of extreme culture media pH in both animal and human IVF. 2. Materials and methods • It is not clear if the pH of the culture media would be measured or confirmed on each day or more frequently. • Only one culture media should be used through the study as different media may affect the results especially the change of media will be in some centres only. • Sample size calculation: there is no justification of using 10%. It is not sure if 93% power is chosen. A much drop-out rate of about 20-30% should be anticipated as patients may not have fresh embryo transfer for a number of reasons such as risk of ovarian hyperstimulation syndrome and high serum progesterone level on the day of hCG etc

REVIEWER	Yanping Kuang Department of Assisted Reproduction, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, People's Republic of China
REVIEW RETURNED	23-Oct-2019

GENERAL COMMENTS	Comments to the Author This manuscript presents a protocol of multicenter, randomized, triple-arm, clinical trial to evaluate the impact of three levels of extracellular pH during human embryo in vitro culture on live birth rate. The idea behind this paper is interesting and RCT with large population is worth for publication, however, this manuscript still exists some concerns about methodological issues and need to be improved before acceptance. major comments 1. What is your objective for this study? Term live birth or live birth? The title showed'term live birth' while the primary outcome was live birth (line 249). Dependent variable in the primary analysis was term live birth (line 293). I am really confused about the main goal of this study. In my understanding of the whole manuscript, it should be live birth rate rather than term live birth. If this, please revised the title and relative content. If not, please give the clear explanation. 2. Why this study set 25% as the basal live birth rate? According to previous studies or their own database? And why to choose 10% difference, rather than 5% or 15%, as the criterion for sample size estimation? These parameters are important and should be explained and discussed in the manuscript. 3. Are there any retrospective studies or other type of clinical studied on the relationship between pH level and clinical outcomes (such as fertilization, embryo development or implantation, etc.)? And why this study divided pH7.2, 7.3 and 7.4 as groups? These should be mentioned in the background part. 4. The detailed analytical methods among the three arms were not clear. The author intended make a pairwise comparison or overall comparison among three groups, in other words, 7.2 vs. 7.3 vs. 7.4 or 7.2 vs.7.3/ 7.2 vs. 7.4 / 7.3 vs. 7.4? If three groups compared using logistic regression, which one should be the reference? I strongly suggest the author to read this published paper (Juszczak, E., Altman, D. G., Hopewell, S., & Schulz, K. (2019). Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 Statement. Jama, 321(16), 1610-1620) and add detailed analytical methods 5. A flowchart of the study design is required to make the protocol easier and more logical for understanding. 6. For the primary analysis, the author used logistic regression, with term live birth event regressed on log(pH), adjusted for study site and participant age (line 294). Why to use log(pH)? Besides, since participant age was an adjusted variable, should it be added as stratified variable in Randomization and Masking part (line 120)? Minor comments 1. Line 31: Please provide the AMH value for inclusion criteria 2. Line 135:Please explain the meaning of 'a previous successful attempt'

	3. Line 262: For cumulative live birth, why the viable neonates were registered only after one fresh plus one vitrified-warmed? and only one year? 4. I suggest the outcomes should be compared and present as rates rather than numbers in outcome measures part. 5. The author should add the detail information about follow-up program and the participants lost follow-up should be considered.

REVIEWER	Ioannis Sfontouris Eugonia IVF Centre, Greece University of Nottingham, UK
REVIEW RETURNED	23-Oct-2019

GENERAL COMMENTS	This is a nicely designed 3-armed RCT. An advantage of the study is having a statistician/methodologist among the authors. However, I have made some comments that the authors should address, aiming to improve the quality and clarity of the protocol. • Line 37: “designer’s wishes”. Please rephrase • Line 72: replace intercellular with intracellular • Lines 96-7: It is important to specify the exact number of participating sites. This is vital for study design and randomization and must be stated prior to study onset. • Line 120 “stratified by trial site”: must know the exact number of trial sites. • Line 131: I assume “10” is the antral follicle count –please specify. Also, specify AMH level. • Line 135: Inclusion criterion 6 is very restrictive. On what basis did the authors choose it? • Exclusion criteria: women with PCO/PCOS should be excluded. Similarly, all women with the excessive response should be excluded (eg >18 follicles on day of trigger) as these women are at high risk of OHSS if triggered with hCG. • Exclude women triggered with GnRH agonist • COS protocol: Using both long and antagonist protocol introduces another source of variability in the study. The protocol should be one, or if both are used they should be controlled for in the statistical analysis. • hCG type: Using two different types (recombinant and urinary) and doses (250 μg or 10,000 IU) introduces another source of variation. Ideally, only one type of hCG should be used. In this respect, the authors must specify how they will trigger high responders at risk for OHSS. Especially 10,000 hCG should be avoided as this dramatically increases the probability of severe OHSS. • Incubators: I understand that using the exact same incubator in all trial sites is difficult. However, all participating sites must use a) benchtop inculcators with b) humidified atmosphere. Big box incubators or dry incubators should be excluded. It has been shown that humidity and incubator volume can impact embryo development so we want to remove this source of bias from the study. • pH measurements: This is very important for the credibility and reliability of the results. Blood gas analysers are acceptable, although their limitation is that they provide a snapshot pH measurement. Ideally, continuous pH monitoring should be used, but I understand it may be difficult for all trial sites to have this type of equipment.

	• Embryo transfer: The ET strategy is unclear and should be described better. If each site transfers different numbers of embryos and on different days of development, this introduces significant bias. How will this be addressed? • In addition, it is not possible to evaluate blastocyst formation rates in patients with Day 3 transfer. The authors must clarify how they will evaluate blastocyst formation and quality (only in patients undergoing Day 5 transfer?). • Secondary outcome 18 (line 273-4): The correct term is “live birth per embryo transferred”. Please replace. • Sample size calculation: A 10% difference in LB appears rather high and unrealistic to achieve. Where did the authors base their assumption for a 10% difference? • Why set the power at 90%? Perhaps, the authors could set an 80% power and assume a smaller difference (less than 10%) for more meaningful results. • Logistic regression: unless the protocol changes, according to my previous comments, the authors should also adjust for more confounders, such as presence of PCOS, type/dose of hCG, number of embryos transferred, day of ET, incubator humidity. • Is there going to be an interim analysis? • Please specify the authors’ roles. • Line 307: I believe the phrase about “manufacturers’ wishes” should be modified and toned down.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Response: Thank you for your recommendations and we hope to respond on them satisfactory.

Major concerns

1. Introduction

• The authors should state the adverse effects of extreme culture media pH in both animal and human IVF.

Response: Thank you for your recommendation. We have amended the Introduction sections to mention the possible dangerous effects of using extreme levels of pH (Highlighted in the revised manuscript: Page 4, Lines 93 to 102).

2. Materials and methods

• It is not clear if the pH of the culture media would be measured or confirmed on each day or more frequently.

Response: The pH measurement will occur twice weekly using blood gas analyser and it will be ensured with daily measurement of CO2. pH also will be measured every new batch of culture media and adjusted accordingly. To account for personal variations, pH will be measured in all centres using

the same calibrated device and the same personnel (Highlighted in the revised manuscript: Page 9, Lines 214–219 and Lines 226–230).

- Only one culture media should be used through the study as different media may affect the results especially the change of media will be in some centres only.

Response: All culture disposables, media and oil will be the same in all centres. If a change occurs, it will be across all centres at the same time, and it will be reported. Since the allocation is stratified by site, any change would be balanced between treatment arms in the study.

- Sample size calculation: there is no justification of using 10%. It is not sure if 93% power is chosen.

Response:

Our study has been designed to detect effects that we consider to be realistic and relevant. This has been partly informed by a recent review of estimated effect sizes and sample sizes in fertility RCTs undertaken by the trial statistician (JW): Stocking, et al., 2019: https://doi.org/10.1093/humrep/dez017 . In light of suggestions made by reviewer 2, we have amended our primary analysis slightly, and this has caused minor changes to the estimated power of the design.

The study will be amongst the largest conducted in IVF (see Stocking, et al., 2019, results, para 1 – note these figures relate to a sample of trials that were the largest conducted for each intervention). A minority of RCTs in this field are well-powered to detect improvements in live birth as large as 20 percentage points. We have powered on the assumption that the difference between the two most discrepant pH groups could be as low as 10 percentage points, with the third pH group differing by about 5 percentage points from either (note that 10 percentage points is not the same as 10% - we believe that the reviewer has intended to refer to the former rather than the latter here – the latter would correspond to an increase, for example, from 20% to 22% - or two percentage points). By simulation, we calculate that the study, with 646 participants per arm, will have very high power to reject the null hypothesis of no effect of pH in the event that the pH effect is as strong or stronger than this (power ~ 99% at a 5% significance level). Note that it makes sense to aim for a high-power value here, reflecting our relative uncertainty in relation to realistic effect size. This ensures that we maintain relatively good power in the event that the spread of birth rates is lower than anticipated. For example, if the birth rates are 26%, 30%, 33% (a spread of just seven percentage points) this sample size yields 86% power against a 5% significance level, and 66% at a 1% significance level. We have also been conservative in our inflation of numbers for drop out (see comments re: dropout below).

We have sent the R code used for power simulation with this response for consideration by the reviewers. The sample size paragraph has been edited in the manuscript.

A much drop-out rate of about 20-30% should be anticipated as patients may not have fresh embryo transfer for a number of reasons such as risk of ovarian hyperstimulation syndrome and high serum progesterone level on the day of hCG etc

Response:

The reviewer appears to assume here that participants who have been randomised, but who do not proceed to fresh transfer, would be excluded from the analysis. This would violate the intention to treat principle, and would be fatal to our ability to make a randomised inference from the trial. All women randomised will included in the analysis according to the groups they were allocated to. This is the intention to treat principle. As such, women who do not proceed to have fresh transfer are included in the analysis, and are recorded as not having a live birth. The live birth rates used in the power calculation are based on actual practice, and incorporate failure to proceed to fresh transfer.

We have allowed for a 'dropout' rate of 5%, to allow for the possibility that a small number of women might withdraw their consent for their data to be used. Barring this, all randomised women will be incorporated in the analysis, and so we actually anticipate having more than 646 women per arm available for analysis.

We would like to thank this reviewer, once again for the recommendations to improve on the manuscript.

Reviewer: 2

Response: Thank you for your recommendations and we hope to respond on them satisfactory.

1. What is your objective for this study? Term live birth or live birth? The title showed 'term live birth' while the primary outcome was live birth (line 249). Dependent variable in the primary analysis was term live birth (line 293). I am really confused about the main goal of this study. In my understanding of the whole manuscript, it should be live birth rate rather than term live birth. If this, please revised the title and relative content. If not, please give the clear explanation.

Response: The primary outcome of the study is stated on page 10 of the manuscript: "Live birth (delivery of one or more viable infants > 20th weeks of gestation)." We apologise for confusion caused by not using consistent wording elsewhere. We have standardised all references to live birth in the manuscript including the title.

2. Why this study set 25% as the basal live birth rate? According to previous studies or their own database? And why to choose 10% difference, rather than 5% or 15%, as the criterion for sample size

estimation? These parameters are important and should be explained and discussed in the manuscript.

Response: Apologies for the lack of detail around this point in the manuscript, which we agree is important. We have amended our primary analysis strategy slightly in light of the reviewer's suggestion, and this has led to some slight changes to the power calculation. Please see the response to Reviewer 1 above, where this is discussed in detail. In relation to the reviewer's query about 25% as an indicative LBR, this has indeed been informed by internal data.

We reiterate that a recent review of power and precision in IVF trials conducted by the trial statistician co-author of the present protocol indicated that such a high level of power against effects of this magnitude is fairly exceptional in this field (Stocking, et al., 2019: https://doi.org/10.1093/humrep/dez017). Per your recommendation, these points are further explained in this revised submission (Highlighted in the revised submission: Page 12 and 13, Lines 301–320).

3. Are there any retrospective studies or other type of clinical studied on the relationship between pH level and clinical outcomes (such as fertilization, embryo development or implantation, etc.)? And why this study divided pH7.2, 7.3 and 7.4 as groups? These should be mentioned in the background part.

Response: Evidence relating to pH levels for human embryo culture is anecdotal or comes from manufactures of culture media. Therefore, we believe we have included most of the relevant data into this revised manuscript. The range of 7.2 to 7.4 has chosen based on the recommendation of manufactures of culture media as most of them recommend being between 7.2 to 7.4 as safe rage with 7.3 is the midpoint in this range. Based on this assumption, we have set our groups. In this version, we amended the manuscript to include this assumption (Highlighted in this revised submission: Page 4, Lines 93 to 102).

4. The detailed analytical methods among the three arms were not clear. The author intended make a pairwise comparison or overall comparison among three groups, in other words, 7.2 vs. 7.3 vs. 7.4 or 7.2 vs.7.3/ 7.2 vs. 7.4 / 7.3 vs. 7.4? If three groups compared using logistic regression, which one should be the reference? I strongly suggest the author to read this published paper (Juszczak, E., Altman, D. G., Hopewell, S., & Schulz, K. (2019). Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 Statement. Jama, 321(16), 1610-1620) and add detailed analytical methods.

Response:

We have revised the text (Pages 13 and 14, Lines 334–351) in light of the reviewer's helpful comments (and note that, as per standard practice, a full Statistical Analysis Plan will be drafted in the opening months of the trial). Having pairwise tests between all combinations of pH as the primary analysis is not feasible – achieving 90% power to compare live birth rates of 25% vs 30% at a 1% significance level would require 2371 participants in each of those two treatment arms, while a comparison of 30 vs 35% would require 2609 per arm, resulting in a trial size of over 6000

participants. While this would be the most informative approach, and would indeed be highly desirable, it is well beyond our ability to realise such a trial. Even if we increase our error rates (e.g. power of 80%, sig level of 1%) we still end up requiring ~ 4000 participants overall.

Our approach then has been to design the most informative analysis given feasibility constraints. In this case, and in light of the reviewer's comments, we have opted to consider a global test of the pH effect as our primary analysis. This can be used to reject the hypothesis that pH in this range is inconsequential.

We will then conduct additional supportive analysis to characterise any pH effect. This will include a test of linear trend in live birth across the groups, conducted by including pH as a continuous covariate in a logistic regression. Detection of a linear trend here would imply increasing (or, depending on the sign of the regression slope, decreasing) live birth rate with increasing pH, indicating the highest or lowest value as the best. We will also conduct exploratory pairwise comparisons of each pH group, with an emphasis on the magnitude and precision of estimated odds ratios.

While a primary analysis based on all pairwise contrasts would no doubt be the most desirable, it is not practicable given the sample size requirements. The analysis has been selected as the most informative analysis that can be practically accomplished.

5. A flowchart of the study design is required to make the protocol easier and more logical for understanding.

Response: Thanks for this recommendation. A flowchart is created and submitted with this revision.

6. For the primary analysis, the author used logistic regression, with term live birth event regressed on log(pH), adjusted for study site and participant age (line 294). Why to use log(pH)? Besides, since participant age was an adjusted variable, should it be added as stratified variable in Randomization and Masking part (line 120)?

Response: Thanks, the reviewer is right to say that taking a log transform of pH is not necessary because pH is already logarithmic in form. We have amended this in the analysis section.

Re: stratification vs adjustment variables: it is true that stratification variables must be adjusted for as covariates in the analysis (hence, the inclusion of site as a covariate) but it is not true that all covariates must also be used to stratify the randomisation. In the case of a large trial such as this, it isn't clear that stratifying for age would have any real value (since a large number of randomised patients makes it unlikely that there would be nontrivial imbalance with respect to age). Nonetheless, it is important to adjust for key prognostic variables such as age to increase power when testing the pH odds ratio in the logistic regression.

See the practical example presented by Raab and colleagues: Raab, G. M., Day, S., & Sales, J. (2000). How to Select Covariates to Include in the Analysis of a Clinical Trial. Controlled Clinical Trials, 21(4), 330–342. doi:10.1016/s0197-2456(00)00061-1

Minor comments

1. Line 31: Please provide the AMH value for inclusion criteria

Response: Thanks for this note. We have specified AMH ≥ 5.4 pmol/L as the lower limit for inclusion.

2. Line 135: Please explain the meaning of 'a previous successful attempt'

Response: it is now "Women undergoing their first ICSI cycle or their second ICSI cycle after previous successful one".

3. Line 262: For cumulative live birth, why the viable neonates were registered only after one fresh plus one vitrified-warmed? and only one year?

Response: This is to get an idea about the cumulative outcome using a unified criteria and time frame. If we wait for completing the transfer of all surplus embryos, this will take a very long time and would prevent the trial from being reported within a reasonable timeframe. However, although very important, it is a secondary outcome.

4. I suggest the outcomes should be compared and present as rates rather than numbers in outcome measures part.

Response: We agree with reviewer and the statistical plan will report rates, where appropriate for the primary and secondary outcomes.

5. The author should add the detail information about follow-up program and the participants lost follow-up should be considered.

Response: We identified a one year from randomization provided that all women have given birth. All lost follow-up participant will be treated as negative in the analysis, in order to realise an intention to treat analysis. Also, our sample size is robust to be maintained with > 90% power even with a 5% loss to follow-up, although this only really becomes a factor if the withdrawn participants withdraw their consent for their data to be analysed (Highlighted in the revised manuscript: Page 12 and 13, Lines 301–320). See comments on this point in response to reviewer 1.

We would like to thank this reviewer, once again for the recommendations to improve on the manuscript.

Reviewer: 3

Response: Thank you for your recommendations and we hope to respond on them satisfactory.

- Line 37: "designer's wishes". Please rephrase

Response: It is now "designers' opinions" – thank you.

- Line 72: replace intercellular with intracellular

Response: corrected in this revised version.

- Lines 96-7: It is important to specify the exact number of participating sites. This is vital for study design and randomization and must be stated prior to study onset.

Response: The study includes only the reported centres thus far. If other centres are included before starting of recruitment, we will amend the randomization per centres and report this in the final report. We clarified this point in this revised submission (Highlighted: Page 5, Lines 108–112).

- Line 120 “stratified by trial site”: must know the exact number of trial sites.

Response: The sites are included in the manuscript and we do not expect to invite more centres. However, if happened, it will occur before randomization or recruitment of any participant and will be reported in the final report of the study.

- Line 131: I assume “10” is the antral follicle count –please specify. Also, specify AMH level.

Response: Thanks for this recommendation. We amended this submission to include the AFC and AMH cut-off values (≥ 5 antral follicles count mean or ≥ 5.4 pmol/L AMH).

- Line 135: Inclusion criterion 6 is very restrictive. On what basis did the authors choose it?

Response: The authors have chosen these groups of patients to reduce the noise that can be introduced by including different categories of patients. We assume these subgroups are of good prognosis and can allow us to draw an intervention related conclusion.

- Exclusion criteria: women with PCO/PCOS should be excluded. Similarly, all women with the excessive response should be excluded (eg >18 follicles on day of trigger) as these women are at high risk of OHSS if triggered with hCG.

Response: We agree with the reviewer to exclude PCOS or any patient with a plan for freeze-all. The current version is amended to include “and 11) Severe PCOS, hyper-responder, OHSS patients, and cycles with agonist trigger or any patient with a plan for a “freeze-all”.

- Exclude women triggered with GnRH agonist

Response: We agree with the reviewer and the protocol is now amended.

- COS protocol: Using both long and antagonist protocol introduces another source of variability in the study. The protocol should be one, or if both are used they should be controlled for in the statistical analysis.

Response: We appreciate this viewpoint, but we do not agree with the reviewer as the current evidence in two meta-analyses (Al-Inany et al 2016 in Cochrane and Lambalk et al 2017 in HRU) suggest that both protocols are equivalent regarding the live birth rate. Both protocols are daily practice in our centres and omitting one of them will make things difficult. The other point to note is that stratification of the randomisation by centre means that any centre-specific protocols relating to stimulation will be balanced across the three study arms.

- hCG type: Using two different types (recombinant and urinary) and doses (250 μg or 10,000 IU) introduces another source of variation. Ideally, only one type of hCG should be used. In this respect, the authors must specify how they will trigger high responders at risk for OHSS. Especially 10,000 hCG should be avoided as this dramatically increases the probability of severe OHSS.

Response: Although we believe that both HCG can work equivalently, we amended this version to include only 10.000 IU hCG and we omitted the use of recombinant form. If we used it due to any circumstances, this will be reported in the final report. Patients amenable for OHSS are excluded from this revised submission of the protocol.

- Incubators: I understand that using the exact same incubator in all trial sites is difficult. However, all participating sites must use a) benchtop inculcators with b) humidified atmosphere. Big box incubators or dry incubators should be excluded. It has been shown that humidity and incubator volume can impact embryo development, so we want to remove this source of bias from the study.

Response: We are planning for using a single brand of incubators within each centre and all participating centres use only benchtop incubators. We are also convinced about the importance of humidity.

- pH measurements: This is very important for the credibility and reliability of the results. Blood gas analysers are acceptable, although their limitation is that they provide a snapshot pH measurement. Ideally, continuous pH monitoring should be used, but I understand it may be difficult for all trial sites to have this type of equipment.

Response: Very much appreciated. We hope the twice weekly measurement of pH using a stringent protocol can make us sure of the pH levels.

• Embryo transfer: The ET strategy is unclear and should be described better. If each site transfers different numbers of embryos and on different days of development, this introduces significant bias. How will this be addressed?

Response: Centres will transfer embryos on day 5 except for one centre has very minimal portion of day-3 transfer. Only a maximum of two embryos will be transferred in all centres using the same transfer medium, catheter and protocol, which can give some control on this issue. The randomisation is stratified per site, so this is nuisance variation, not bias (the latter refers to systematic error effecting the arms in a differential manner); procedures will be divided equally between arms. Adjusting for site in the analysis then removes this nuisance variation from the estimate of the treatment effect.

• In addition, it is not possible to evaluate blastocyst formation rates in patients with Day 3 transfer. The authors must clarify how they will evaluate blastocyst formation and quality (only in patients undergoing Day 5 transfer?).

Response: The portion of day 3 transfer is very limited and the implantable embryos from this portion will be included as formed and high-quality blastocyst, while embryo that will not implant from day 3 portion will be considered blocked at day three. However, all the embryos transferred on day 3 or 5 as well as those cryopreserved will be considered in the utilizable embryo rate. This version of the protocol is amended to include this information (Highlighted: Page 14, Lines 339–351). Another point to be considered is that although embryo development parameters are important, the study identified a clinical outcome (live birth rate) as a primary endpoint, which is the ultimate goal of an IVF procedure.

• Secondary outcome 18 (line 273-4): The correct term is “live birth per embryo transferred”. Please replace.

Response: corrected and highlighted in the manuscript.

• Sample size calculation: A 10% difference in LB appears rather high and unrealistic to achieve. Where did the authors base their assumption for a 10% difference?

Response: Please see our responses to reviewer 1 and 2 on the subject of power. We apologise for any confusion around this point. The power calculation does not suppose a 10 percentage point difference between each group; it allows for 5 percentage point differences between groups and the 10 percentage point value refers to the difference between the best and worst of the three groups. As noted in the response to reviewer 1, our analysis allows us to detect a pH effect with reasonable power even if the range of live birth rates is somewhat less than this. Moreover, our assumptions re: withdrawn participants are conservative, and we will increase our power by adjusting for prognostic covariates.

- Why set the power at 90%? Perhaps, the authors could set an 80% power and assume a smaller difference (less than 10%) for more meaningful results.

Response: Despite common beliefs around this point, this isn't really how power calculations work. A given sample size for a fixed alpha level represents an infinite range of power levels – one for each conceivable effect size. Choosing one pair of these to report rather than another does not result in any changes to the study, to the analysis, or to the inference that can be made from a study. If the reviewer means to say here that we should consider a larger sample size, then hopefully the responses we have detailed above make it clear that this is not practical – the sample size suggested already puts this amongst the largest RCTs to have been conducted in the field – see the review on this topic by the trial statistician JW (Stocking, et al., 2019: https://doi.org/10.1093/humrep/dez017).

- Logistic regression: unless the protocol changes, according to my previous comments, the authors should also adjust for more confounders, such as presence of PCOS, type/dose of hCG, number of embryos transferred, day of ET, incubator humidity.

Response: While we appreciate the sentiment here, adjusting for the covariates listed by the reviewer here pose a high risk of rendering the analysis noninformative, or even statistically invalid. The first point to note is that it is common, but not correct, to refer to other sources of variation as 'confounders' in RCTs. This isn't pedantry – it has material consequences for how we treat these variables in an analysis. In an RCT, we adjust for covariates not to control for 'confounding', but to increase power and precision of the test and estimation of the treatment effect. When the outcome is continuous, this is straightforward – we know that adjusting for predictive covariates in a linear model improves power and precision. When we have a binary outcome, and have to use a nonlinear model such as logistic regression, the situation is not so straightforward. Adjusting for weakly predictive covariates can actually increase imprecision, and, because the study is randomised, offers no benefit in relation to 'confounding'. This is not a recent discovery – see, for example, Robinson and Jewell 1991: https://www.jstor.org/stable/pdf/1403444.pdf?seq=1#page_scan_tab_contents .

Moreover, some of the variables mentioned by the reviewer here are post-randomisation variables. It is not a valid statistical strategy to adjust for post-randomisation variables in an RCT – indeed, doing so discards the benefit of randomisation. See, for example, the guidance document Guideline on adjustment for baseline covariates in clinical trials from the European Medicines Agency: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf : "[this document] cautions against adjusting for 'covariates measured after randomisation because they may be affected by the treatments'" (Introduction, pg 4). For these reasons, it is recommended to adjust for the stratification variables and for strongly predictive covariates, which must be measured prior to randomisation. This is the strategy we adopt.

- Is there going to be an interim analysis?

Response: No interim analysis will be conducted by the trialists. However, the study Data monitoring Committee will have full access to the unblinded data throughout the trial, and will monitor for any concerning developments.

- Please specify the authors' roles.

Response: We have specified the authors' roles in this revised manuscript (Highlighted in this revised submission: Page 15, Lines 360–364.

- Line 307: I believe the phrase about "manufacturers' wishes" should be modified and toned down.

Response: It is now changed and read "recommendations of manufactures"

We would like to thank this reviewer, once again for the recommendations to improve on the manuscript.

VERSION 2 – REVIEW

REVIEWER	Professor Ernest HY NG Department of O & G, LKS Faculty of Medicine, the University of Hong Kong
REVIEW RETURNED	20-Nov-2019

GENERAL COMMENTS	The aim of this protocol paper was to examine if there is an effect on live birth rate using three levels of pH of culture media for IVF. Strength of the paper 1. There is no randomized study comparing different culture media pH on the pregnancy outcomes especially the live birth rate. 2. It is double blind study as the patients and the physicians would not be aware of the randomized arms. 3. There is clear description of the sample size calculation and randomization plan in the text. Major concerns 1. Introduction • The authors should state the adverse effects of extreme culture media pH in both animal and human IVF. 2. Materials and methods • It is not clear if the pH of the culture media would be measured or confirmed on each day or more frequently. • Only one culture media should be used through the study as different media may affect the results especially the change of media will be in some centres only. • Sample size calculation: there is no justification of using 10%. It is not sure if 93% power is chosen. A much drop-out rate of about 20-30% should be anticipated as patients may not have fresh embryo transfer for a number of reasons such as risk of ovarian hyperstimulation syndrome and high serum progesterone level on the day of hCG etc

REVIEWER	Ioannis Sfontouris Eugonia ART Unit, Greece
REVIEW RETURNED	03-Dec-2019

GENERAL COMMENTS	The authors have done a good job addressing my comments. A couple of minor observations from my side: Lines 171-2: Severe PCOS, hyper-responder, OHSS patients, and cycles with agonist trigger or any patient with a plan for a “freeze-all”. • I do not think “severe PCOS” in a universally used term. Please replace with “PCOS” • Please define hyper-responders (eg with >18 follicles on trigger day?) In the current or previous cycle? • OHSS patients: I assume you mean patients who developed OHSS in a previous cycle. In this case, will you exclude only patients with severe OHSS or with moderate OHSS as well? Lines343-6: The authors state that “ In the analysis of number of usable embryos, implanted embryos arising from the day 3 transfer will be included as formed and good quality blastocysts, while those that do not implant in this portion will be considered blocked at day 3." This is an arbitrary calculation of blastocyst formation. I am happy for this assumption to remain as is, provided it is adequately discussed in the discussion and the limitations.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Response: Thank you for your recommendations and we hope to respond on them satisfactory.

Major concerns

1. Introduction

• The authors should state the adverse effects of extreme culture media pH in both animal and human IVF.

Response: We appreciate this comment, in this revised submission, we have discussed the possible harms of using extreme levels of pH (Highlighted in the revised manuscript: Page 4, Lines 92 to 106). The harms of using extreme levels of pH are beyond the scope of this trial as we are trying to identify which is a better level within a clinically proven safe range (7.2 to 7.4). Perhaps further trials can find whether these extreme levels of pH (≤7 or ≥ 7.5) can be really harmful.

2. Materials and methods

- It is not clear if the pH of the culture media would be measured or confirmed on each day or more frequently.

Response: pH levels will be measured twice weekly using blood gas analyser. Once the level of pH is adjusted, it will be ensured with daily measurement of CO2. pH also will be measured every new batch of culture media and adjusted accordingly. To account for personal variations, pH will be measured in all centres using the same calibrated device and the same personnel (Highlighted in the revised manuscript: Page 9, Lines 219–220, Page 10, Lines 231–236).

- Only one culture media should be used through the study as different media may affect the results especially the change of media will be in some centres only.

Response: We agree with the reviewer and confirm that all culture disposables, media and oil will be the same in all centres. If a change occurs, it will be across all centres at the same time, and it will be reported. Since the allocation is stratified by site, any change would be balanced between treatment arms in the study.

- Sample size calculation: there is no justification of using 10%. It is not sure if 93% power is chosen.

Response: Our study has been designed to detect effects that we consider to be realistic and relevant. We have chosen to try to detect a 10-percentage points difference across the three levels of pH (7.2, 7.3, and 7.4). This is realistic range as specifying a wider difference for live birth rate is not practical, although it can help us reducing the sample size. In the contrary, using a narrower range of percentage-points difference (5% for example) will make the sample size huge and beyond the conduct. Therefore, we have powered on the assumption that the difference between the two most discrepant pH groups could be as low as 10 percentage points, with the third pH group differing by about 5 percentage points from either. By simulation, we calculate that the study, with 646 participants per arm, will have very high power to reject the null hypothesis of no effect of pH in the event that the pH effect is as strong or stronger than this (power ~ 99% at a 5% significance level). Note that it makes sense to aim for a high-power value here, reflecting our relative uncertainty in relation to realistic effect size. This ensures that we maintain relatively good power in the event that the spread of birth rates is lower than anticipated. For example, if the birth rates are 26%, 30%, 33% (a spread of just seven percentage points) this sample size yields 86% power against a 5% significance level, and 66% at a 1% significance level. We have also been conservative in our inflation of numbers for drop out (see comments re: dropout below). This has been partly informed by a recent review of estimated effect sizes and sample sizes in fertility RCTs undertaken by the trial statistician (JW): Stocking, et al., 2019: https://doi.org/10.1093/humrep/dez017 . In light of suggestions made by reviewer 2, we have amended our primary analysis slightly, and this has caused minor changes to the estimated power of the design.

The study will be amongst the largest conducted in IVF (see Stocking, et al., 2019, results, para 1 – note these figures relate to a sample of trials that were the largest conducted for each intervention). A

minority of RCTs in this field are well-powered to detect improvements in live birth as large as 20 percentage points. The mentioned could justify why 93% power was used as it resulted from the power simulation.

We have sent the R code used for power simulation with this response for consideration by the reviewers. The sample size paragraph has been edited in the manuscript.

A much drop-out rate of about 20-30% should be anticipated as patients may not have fresh embryo transfer for a number of reasons such as risk of ovarian hyperstimulation syndrome and high serum progesterone level on the day of hCG etc

Response:

The reviewer appears to assume here that participants who have been randomised, but who do not proceed to fresh transfer, would be excluded from the analysis. This would violate the intention to treat principle, and would be fatal to our ability to make a randomised inference from the trial. All women randomised will included in the analysis according to the groups they were allocated to. This is the intention to treat principle. As such, women who do not proceed to have fresh transfer are included in the analysis, and are recorded as not having a live birth. The live birth rates used in the power calculation are based on actual practice, and incorporate failure to proceed to fresh transfer.

We have allowed for a 'dropout' rate of 5%, to allow for the possibility that a small number of women might withdraw their consent for their data to be used. Barring this, all randomised women will be incorporated in the analysis, and so we actually anticipate having more than 646 women per arm available for analysis.

We would like to thank this reviewer, once again for the recommendations to improve on the manuscript.

Reviewer: 3

Response: Thank you for your comments and we hope to respond on your recommendations satisfactory.

Lines 171-2: Severe PCOS, hyper-responder, OHSS patients, and cycles with agonist trigger or any patient with a plan for a "freeze-all".

- I do not think "severe PCOS" in a universally used term. Please replace with "PCOS"

Response: We amended it and it is now read “PCOS”

• Please define hyper-responders (eg with >18 follicles on trigger day?) In the current or previous cycle?

We have removed this vague term “Hyper responder” from the exclusion criteria as OHSS is enough.

• OHSS patients: I assume you mean patients who developed OHSS in a previous cycle. In this case, will you exclude only patients with severe OHSS or with moderate OHSS as well?

Response: We amended the manuscript to include the term “women with history of severe OHSS” instead of “OHSS” as patients who will develop OHSS after randomization will be included in the analysis as negative outcome. This is the concept of intention-to-treat we will use.

Lines343-6: The authors state that “ In the analysis of number of usable embryos, implanted embryos arising from the day 3 transfer will be included as formed and good quality blastocysts, while those that do not implant in this portion will be considered blocked at day 3."

This is an arbitrary calculation of blastocyst formation. I am happy for this assumption to remain as is, provided it is adequately discussed in the discussion and the limitations.

Response: We included one sentence regarding this limitation in the limitation section and we added this assumption to the discussion (Highlighted: Page 3, line 61, and Page 15, Lines 361 to 367)

VERSION 3 - REVIEW

REVIEWER	Professor Ernest HY NG Department of O&G, The University of Hong Kong, HKSAR
REVIEW RETURNED	12-Dec-2019

GENERAL COMMENTS	The drop out rate is too low. If a higher drop out rate is encountered, the sample size may not be enough to show the proposed difference. The sample size should be inflated by at least 20% based on previous experience.

REVIEWER	Ioannis Sfontouris Eugonia IVF Unit, Greece
REVIEW RETURNED	06-Dec-2019

GENERAL COMMENTS	I am happy for the manuscript to be published.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 3

Reviewer Name: Ioannis Sfontouris

Institution and Country: Eugonia IVF Unit, Greece

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

I am happy for the manuscript to be published.

Response: Thank you for the efforts made by this reviewer to improve on this protocol.

Reviewer: 1

Reviewer Name: Professor Ernest HY NG

Institution and Country: Department of O&G, The University of Hong Kong, HKSAR

Please state any competing interests or state ‘None declared’: Nil

Please leave your comments for the authors below

The dropout rate is too low. If a higher drop-out rate is encountered, the sample size may not be enough to show the proposed difference. The sample size should be inflated by at least 20% based on previous experience.

Response: We do not plan to consider deviation of protocol by freeze-all or any other medical indication as dropout. We assume that in a proper intention to treat analysis, any protocol deviation or unreached participants should be considered as negative results and included in the analyses. The dropout rate we have considered in this trial is for participants who will withdraw their consent to participation. We believe that this will not exceed 5% as written in the protocol. We screened the facilities that will conduct this trial for their dropout rate, and they respond that it is 5% on average in their database. Also, although the worst scenario of 20% dropout appears rare in our facilities, we can assure the reviewer that 680 participants per arm is still robust to give > 85% power at 1% alpha level, which is a very valid power for the trial’s results. Attached is R code of power calculation that we used to check for all of these scenarios. The sample size and power of this trial are one of the highest in IVF trials.