OBSTETRICS & GYNECOLOGY

NOTICE: This document contains correspondence generated during peer review and subsequent revisions but before transmittal to production for composition and copyediting:

- Comments from the reviewers and editors (email to author requesting revisions)
- Response from the author (cover letter submitted with revised manuscript)*

**The corresponding author has opted to make this information publicly available.*

Personal or nonessential information may be redacted at the editor's discretion.

Questions about these materials may be directed to the *Obstetrics & Gynecology* editorial office: obgyn@greenjournal.org.

Date: Nov 02, 2021
To: "Esther H Chung" [redacted]
From: "The Green Journal" em@greenjournal.org
Subject: Your Submission ONG-21-1959

RE: Manuscript Number ONG-21-1959

Virtual Versus In-Clinic Transvaginal Ultrasound for Ovarian Reserve Assessment: A Non-inferiority Trial

Dear Dr. Chung:

Your manuscript has been reviewed by the Editorial Board and by special expert referees. Although it is judged not acceptable for publication in Obstetrics & Gynecology in its present form, we would be willing to give further consideration to a revised version.

If you wish to consider revising your manuscript, you will first need to study carefully the enclosed reports submitted by the referees and editors. Each point raised requires a response, by either revising your manuscript or making a clear and convincing argument as to why no revision is needed. To facilitate our review, we prefer that the cover letter include the comments made by the reviewers and the editor followed by your response. The revised manuscript should indicate the position of all changes made. We suggest that you use the "track changes" feature in your word processing software to do so (rather than strikethrough or underline formatting).

Please be sure to address the Editor comments (see "EDITOR COMMENTS" below) in your point-by-point response.

Your paper will be maintained in active status for 21 days from the date of this letter. If we have not heard from you by Nov 23, 2021, we will assume you wish to withdraw the manuscript from further consideration.

REVIEWER COMMENTS:

Reviewer #1:

This was a non-randomized, non-inferiority trial in which all patients has both an at-home TVUS guided by a remote ultrasound technologist, and then following this has an in-clinic TVUS performed by an ultrasound technologist.

Strengths:
* It is certainly a very interesting idea that patients can do their own vaginal ultrasounds at home, and minimize the sense of invasiveness and cost associated with clinical ultrasounds.
* Patients served as their own controls, allowing there to be limited recruitment and selection bias between the two interventions.
* 2-3 blinded raters assessed each cine clip, so they did not know if the images were taken by patient guidance or in the clinic and minimized assessor bias.

Limitations:
* The dichotomized outcome of "clinical quality" of images (yes or no) is highly subjective and may not correlate to clinically meaningful results. The clinically impactful outcomes (follicle counts, number of submucosal fibroids) could not or were not powered to detect non-inferiority.
* The primary outcome was changed mid-trial, limiting the veracity of the results as the sample size and planning of the outcome was not truly done a priori.
* The study is funded by a company with a vested interest in one of the interventions (at-home imaging), and this company was highly involved with study design and procedures.

Comments for authors by section:

Abstract:
* Line 24: Equivalence is not the same as saying there is no significant difference. You can say "similar" or "non-significant difference", but cannot utilize the word "equivalent" unless you have set an equivalence margin and powered for this outcome, which is not the case.
* Line 26-27: How superior was the NPS? The readers should know from the abstract how much better patient satisfaction was.

Introduction:

* Line 58-59: There should be a clear hypothesis stated that indicates the main outcome or outcomes that are being assessed. "Non-inferior" does not tell the reader what is being assessed as the measure of "inferiority" or "non-inferiority".

Methods:
* Line 18-185 and Line 203: If the authors were going to dichotomize into "clinical quality" or not, why did they have them rate on a scale with criteria? When a scale is being utilized, there should be some justification for why this was then dichotomized further.
* Line 218-221: The authors must present the power they wanted to achieve and the sample size needed to do so, in addition to presenting the non-inferiority margin and the alpha statistic.
* Line 218-221: Why was an alpha of 2.5% chosen, as opposed to 5%? I was merely curious, and I imagine readers may be as well, as this is unusual.
* Line 230-233: If multiple organs from each subject contributed to the primary outcome, were they counted separately as endpoints? Or did each contribute to a "clinical quality" stamp for that one patient TVUS event? This should be made more clear, as it is hard to determine if appropriate statistical analyses were used if the clinical quality of one organ was analyzed as being independent from the clinical quality of another organ on the same patient (which would not be appropriate).
* Line 263-275: I have never before read a study utilizing the NPS, so this is interesting. I would like to see more explanation of what difference between the NPS of one intervention (in-clinic TVUS) versus another intervention (home TVUS, in this case) is considered clinically meaningful. This would help put the results presented later into context.

Results:
* Line 294: Please explain better in the Methods section what was being timed for the duration of the exam. Was this from the time the transducer went in the vagina to when it went out? Or the cumulative time of all cine clips from the patient? This should be more clearly defined for in-office and at-home exams.
* Line 295-295: Please give a p-value for this difference or a mean difference with confidence intervals.
* Line 310-314: Give the actual number of women with high BMI here, as opposed to just saying "34% of the population", and give the confidence intervals around the percent difference or the p-value for the 93% vs. 98%.
* Line 350: The definition and explanation of "nonreplicable findings" is not clearly explained in the methods, although it is mentioned in the abstract and the Results. This should have been more clearly delineated in the Methods section.
* Line 334: Again, you should give a p-value for the difference between NPS between the two interventions and state was that p-value reflects (difference in percent that are promoters? Difference in score?)
* Table 2 gives no comparative statistics between at-home and in-clinic TVUS groups, although it gives values of outcomes for each group.

Discussion:
* Line 368: Were these populations that were getting fertility workup, or other types of ultrasound assessments? There should be a bit more discussion about what these past studies left unanswered, and therefore how this fits in the literature.
* Line 386-394: It should also be mentioned that the study was underpowered to assess important clinical outcomes, or power for these outcomes was not assessed.

Reviewer #2:

Abstract
* Line 21: What was the population of women who were recruited for this study?
* Lines 33-35: Patients have to come in for a sono or HSG anyway and it only takes an additional minute to perform an AFC at the time of HCG or sono, so does at-home TVUS for AFC really save time and money?

Introduction
* Lines 42-44: "…limited public awareness of age-related fertility decline." Followed by: "interest in at-home fertility testing has emerged as a cost-considerate, easy-to-access initial option." These statements seem contradictory. If people are not aware of age-related fertility decline, how are they going to know about at home TVUS for AFC?

Role of the Funding Source
* Lines 64-54: Turtle Health's involvement in the study's design, data analysis, and reporting introduces bias. How was this accounted for?

Methods
* Lines 96-97: Did you try to recruit patients with known submucoal fibroids? Did this introduce bias? Were other patients not approached to participate in the study because they did not have known submucosal fibroids?
* Lines 125-126: "…otherwise in the professional judgement of the Primary Investigator (PI) able to give informed consent." What does this mean? The PI decided who was or was not able to give informed consent and thus participate in

the study? Did this introduce bias?
* Line 139: 64 patients were consented by only 56 ultimately participated in the study. What happened to the other 8 participants?
* Line 142: If subjects had oligomenorrhea or amenorrhea were visits still scheduled on consecutive days, or just whenever the subject had availability? How did this affect your results if the visits were not on consecutive days for this group of patients?
* Line 162: How often were there discrepant ratings on clinical quality requiring a tie-breaker?
* Line 164: How much were participants compensated for their participation?
* Lines 177-179: This is confusing. Was a rater given all the cine clips from one patient for each location to view back to back? Ie, did the rater view, for one patient, all the cine clips from the at home TVUS immediately followed by all the cine clips from the office TVUS? Or did the rater view one cine clip from one patient then another random cine clip from another patient?
* Lines 186-188: This is confusing. Perhaps the grammar makes this sentence difficult to follow. What are you trying to describe?
* Lines 193-194: At this point, it seems clear that each cine has its own folder. However, this contradicts with lines 173-175 as well as line 198. Please clarify what was actually in each folder. Is there an ovary cine in each folder? Or do some folders only have uterus cines.
* Lines 224-228: Confusing, run on sentence, please reword
* Line 229: Do you mean "rating" rather than "rate"?
* Lines 233-235: How often did this happen? Was it with the at home or clinical scan more often?
* Line 236: Did you look at time to image for at home vs clinical scans?
* Line 246: What do you mean by fractional?
* Lines 252-257: Please make discussion of the null hypothesis more concise
* Line 268-269: "in the its ability" please fix grammar
* Lines 272-273: What about patients who did not have previous in-clinic TVUS experience? How was their view of at home TVUS experience evaluated if they didn't have anything to compare it to?
* Overall: Please shorten the methods section and make it more concise

Results
* Line 283: What are the standard deviations or ranges for age and BMI?
* Line 287: "higher than the original 25% target for the trial". What does this mean? You were targeting patients with a BMI around 25? Does this introduce bias?
* Line 291: OCPs stands for oral contraceptive pills
* Line 338: What do the numbers "10" and "26" correspond to?
* Line 348-349: While this is technically true because there were no false positives, there were no true positives either. This statement is a bit of a stretch.

Discussion
* Line 375: "The NPS for…" This is confusing. Please specify that this is the NPS for at-home TVUS
* Lines 376-377: "likely representing patients who would hesitate to seek care in-person." This is speculation. Are patients really not going to seek care in person? Unfortunately, infertility treatment involves in-person care, in-person TVUS, in-person blood work, and in-person procedures. While consults can be done remotely, it would be impossible to treat a patient for infertility without seeing them in person.
* Line 380: "characteristics" in pleural. What besides BMI was physically diverse?
* Line 381: "persons" Persons or females? Sex assigned at birth females were included, not men, correct?
* Lines 388-389: Specificity was not actually assessed in your study so it is hard to say the specificity is 100%.
* Line 405: Write what "FP" stands for
* Line 406: Write what "RE" stands for

Figures
* Table 2: Remove line regarding specificity as this was not actually assessed

Reviewer #3:

This study examined the use of virtual transvaginal ultrasound performed by patient compared to in clinic transvaginal ultrasound performed by the UT with the primary outcome of image quality and AFC. This study is well designed to understand the acceptability of this innovative approach to the patient and physician.

Methods
Page 8 - the authors should provide more information on the vaginal probe including MHz, dimensions etc. Were any modification made to the probe to improve ease of use for the patient?
Page 9 - exclusion criteria - I am surprised that vaginismus was not an exclusion criterion. Also, absence of an ovary

should have been an exclusion criterion when each organ is being evaluated in the analysis. 8 subjects did not participate after consenting and the authors note that this was for personal reasons. Please elaborate these reasons.
Page 9 - what instructions were provided to the patients on how to perform the ultrasound? This can be added as a supplement. On what day of the cycle were the scans performed? What materials were supplied in the package i.e. gel, probe cover etc What training was provided to the UT to guide the patients and acquire optimal images. Did any patient have a retroverted uterus? Were any ovaries difficult to visualize due to bowel gas?
Page 10 - US in the office was performed using the same probe. Was it attached to a regular office US machine or were the images sent wirelessly?

STATISTICS EDITOR COMMENTS:

Abstract: Need to format the primary outcome and its evaluation in Results in terms of non-inferiority. That is, the reader cannot tell from the Abstract what non-inferiority difference was posited and therefore how that difference related to the difference in clinical quality of the two cohorts. The conclusion needs a re-write: the non-inferiority test was only regarding the quality of the U/S images, not specifically for assessing a metric of ovarian reserve.

lines 218-222 and reference # 18: The reference cited does suggest using alpha = 0.025 for inference testing, but does not stipulate or even mention 18% as the margin, so the statement is incomplete. What was the basis for asserting that a difference of as much as 18% in clinical quality would be acceptable? Seems that the margin is wide, even by the Authors' admission on lines 224-228, hence the study would require larger samples. There is a potential downside to having a higher proportion of unacceptable quality studies.

Table 1: Need units for age.

Table 2: Need to clearly separate the primary from the secondary outcomes. Only the primary was factored into the sample size calculation, the secondaries were not. Regardless of whether the secondary outcomes were statistically significant, they are of interest, but second tier compared to the primary outcome. It would also be preferable to format the primary outcome in the usual graphical display of a non-inferiority study, so that the reader can easily see how the difference between the cohorts compares with the non-inferiority margin.

EDITORIAL OFFICE COMMENTS:

1. The Editors of Obstetrics & Gynecology have increased transparency around its peer-review process, in line with efforts to do so in international biomedical peer review publishing. If your article is accepted, we will be posting this revision letter as supplemental digital content to the published article online. Additionally, unless you choose to opt out, we will also be including your point-by-point response to the revision letter. If you opt out of including your response, only the revision letter will be posted. Please reply to this letter with one of two responses:

A. OPT-IN: Yes, please publish my point-by-point response letter.
B. OPT-OUT: No, please do not publish my point-by-point response letter.

2. When you submit your revised manuscript, please make the following edits to ensure your submission contains the required information that was previously omitted for the initial double-blind peer review:
* Include your title page information in the main manuscript file. The title page should appear as the first page of the document. Add any previously omitted Acknowledgements (ie, meeting presentations, preprint DOIs, assistance from non-byline authors).
* Funding information (ie, grant numbers or industry support statements) should be disclosed on the title page and in the body text. For industry-sponsored studies, the Role of the Funding Source section should be included in the body text of the manuscript.
* Include clinical trial registration numbers, PROSPERO registration numbers, or URLs at the end of the abstract (if applicable).

* Name the IRB or Ethics Committee institution in the Methods section (if applicable).
* Add any information about the specific location of the study (ie, city, state, or country), if necessary for context.

3. Obstetrics & Gynecology uses an "electronic Copyright Transfer Agreement" (eCTA), which must be completed by all authors. When you uploaded your manuscript, each co-author received an email with the subject, "Please verify your authorship for a submission to Obstetrics & Gynecology." Please check with your coauthors to confirm that they received and completed this form, and that the disclosures listed in their eCTA are included on the manuscript's title page.

4. For studies that report on the topic of race or include it as a variable, authors must provide an explanation in the manuscript of who classified individuals' race, ethnicity, or both, the classifications used, and whether the options were defined by the investigator or the participant. In addition, the reasons that race/ethnicity were assessed in the study also should be described (eg, in the Methods section and/or in table footnotes). Race/ethnicity must have been collected in a formal or validated way. If it was not, it should be omitted. Authors must enumerate all missing data regarding race and ethnicity as in some cases, missing data may comprise a high enough proportion that it compromises statistical precision and bias of analyses by race.

Use "Black" and "White" (capitalized) when used to refer to racial categories. The nonspecific category of "Other" is a convenience grouping/label that should be avoided, unless it was a prespecified formal category in a database or research instrument. If you use "Other" in your study, please add detail to the manuscript to describe which patients were included in that category.

5. Clinical trials submitted to the journal as of July 1, 2018, must include a data sharing statement. The statement should indicate 1) whether individual deidentified participant data (including data dictionaries) will be shared; 2) what data in particular will be shared; 3) whether additional, related documents will be available (eg, study protocol, statistical analysis plan, etc.); 4) when the data will become available and for how long; and 5) by what access criteria data will be shared (including with whom, for what types of analyses, and by what mechanism). Responses to the five bullet points should be provided in a box at the end of the article (after the References section).

6. Standard obstetric and gynecology data definitions have been developed through the reVITALize initiative, which was convened by the American College of Obstetricians and Gynecologists and the members of the Women's Health Registry Alliance. Obstetrics & Gynecology has adopted the use of the reVITALize definitions. Please access the obstetric data definitions at https://www.acog.org/practice-management/health-it-and-clinical-informatics/revitalize-obstetrics-data-definitions and the gynecology data definitions at https://www.acog.org/practice-management/health-it-and-clinical-informatics/revitalize-gynecology-data-definitions. If use of the reVITALize definitions is problematic, please discuss this in your point-by-point response to this letter.

7. Because of space limitations, it is important that your revised manuscript adhere to the following length restrictions by manuscript type: Original Research should not exceed 5,500 words. Stated word limits include the title page, précis, abstract, text, tables, boxes, and figure legends, but exclude references.

8. Specific rules govern the use of acknowledgments in the journal. Please note the following guidelines:

* All financial support of the study must be acknowledged.
* Any and all manuscript preparation assistance, including but not limited to topic development, data collection, analysis, writing, or editorial assistance, must be disclosed in the acknowledgments. Such acknowledgments must identify the entities that provided and paid for this assistance, whether directly or indirectly.
* All persons who contributed to the work reported in the manuscript, but not sufficiently to be authors, must be acknowledged. Written permission must be obtained from all individuals named in the acknowledgments, as readers may infer their endorsement of the data and conclusions. Please note that your response in the journal's electronic author form

verifies that permission has been obtained from all named persons.
* If all or part of the paper was presented at the Annual Clinical and Scientific Meeting of the American College of Obstetricians and Gynecologists or at any other organizational meeting, that presentation should be noted (include the exact dates and location of the meeting).
* If your manuscript was uploaded to a preprint server prior to submitting your manuscript to Obstetrics & Gynecology, add the following statement to your title page: "Before submission to Obstetrics & Gynecology, this article was posted to a preprint server at: [URL]."

9. The most common deficiency in revised manuscripts involves the abstract. Be sure there are no inconsistencies between the Abstract and the manuscript, and that the Abstract has a clear conclusion statement based on the results found in the paper. Make sure that the abstract does not contain information that does not appear in the body text. If you submit a revision, please check the abstract carefully.

In addition, the abstract length should follow journal guidelines. The word limit for Original Research articles is 300 words; Reviews is 300 words; Case Reports is 125 words; Current Commentary articles is 250 words; Executive Summaries, Consensus Statements, and Guidelines are 250 words; Clinical Practice and Quality is 300 words; Procedures and Instruments is 200 words. Please provide a word count.

10. Abstracts for all randomized, controlled trials should be structured according to the journal's standard format. The Methods section should include the primary outcome and sample size justification. The Results section should begin with the dates of enrollment to the study, a description of demographics, and the primary outcome analysis. Please review the sample abstract that is located online here: http://edmgr.ovid.com/ong/accounts/sampleabstract_RCT.pdf. Please edit your abstract as needed.

11. Only standard abbreviations and acronyms are allowed. A selected list is available online at http://edmgr.ovid.com /ong/accounts/abbreviations.pdf. Abbreviations and acronyms cannot be used in the title or précis. Abbreviations and acronyms must be spelled out the first time they are used in the abstract and again in the body of the manuscript.

12. The journal does not use the virgule symbol (/) in sentences with words. Please rephrase your text to avoid using "and/or," or similar constructions throughout the text. You may retain this symbol if you are using it to express data or a measurement.

13. ACOG avoids using "provider." Please replace "provider" throughout your paper with either a specific term that defines the group to which are referring (for example, "physicians," "nurses," etc.), or use "health care professional" if a specific term is not applicable.

14. In your Abstract, manuscript Results sections, and tables, the preferred citation should be in terms of an effect size, such as odds ratio or relative risk or the mean difference of a variable between two groups, expressed with appropriate confidence intervals. When such syntax is used, the P value has only secondary importance and often can be omitted or noted as footnotes in a Table format. Putting the results in the form of an effect size makes the result of the statistical test more clinically relevant and gives better context than citing P values alone.

If appropriate, please include number needed to treat for benefits (NNTb) or harm (NNTh). When comparing two procedures, please express the outcome of the comparison in U.S. dollar amounts.

Please standardize the presentation of your data throughout the manuscript submission. For P values, do not exceed three decimal places (for example, "P = .001"). For percentages, do not exceed one decimal place (for example, 11.1%").

15. Your manuscript contains a priority claim. We discourage claims of first reports since they are often difficult to prove. How do you know this is the first report? If this is based on a systematic search of the literature, that search should be described in the text (search engine, search terms, date range of search, and languages encompassed by the search). If it is not based on a systematic search but only on your level of awareness, it is not a claim we permit.

16. Please review the journal's Table Checklist to make sure that your tables conform to journal style. The Table Checklist is available online here: http://edmgr.ovid.com/ong/accounts/table_checklist.pdf.

17. Please review examples of our current reference style at http://ong.editorialmanager.com (click on the Home button in the Menu bar and then "Reference Formatting Instructions" document under "Files and Resources). Include the digital object identifier (DOI) with any journal article references and an accessed date with website references. Unpublished data, in-press items, personal communications, letters to the editor, theses, package inserts, submissions, meeting presentations, and abstracts may be included in the text but not in the reference list.

In addition, the American College of Obstetricians and Gynecologists' (ACOG) documents are frequently updated. These documents may be withdrawn and replaced with newer, revised versions. If you cite ACOG documents in your manuscript, be sure the references you are citing are still current and available. Check the Clinical Guidance page at https://www.acog.org/clinical (click on "Clinical Guidance" at the top). If the reference is still available on the site and isn't listed as "Withdrawn," it's still a current document.

If the reference you are citing has been updated and replaced by a newer version, please ensure that the new version supports whatever statement you are making in your manuscript and then update your reference list accordingly (exceptions could include manuscripts that address items of historical interest). If the reference you are citing has been withdrawn with no clear replacement, please contact the editorial office for assistance (obgyn@greenjournal.org). In most cases, if an ACOG document has been withdrawn, it should not be referenced in your manuscript.

18. When you submit your revision, art saved in a digital format should accompany it. If your figure was created in Microsoft Word, Microsoft Excel, or Microsoft PowerPoint formats, please submit your original source file. Image files should not be copied and pasted into Microsoft Word or Microsoft PowerPoint.

When you submit your revision, art saved in a digital format should accompany it. Please upload each figure as a separate file to Editorial Manager (do not embed the figure in your manuscript file).

If the figures were created using a statistical program (eg, STATA, SPSS, SAS), please submit PDF or EPS files generated directly from the statistical program.

Figures should be saved as high-resolution TIFF files. The minimum requirements for resolution are 300 dpi for color or black and white photographs, and 600 dpi for images containing a photograph with text labeling or thin lines.

Art that is low resolution, digitized, adapted from slides, or downloaded from the Internet may not reproduce.

19. Authors whose manuscripts have been accepted for publication have the option to pay an article processing charge and publish open access. With this choice, articles are made freely available online immediately upon publication. An information sheet is available at http://links.lww.com/LWW-ES/A48. The cost for publishing an article as open access can be found at https://wkauthorservices.editage.com/open-access/hybrid.html.

If your article is accepted, you will receive an email from the editorial office asking you to choose a publication route (traditional or open access). Please keep an eye out for that future email and be sure to respond to it promptly.

If you choose open access, you will receive an Open Access Publication Charge letter from the Journal's Publisher, Wolters Kluwer, and instructions on how to submit any open access charges. The email will be from

publicationservices@copyright.com with the subject line, "Please Submit Your Open Access Article Publication Charge(s)." Please complete payment of the Open Access charges within 48 hours of receipt.

***

If you choose to revise your manuscript, please submit your revision through Editorial Manager at http://ong.editorialmanager.com. Your manuscript should be uploaded as a Microsoft Word document. Your revision's cover letter should include the following:

* A confirmation that you have read the Instructions for Authors (http://edmgr.ovid.com/ong/accounts/authors.pdf), and
* A point-by-point response to each of the received comments in this letter. Do not omit your responses to the Editorial Office or Editors' comments.

If you submit a revision, we will assume that it has been developed in consultation with your co-authors and that each author has given approval to the final form of the revision.

Again, your paper will be maintained in active status for 21 days from the date of this letter. If we have not heard from you by Nov 23, 2021, we will assume you wish to withdraw the manuscript from further consideration.

Sincerely,

John O. Schorge, MD
Associate Editor, Gynecology

2020 IMPACT FACTOR: 7.661
2020 IMPACT FACTOR RANKING: 3rd out of 83 ob/gyn journals

_____

Oct 23, 2021

The Editor-in-Chief
Dr. Dwight J. Rouse
c/o *Obstetrics & Gynecology*

Dear Dr. Rouse and the Editorial Staff,

We would like to thank the reviewers for their insightful comments and suggestions for our manuscript entitled, "**Virtual Versus In-Clinic Transvaginal Ultrasound for Ovarian Reserve Assessment: A Non-inferiority Trial"**. We have reviewed the comments of the reviewers and have incorporated changes into the manuscript as suggested. We have also included a point by-point response to the reviewers' comments and suggestions below with references to changes in the manuscript where appropriate. Line references refer to the CLEAN (non-tracked changes) version of the manuscript. I would like to state that all authors concur with the edits of this revised resubmission. I would also like to verify that the Instructions for Authors have been reviewed.

We would like to express our gratitude for your continued consideration of this manuscript for publication in *Obstetrics and Gynecology*. We look forward to your feedback. If there are any questions or concerns, please feel free to contact us.

Warm regards,

Esther H. Chung, MD
Duke University Medical Center
Department of Obstetrics & Gynecology
Durham, NC, 27705

Response to Reviewers:

Reviewer #1:

This was a non-randomized, non-inferiority trial in which all patients has both an at-home TVUS guided by a remote ultrasound technologist, and then following this has an in-clinic TVUS performed by an ultrasound technologist.

Strengths:
***It is certainly a very interesting idea that patients can do their own vaginal ultrasounds at home, and minimize the sense of invasiveness and cost associated with clinical ultrasounds.**

Response: Thank you for this comment. We completely agree with this insight that the at-home ultrasound has the potential to minimize the invasiveness and cost associated with in-clinic ultrasounds.

***Patients served as their own controls, allowing there to be limited recruitment and selection bias between the two interventions.**

Response: We appreciate this thoughtful comment and agree that this was the benefit of having patients serve as their own controls.

***2-3 blinded raters assessed each cine clip, so they did not know if the images were taken by patient guidance or in the clinic and minimized assessor bias.**

Response: Thank you for this comment, and we agree that the study allowed for randomized assessment of the images taken at home and in-clinic by the independent raters.

Limitations:
***The dichotomized outcome of "clinical quality" of images (yes or no) is highly subjective and may not correlate to clinically meaningful results.**

Response: Thank you for this insightful feedback. The subjectivity of interpretation of TVUS images is an important concern. We had performed an extensive literature search that unfortunately did not yield a clear consensus TVUS evaluation scale or rubric. We agree that the lack of a clear consensus scale for image quality is a general limitation for research in this area. We have added statements to this effect in the limitations section in the Discussion (see lines 408-411):

> The lack of previously validated standards for clinical quality TVUS imaging is a general limitation for research in this area. We hope that the endpoint used in this trial—which was found to be intuitive and representative of clinical practice by the blinded raters—may inform future studies.

Given this limitation, we designed an image quality endpoint that attempted to faithfully replicate the actual clinical decision process used to evaluate real-world ultrasound images: an initial subjective assessment of suitability for interpretation, followed by the conduct of specific diagnostic tasks (e.g. AFC estimation). Interviews with REIs and generalist OB/GYNs responsible for ultrasound instruction mapped that process to the four-point scale used in the trial, which is similar to scales used in other imaging studies. In practice, independent raters found the scale representative of their clinical practice, and for that reason, likely to correlate to clinically meaningful results. We have added a statement to this effect to the Methods section (see Lines 207-212):

> Given lack of an existing consensus for TVUS image quality criteria[26–28], a four-point scale was developed from interviews with expert physicians who regularly assess images (Supplementary Table 1). During scale development, emphasis was placed on replicating and codifying the thought process from usual clinical practice and ensuring all clips could be intuitively categorized by an expert physician. The scales for imaging quality assessment used by Hausleiter et al were used as a guide for quality rating[29].

From the original protocol (for reference): "There does not appear to be a single standard rating scale for ultrasound quality. Interviews with healthcare professionals yielded a 4 point scale, which has precedence in relevant literature. See, for instance, Hausleiter et al (2010). Image quality and radiation exposure with a low tube voltage protocol for coronary CT angiography results of the PROTECTION II Trial."

***The clinically impactful outcomes (follicle counts, number of submucosal fibroids) could not or were not powered to detect non-inferiority.**

Response: Thank you for your feedback here. Sample size was determined by powering the primary endpoint, though secondary endpoints were expected to be powered under a range of reasonable assumptions. The AFC equivalence hypothesis was also 90% powered for a mean AFC per ovary of eight and a standard deviation of AFC per ovary as high as seven, which are conservative assumptions based on published AFC nomograms. Particularly because of the novelty of the imaging method, and due to some uncertainty that existed prior to the trial about how intra-cycle, inter-modality, and intra-rater variability in AFC estimation might impact the variation of individual estimates, such a conservative estimate appeared reasonable. In practice, the study was >95% powered to detect AFC equivalence given the observed variation in-sample and the 2.65 equivalence margin. Further clarification is provided in our Statistical Methods section (lines 242-247)

> Sample size was selected to ensure >90% power to demonstrate non-inferiority of the clinical quality image rate, given a non-inferiority margin of 18%, a true margin of up to 7%, a 95% clinical quality image rate in-clinic, an adjustment for up to a 60% intra-patient correlation between organ images, and an alpha of 2.5%. This alpha was chosen based on convention for one-sided tests, utilizing FDA guidance19. AFC equivalency was well-powered (>90%) under reasonable assumptions.

The challenges recruiting submucosal subjects were indeed a limitation of the trial, despite the best efforts of an experienced recruitment team, which included both an experienced trialist PI, as well as the former head of digital recruitment from Sanofi. This is further clarified throughout the manuscript (see Lines 114-116 and 361-362).

> Despite three months of recruitment, no patients with known submucosal fibroids were identified. As a result, the fibroid study arm was eliminated.
>
> No patient was identified as a having a submucosal fibroid, despite the raters being prompted on their survey.

***The primary outcome was changed mid-trial, limiting the veracity of the results as the sample size and planning of the outcome was not truly done a priori.**

Thank you for this insightful comment. We completely agree with caution in interpreting ex-post amendments. In this case, the timing of the amendment was triggered by late-breaking feedback from FDA, which was out of the control of the Sponsor (the amendment to modify submucosal targets had been prepared, which was then held to additionally add on a statistical test of organ independence to the primary endpoint at the FDA's request, leading to the amendment being submitted later than otherwise planned). It is important to note, however, that despite this delay, no data was analyzed prior to the amendment, so the planning of the outcome remained *a priori*.

As mitigation for this concern and in response to this feedback, the primary endpoint was re-analyzed as specified in the original protocol (no intra-user dependence adjustment, only assessing the first 30 participants). Doing so obtains a (naive) 97.5% CI of the difference in clinical quality imaging rate between settings of (-0.071, 0.027), essentially the same as the (appropriately adjusted) one obtained in the trial: (-0.064, 0.015). This suggests that the amendment served its intended purpose in maintaining study power while reducing the risk of biased confidence intervals. The amendment eliminated the submucosal fibroid arm of the trial due to inability to recruit patients with known pathology, despite best efforts of an experienced recruitment team. As noted in other comments, we agree the inability to recruit patients with submucosal fibroids is a limitation of the trial. Submucosal fibroid detection was, however, a secondary endpoint, and specificity remained powered. We have refrained from commenting about FDA's delay in sending feedback in the manuscript, but the Sponsor is able to provide any necessary documentation to *Obstetrics and Gynecology*. We have, however, clarified further in the text that the planning of the outcome was done a priori (see Lines: 121-122)

> This amendment was submitted prior to the review of any cine clips by the independent raters, and thus the planning of analysis remained *a priori*.

***The study is funded by a company with a vested interest in one of the interventions (at-home imaging), and this company was highly involved with study design and procedures.**

Response: We appreciate that this was a funded industry study, and all the authors recognize this is a notable limitation and introduces potential risks. We have made several mitigations: similar to other prior sponsored studies that have been published in this journal (e.g., Burke *et al* 2019), all conflicts were fully disclosed as requested by the Journal. None of the authors received equity compensation. In addition, the FDA has independently reviewed the underlying source data (cine clips). The reviewers and editors at *Obstetrics and Gynecology* are fully welcome to review any or all cine clips as desired, for further independent verification.

Comments for authors by section:

Abstract:
***Line 24: Equivalence is not the same as saying there is no significant difference. You can say**

"similar" or "non-significant difference", but cannot utilize the word "equivalent" unless you have set an equivalence margin and powered for this outcome, which is not the case.

Response: Thank you for this comment. We completely agree with the provided definition of equivalence and regret our initial submission was not clear on this point. We use the word equivalence to mean the 95% CI of the difference falling entirely within the prespecified equivalence margin of 2.65 follicles. As mentioned, the AFC equivalency endpoint was >90% powered. See clarified modification (lines 41-42):

> AFC was equivalent across settings, with a difference in follicles, 0.23 (95% CI [-0.36, 0.82] within the equivalence margin (2.65).

***Line 26-27: How superior was the NPS? The readers should know from the abstract how much better patient satisfaction was.**

Response: Thank you for this feedback. We agree with this comment and thus have added the difference in NPS and 97.5% CI of the difference to the abstract (line 42-44):

> Virtual exams had superior NPS (58.1 points, 97.5% CI of difference [37.3,79.0]), indicating greater satisfaction with the virtual experience.

Introduction:
***Line 58-59: There should be a clear hypothesis stated that indicates the main outcome or outcomes that are being assessed. "Non-inferior" does not tell the reader what is being assessed as the measure of "inferiority" or "non-inferiority".**

Response: Thank you for this helpful feedback. We have clarified the primary hypothesis in the introduction as recommended (Lines 72-76):

> Our study evaluated the ability of the virtual TVUS to consistently produce clinical quality images sufficient to evaluate ovarian reserve. Specifically, the study primarily assessed whether the rate of clinically interpretable images was non-inferior to that of traditional in-clinic ultrasound, and whether the resulting estimates of AFC were equivalent.

Methods:
***Line 18-185 and Line 203: If the authors were going to dichotomize into "clinical quality" or not, why did they have them rate on a scale with criteria? When a scale is being utilized, there should be some justification for why this was then dichotomized further.**

Response: Thank you again for pointing out this important comment. We agree this justification could be more detailed. The four-point scale was driven by physician input to try to mirror existing clinical logic. Given that physician logic appeared to have four primary steps, this resulted in a scale that was further dichotomized. Ultimately, we designed an image quality endpoint that attempted to faithfully replicate the actual clinical decision process used to evaluate real-world ultrasound images: an initial subjective assessment of suitability for interpretation, followed by the conduct of specific diagnostic tasks (e.g. AFC estimation). Interviews with REIs and generalist OB/GYNs responsible for ultrasound instruction mapped that process to the four-point scale used in the trial, which is similar to scales used in other imaging studies. In practice, independent raters found the scale representative of their clinical practice, and for that reason, likely to correlate to clinically meaningful results. We have added a statement to this effect to the Methods section (see Lines 207-212).

> Given lack of an existing consensus for TVUS image quality criteria[26–28], a four-point scale was developed from interviews with expert physicians who regularly assess images (Supplementary Table 1). During scale development, emphasis was placed on replicating and codifying the thought process from usual clinical practice and ensuring all clips could be intuitively categorized by an expert physician. The scales for imaging quality assessment used by Hausleiter et al were used as a guide for quality rating[29].

From the original protocol (for reference): "There does not appear to be a single standard rating scale for ultrasound quality. Interviews with healthcare professionals yielded a 4 point scale, which has precedence in relevant literature. See, for instance, Hausleiter et al (2010). Image quality and radiation exposure with a low tube voltage protocol for coronary CT angiography results of the PROTECTION II Trial."

***Line 218-221: The authors must present the power they wanted to achieve and the sample size needed to do so, in addition to presenting the non-inferiority margin and the alpha statistic.**

Response: We appreciate this feedback around power. We have added the power calculations that informed sample size calculation, along with associated non-inferiority margins and alphas. Under reasonable a priori assumptions, the study was >90% powered for the primary image quality and AFC equivalence endpoint, though only the former drove sample size considerations. The primary image quality endpoint would also have been >85% powered for a more conservative 10% non-inferiority margin under performance assumptions that matched results obtained from the trial. This was further described in the Statistical Methods section (lines 242-254):

> Sample size was selected to ensure >90% power to demonstrate non-inferiority of the clinical quality image rate, given a non-inferiority margin of 18%, a true margin of up to 7%, a 95% clinical quality image rate in-clinic, an adjustment for up to a 60% intra-patient correlation between organ images, and an alpha of 2.5%. This alpha was chosen based on convention for one-sided tests, utilizing FDA guidance[19]. AFC equivalency was well-powered (>90%) under reasonable assumptions.
>
> Power calculations for sample size and confidence intervals were based on a non-inferiority test for paired binary data using a RMLE-based test-statistic[19,22]. The non-inferiority margin was chosen utilizing FDA guidance that the maximum acceptable margin (18%) is justified in cases of major improvements in tolerability, access to care, and lack of irreversible outcomes[19,36]. In practice, such a generous margin proved unnecessary—the trial was still >85% powered to demonstrate non-inferiority at a 10% non-inferiority margin, given the observed delta of 2.4%.

***Line 218-221: Why was an alpha of 2.5% chosen, as opposed to 5%? I was merely curious, and I imagine readers may be as well, as this is unusual.**

Response: Thank you for this question! This is the alpha suggested for use by FDA guidance for non-inferiority tests. It is conventional for one-sided tests (such as those for a non-inferiority hypothesis) to have half the type I error rate of two-sided tests. We have added a sentence to this effect in the Statistical Methods section (see Lines 298-313). The pertinent FDA guidance may be found at this link: https://www.fda.gov/media/78504/download.

***Line 230-233: If multiple organs from each subject contributed to the primary outcome, were they counted separately as endpoints? Or did each contribute to a "clinical quality" stamp for that one patient TVUS event? This should be made more clear, as it is hard to determine if appropriate statistical analyses were used if the clinical quality of one organ was analyzed as being independent from the clinical quality of another organ on the same patient (which would not be appropriate).**

Response: Thank you for this comment. This section has been largely re-written for clarity in response to your feedback (See Lines 255-266 below). Each organ from each patient was assessed for clinical quality and counted separately, i.e., each woman contributed six randomization events that were assessed separately by the raters, with the exception of one woman with prior unilateral oophorectomy who contributed four. However, multiple organs from the same patient were not treated as fully independent. Instead, we followed FDA's feedback to statistically test the independence of organs (which could have been completely independent, not independent at all, or partially-correlated). While finding each organ on a home ultrasound is practically a separate clinical exercise, factors like body habitus may drive some level of correlation requiring adjustment.

This adjustment was accomplished be re-weighting each observation based on the number of events contributed by that subject. The nature of the RMLE test statistic allows for this re-weighting to produce interpretable results. As expected, this correlation was ultimately low but positive (0.32), allowing for effective control of intra-patient correlation as demonstrated by empirical testing on simulated data. The details of this adjustment, which were too extended to cover in the paper, but were included in the study protocol, follow. The authors would be willing to submit this detailed explanation as part of the supplementary materials if deemed appropriate.

Of note, these adjustments proved unnecessary. If each organ imaged (uterus, left ovary, right ovary) had been assessed as a separate endpoint, with one at-home/in-clinic image pair provided per patient and no other adjustments, all three would have demonstrated non-inferiority in clinical quality imaging rate to in-clinic scanning (97.5% CI lower bounds -6.3%, -6.2%, and -12.4%, respectively). The authors found this approach unnecessarily conservative. See Lines 255-266 below.

> The primary endpoint analysis treated each organ from each subject as a separate data point. Importantly, independence of intra-subject image ratings was not assumed. Multiple randomization events from the same women were down-weighted based on the intra-patient correlation in clinical image quality (0.32). This adjustment formula was defined *a priori* in the study protocol in response to FDA feedback to test independence of ratings. Additionally, organs obtained at-home were excluded if acquired in less time than was deemed to be a credible independent process from the previous cine clip (<30 seconds to locate the subsequent organ). This process excluded nine organs (5% of images obtained at-home), which had all been rated "clinical quality." After adjustment, each subject contributed, on average, the statistical equivalent of 2.05 independent observations to the primary endpoint across the three organs imaged. Despite their complexity, these adjustments had little impact on the study's conclusions as each organ met the pre-defined noninferiority endpoint when assessed separately.

Please see below for additional statistical details on the organ independence test. Of note, CONSORT guidance appeared to suggest including most of this in the manuscript, which we originally did; however, in response to another reviewer comment to the length of this Methods section, and because it has little impact on study conclusions, we have decided to leave this in the Response to Reviewers here in blue:

As ovaries are intrinsically mobile in their *in situ* pelvic position, the processes for imaging each organ within a given patient are likely to be independent of one another. Successful imaging requires variable angles and depths of insertion of the vaginal probe ultrasound for adequate visualization at different points in time, and may be visualized through different propagation of ultrasound. Therefore, based on clinical experience, the acquisition and successful imaging of each organ is a separate and independent process, albeit a process that takes place on a woman with the same body habitus. Significantly, this is different than other common imaging modalities, e.g., mammography, where the location of the organ in question is generally known and the acquisition of subsequent images is largely a replication of acquiring the initial image.

After recruitment of the study began, FDA provided feedback on this assumption of independence. While FDA expressed no objection to this clinical rationale, there was specific concern that this intra-user image independence assumption receive quantitative testing and appropriate control for any correlation present. Any observed intra-user correlation must originate from one of five distinct sources, with aggregate correlation being the sum of the five effects:

1) Imaging process factors: If the successful acquisition of one organ capture mechanistically makes it more likely to successfully capture the next organ, likely due to minimal changes or effort required in acquiring the next image.
2) Technological factors: As very poor internet or data connectivity can cause lag in the streaming ultrasound image viewed by the ultrasound technologist (though not in the captured image itself) it is possible that those users may face a consistently more difficult process of organ acquisition.
3) Ultrasound technologist factors: If different technologists have different proficiency, technologists effects may be captured in intra-user correlation, as all images from one user would be created by the same technologist.

4) Subgroup factors: It may be the case that certain subgroups, such as high-BMI patients, have consistently lower or higher quality images.
5) Patient-specific factors: For whatever reason, some individuals may be able to consistently obtain higher or lower quality images of their own organs.

In order to address FDA's concern, and ensure that no assumptions were violated in the analysis of the primary endpoint, each factor was assessed explicitly:

(1) Imaging process factors were addressed by excluding any at-home scan that was obtained in under 30 seconds, suggesting, according to interviews with ultrasound technologists, that the organ was located during the scan of a previous organ and did not require a separate process to obtain. A sub-analysis of time to acquire showed no significant difference in time to image each subsequent organ, suggesting no change in difficulty over the course of an exam.

(2) Technological factors turned out not be significant in the trial. There was initial concern that areas of low internet connectivity might prevent the successful completion of exams. Technical issues and lag were noted qualitatively by the ultrasound technologists, and while there were a handful of minor issues reported, none prevented an exam from occurring, nor were they correlated with image quality scores As clinical images are stored at full quality locally, artifacts present for a remote technologist are not present to the clinician reviewing the scan itself.

(3 + 4) Correlation due to the subgroup and technologist factors is inherent to any study design with either multiple technologists or multiple subgroups. Subgroup analyses were performed and are discussed in the body of the text. No subgroup showed significantly lower performance than the overall population.

(5) Pure intra-user image correlation can increase variance in test statistics as additional images are treated as new data when, in fact, they partially restate existing findings. If intra-user correlation is <1.0, however, additional images from the same user do, in fact, provide new information - though less than what a completely independent image would provide. Due to the nature of the RMLE-based test statistic, however, it is possible to treat individual organ ratings as fractional while still producing a reliable and interpretable result. Analysis of simulated data confirmed that an unbiased test statistic could be generated via adjustment, and that that adjustment could be controlled to further ensure a strictly *lower* type I error rate across repeated simulations – ensuring the adjustment is somewhat conservative compared to an 'ideal'. The fractional observation adjustment used is as follows:

As pre-defined in the protocol, first intra-user correlation is calculated based on a linear regression of the average quality score of an organ image by the average score of the previous image obtained from that user (if any), binary fixed effects for (1) BMI >30, (2) BMI <20, (3) previous ultrasound experience, (4) Age >35 (but up to and including 38), (5) Age <27, and (6) the performing ultrasound technologist. Images that took less than 30 seconds to capture or were identified as poor due to technological problems will not be included in this analysis. The degree to which the previous scan quality rating correlates with subsequent scan, after controlling for fixed effects (e.g., how well it correlates with the residual in a linear regression including the fixed effects) was the estimated correlation used for the adjustment factor. In practice, none of these fixed-effects were significant and their inclusion had no impact.

Next, an adjustment factor is calculated for data points from women who provided three organ-rating pairs and a separate factor is calculated for data points from women who provided only two organ-rating pairs. The more organ rating pairs provided, the greater the adjustment and the less each individual data point from that woman contributed to the final analysis. The formula for weighting each imaging rating from a woman who provided three pairs of organs to the analysis was:

$$\frac{1 + (1 - [correlation]) + (1 - [correlation])^2}{3}$$

The formula for weighting each imaging rating from a woman who provided two pairs of organs to the analysis was:

$$\frac{1 + (1 - [correlation])}{2}$$

These weightings were used instead of '1' for each observation when populating the confusion matrix used for the RMLE-based test statistic. If only one image was included from a subject for a given analysis then each adjustment is required.

The net effect of this process was to reduce the effective sample size of the study to 114.9, though all 167 organ-pair randomization events were still analyzed.

***Line 263-275: I have never before read a study utilizing the NPS, so this is interesting. I would like to see more explanation of what difference between the NPS of one intervention (in-clinic TVUS) versus another intervention (home TVUS, in this case) is considered clinically meaningful. This would help put the results presented later into context.**

Response: As a reliable indicator of patient preference, we believe that any statistically significant difference in NPS between modalities is clinically meaningful. That said, the size of the difference obtained, 58.1 points, 97.5% CI of difference (37.3, 79.0), was frankly a surprise even to the investigators. For reference, the difference in average patient NPS between total knee replacement and total hip replacement surgeries is only 22 points, despite significant differences in patient tolerability and recovery time (Hamilton *et al.* 2014, The Bone & Joint Journal). The difference between at-home and in-clinic experience is driven by patient dissatisfaction with in-clinic ultrasound, whose NPS of negative 12 is lower than any other benchmark we have identified in healthcare literature. Low and negative scores are typically rare, as there are not many services that can withstand such a high level of dissatisfaction. In response to this comment, we have added context on what a 'good' NPS would be to the methods section (Lines 281-285). Further interpretation of the magnitude of the difference occurs in the discussion section (Lines 390-395).

> The value of this measure lies in its ability to compare between two similar options, with the higher NPS likely translating to significantly better word-of-mouth referrals for a certain product, experience or company, over another in the same arena. Any NPS from -100 to 0 would warrant improvement, while an NPS between 30 to 100 would be deemed a positive to great experience[32].

> Many patients found the virtual experience to be one they would highly recommend to others, with an NPS well above the healthcare average[31–35,51] and significantly superior to that of the in-clinic experience. As a point of reference, the NPS difference between at-home and in-clinic exams (58.1 points) is nearly three times greater than the NPS difference between a total knee and a total hip replacement, procedures known to have significant differences in patient tolerability[31].

Results:

***Line 294: Please explain better in the Methods section what was being timed for the duration of the exam. Was this from the time the transducer went in the vagina to when it went out? Or the cumulative time of all cine clips from the patient? This should be more clearly defined for in-office and at-home exams.**

Response: Thank you for this comment. We agree this was ambiguous as written. The number reported is the total time actively scanning the patient, from insertion into the vagina to probe removal. It represents the full time of direct exposure to ultrasound. It includes the time in-between organs during internal scanning, as well as cine clip capture of each organ. We have clarified the language in this section (see Lines 312-313).

> The mean duration of the virtual exam, accounting for the entirety of the time the probe was inserted into the vagina, was 13 minutes 40 seconds (SD 3.93 minutes).

***Line 295-295: Please give a p-value for this difference or a mean difference with confidence intervals.**

Response: Thank you for pointing this out. We agree this section benefits from additional detail. The mean difference in active scanning time was 5 minutes 10 seconds, 95% CI (3.93, 6.37). We have added this detail to the results section as noted in Line 314-315.

***Line 310-314: Give the actual number of women with high BMI here, as opposed to just saying "34% of the population", and give the confidence intervals around the percent difference or the p-value for the 93% vs. 98%.**

Response: Pertaining to your comment on BMI, we have clarified the patient count. 19/56 women had higher BMI (Line 305). Regarding your comment around the CI around the % difference, this statement (Line 329-330) has also been clarified (p=0.02).

> The mean age of participants was 27.6 (19-35) years old and mean BMI was 23.6 kg/m^2 (19.5-33.9 kg/m^2), 19 (34%) of whom had a BMI $\geq$ 25 kg/m^2.
>
> In another subgroup analysis examining subjects with BMI > 25 kg/m^2, the difference also remained non-inferior, with a difference of 5.6% (*P*=0.02).

***Line 350: The definition and explanation of "nonreplicable findings" is not clearly explained in the methods, although it is mentioned in the abstract and the Results. This should have been more clearly delineated in the Methods section.**

Response: Thank you for highlighting this oversight. This was an editing error on the part of the authors. The definition of nonreplicable finding, which was included in the registered study protocol, has been re-inserted (see Lines 291-295).

> Nonreplicable findings were cases where a significant finding on an at-home exam was not seen during subsequent in-clinic scan. A significant finding was defined as either at least one rater noting they 'definitely' saw a finding for which follow-up was recommended, or two raters noting they were 'almost certain' about it. Any significant finding at-home not seen in-clinic was considered non-replicable.

As noted in the revised text, non-replicable findings were cases where a significant finding on an exam was not reproducible in-clinic. A significant finding was defined as either (1) at least one rater provided a rating of "Yes, definitely" to question "Do you observe any (other) observable pathologies or abnormalities?" AND 'Yes' to question "If you answered something other than 'No' above, do you recommend the subject receive follow-up care for this pathology or abnormality (if not yet receiving care for this condition" or (2) at least two raters responded "Almost certain" to the first question and 'Yes' to the latter question. As such, any significant finding at-home would also need to be seen in-clinic, or be considered a non-replicable finding for this endpoint. In-clinic significant findings were only deemed non-replicable if follow-up care provided by a physician could not replicate the finding.

***Line 334: Again, you should give a p-value for the difference between NPS between the two interventions and state was that p-value reflects (difference in percent that are promoters? Difference in score?)**

Response: The 97.5% CI of the difference in NPS was (37.3,79.0) and had a p-value of <0.01. We have pulled the data forward for clarity and emphasized that the difference is in net NPS, calculated using the methods described in the methods section. See Lines 351-352.

> Across all subjects, virtual scanning's NPS was 58.1 points higher than recalled in-clinic scanning, 97.5% CI (37.3,79.0), *P*<0.01.

***Table 2 gives no comparative statistics between at-home and in-clinic TVUS groups, although it gives values of outcomes for each group.**

Response: Thank you for this comment. We have reformatted the table to more clearly highlight the comparative statistics on the same row as the in-group outcomes. It is important to note that the same population is represented in both virtual and in-clinic settings. Please see Table 2.

Discussion:
***Line 368: Were these populations that were getting fertility workup, or other types of ultrasound assessments? There should be a bit more discussion about what these past studies left unanswered, and therefore how this fits in the literature.**

Response: This is really helpful context to add to the Discussion. In response to your feedback, we have added clarification to this section in the text (see lines: 378-379). The inclusion of these references was intended to provide qualitative context for the overall performance of the device via comparisons to rates deemed reasonable in real world settings. The comparison is neither direct or statistical: The citations represent a diversity of patient populations, from normal women recruited specifically for the study, to all TVUS performed in emergency rooms, to a retrospective analysis of ovarian cancer screening data, to an early pregnancy population, to recommended guidelines for sonographer evaluations for use by OB/GYN practices. Our literature search did not reveal any published data on successful imaging rates in women receiving a fertility workup, nor any studies using the same device used in the trial, or assessing imaging quality in an at-home setting.

While recognizing those significant caveats, the published data supported an interpretation that at-home quality rates are not qualitatively different than in-clinic quality rates for other common TVUS use cases. The total range of visualization rates reported (requiring only that the full organ be present in a clip—a lower bar than that used in the trial), was 90-99%. That range covered cases that appeared to have a more consistent visualization rate (early pregnancy, 99%) and lower rates (minimum sonographer performance standards for all OB/GYN imaging, 90%). Looking at only imaging rates for the ovaries and uterus in reproductive age women reported in these studies, that range narrowed, with reported average visualization rates ranging from 92-97%, essentially the same as the 97.5% confidence interval for at-home device clinical quality imaging rates (92%-99%).

> Virtual ultrasound produced similar clinical quality images to those obtained in in-office practice, where reported rates of optimal pelvic organ visualization range between 90-99%, with average visualization rates of 92-97% seen in reproductive-aged women similar to those represented in this trial[44–47]. While not direct comparisons, as they do not represent a population of women seeking ovarian reserve information or the device used in the trial, they do suggest that the absolute at-home imaging quality rate, 96%, 97.5% CI (92%, 99%), is not qualitatively different than the in-office quality rates seen for other common TVUS use cases. These high imaging quality rates persisted after excluding HCP-subjects and after specifically examining subjects within the highest BMI ranges. While higher BMI was associated with slightly lower image quality rates, performance of the virtual TVUS in this group (93%) was consistent with prior studies utilizing traditional TVUS[48–50].

***Line 386-394: It should also be mentioned that the study was underpowered to assess important clinical outcomes, or power for these outcomes was not assessed.**

Response: Thank you for this comment. We have provided additional detail around power considerations in our Statistical Methods section (see Lines 242-254). The study was generally powered to ensure >90% power for the primary outcome of non-inferiority of the clinical quality image rates and generally expected to be powered for the secondary endpoints, under a range of reasonable expectations. The AFC equivalence hypothesis was also 90% powered for a mean AFC per ovary of 8 and a standard deviation as high as 7, which are conservative assumptions based on published AFC nomograms. Particularly because of the novelty of the imaging method, and due to some uncertainty that existed prior to the trial about how intra-cycle, inter-modality, and intra-rater variability in AFC estimation might impact the

variation of individual estimates, such a conservative estimate appeared reasonable. In practice, the study was >95% powered to detect AFC equivalence given the observed variation in-sample and the 2.65 equivalence margin.

Reviewer #2:

Abstract
***Line 21: What was the population of women who were recruited for this study?**

Response: Thank you for this clarifying question. Recruitment focused on professional women from a diverse range of careers with an interest in learning more about ovarian reserve and fertility potential. The population likely included those who may never have attempted conception in the past, those who were not actively trying to conceive, and those who may meet the clinical definition of infertility. To our knowledge, no participants were actively undergoing a medical evaluation or treatment for infertility. Recruitment channels included emails to professional networks, Facebook groups, social media advertising, and OB/GYN referrals. Search terms included items like "fertility testing" and "how many eggs do I have left." (see Lines 30-32 and Lines 38-39).

> Subjects were women in the greater Boston area interested in evaluating ovarian reserve and recruited via social media, healthcare referrals, and professional networks.
>
> Fifty-six women participated from December 2020 to May 2021. Subjects varied in age (19-35 years), BMI (19.5-33.9 kg/m^2), and occupation.

***Lines 33-35: Patients have to come in for a sono or HSG anyway and it only takes an additional minute to perform an AFC at the time of HSG or sono, so does at-home TVUS for AFC really save time and money?**

Response: Thank you for this thoughtful perspective. In this population of patients who are seeking ovarian reserve assessment, uterine cavity/tubal evaluation is not always performed initially given their current aim is not to conceive or undergo treatments to conceive in the short term. A TVUS is a common, less-invasive first step for OB/GYNs seeing fertility consults in the gynecology clinics, and given the possible risks, discomfort, and lack of easy access or availability with HSG, there seem to be diverse opinions by OB/GYN and REI providers on the use of HSG as the primary diagnostic test as opposed to TVUS. Uniquely, virtual TVUS provides the opportunity for a less cost-intensive ovarian reserve assessment compared to office visits for patients who may not have insurance coverage. Also, for patients in underserved areas, virtual TVUS would obviate the need for long distance travel to metropolitan areas or the use of public transportation for those within urban settings.

If there is suspicion for tubal factors (whether due to a more extensive length of time to conceive or tubal/uterine factor risks), then at-home TVUS may not remove a visit and therefore may be less time/money saving than going straight to an HSG. We do recognize that clinical practice varies in this regard, and some HCPs do feel strongly about performing HSGs or sonohysterograms on all patients.

Introduction
***Lines 42-44: "…limited public awareness of age-related fertility decline." Followed by: "interest in at-home fertility testing has emerged as a cost-considerate, easy-to-access initial option." These statements seem contradictory. If people are not aware of age-related fertility decline, how are they going to know about at home TVUS for AFC?**

Response: You do bring up a great point. This comment touches on an important issue that we have attempted to clarify. While women do systematically underestimate the impact of age on fertility, there is expanding interest in knowing fertility status. The US market for at-home AMH tests is growing rapidly - the market leader in this space, Modern Fertility, reported annual sales of $14.3M on an average price of $159/test, implying ~90,000 tests purchased from them alone

(https://growjo.com/company/Modern_Fertility). There are at least three other competitors (Everlywell, LetsGetChecked, MyLabBox) offering similar testing, implying a minimum annual test rate well over 100,000 women per year. This number is up from ~0 as recently as 2014. There are, however, currently around 38 million women of reproductive age in the United States, leaving significant opportunity to increase awareness and access.

At the same time, and despite consistently high levels of concern about fertility among women of reproductive age, as reported in various surveys (e.g. https://www.businesswire.com/news/home/20170504005519/en/Attitudes-of-Millennials%C2%A0Regarding-Fertility-Revealed), the impact of age-related fertility decline is consistently underestimated (see for instance, C. Lampic et al. Fertility awareness, intentions concerning childbearing, and attitudes towards parenthood among female and male academics, *Human Reproduction*, Volume 21, Issue 2, February 2006, Pages 558–564, https://doi.org/10.1093/humrep/dei367). Our takeaway from this data is that there still a need to make medical information more accessible and interpretable for patients planning their fertility. We do agree, however, that it will take time for many women to become aware of their expanding options. See lines 55-60 for our edits per your helpful feedback.

> Infertility impacts one in eight reproductive-aged couples in the United States (US)[1]. However, diagnostic testing for this disorder is often not pursued due to cost, lack of insurance coverage, and underestimation of age-related fertility decline among women of reproductive age[2]. Recently, with increasing average maternal age and more public discussion of infertility, there is expanding interest in at-home fertility testing has emerged as a cost-considerate, easy-to-access[3,4].

Role of the Funding Source
***Lines 64-54: Turtle Health's involvement in the study's design, data analysis, and reporting introduces bias. How was this accounted for?**

Response: We appreciate that this was a funded industry study, and all the authors recognize this is a limitation and introduces potential risks. We have made several mitigations: similar to other prior sponsored studies that have been published in this journal (e.g., Burke *et al* 2019), all conflicts were fully disclosed as requested by the Journal. None of the authors received equity compensation. In addition, the FDA has independently reviewed the underlying source data (cine clips). The reviewers and editors at *Obstetrics and Gynecology* are fully welcome to review any or all cine clips as desired, for further independent verification.

Methods
***Lines 96-97: Did you try to recruit patients with known submucoal fibroids? Did this introduce bias? Were other patients not approached to participate in the study because they did not have known submucosal fibroids?**

Response: Thank you for this question. Extensive efforts were undertaken to recruit submucosal subjects. This included outreach to reproductive surgeons in the Boston area via the PI, social media advertising, partnering with local resident physician clinics and reaching out to several infertility and fibroid patient groups. The Sponsor trial team includes two recruitment experts (former Pfizer and Sanofi). The lack of submucosal population recruitment likely reflects a combination of COVID conditions, lack of direct benefit of participation to this population (since they are already diagnosed), and, based on feedback with a fibroid patient advocacy group, historical distrust of the healthcare system and therefore lower willingness to participate in trials.

Patients without submucosal fibroids were successfully recruited, with separate IRB-approved recruitment materials used for submucosal and non-submucosal subjects to ensure appropriate recruitment approaches toward each population. This was further clarified in the Methods (see Lines 112-122).

> Two study arms were planned: (1) N=30 healthy patients and (2) N=15 patients with a history of submucosal fibroids, with IRB-approved recruitment materials designed for each population. Despite three months of recruitment, no patients with known submucosal fibroids were identified. As a result, the fibroid

> study arm was eliminated. Additionally, the methods were updated to address the FDA's feedback that the statistical independence of organs needed to be tested for intra-subject correlation. Finally, in response to the changes, the total sample size of the healthy trial arm was increased to a target of N=55, based on the number of subjects required to fully power the primary endpoint and projected capacity of the trial site. Power calculations remained robust and the recruitment materials were unaffected. This amendment was submitted prior to the review of any cine clips by the independent raters, and thus the planning of analysis remained *a priori*.

***Lines 125-126: "…otherwise in the professional judgement of the Primary Investigator (PI) able to give informed consent." What does this mean? The PI decided who was or was not able to give informed consent and thus participate in the study? Did this introduce bias?**

Response: Thank you for this comment. In practice, this provision was never used (e.g., all patients who met the other inclusion/exclusion criteria and wanted to consent were invited to do so). As such, this did not introduce bias. This condition was originally included in the protocol to cover the unpredictable range of possible scenarios present in a COVID environment where the PI lacked the normal in-person cues and information typically used to evaluate whether informed consent is being freely given. See Lines 139-144 for clarification per your feedback:

> Eligibility criteria included females between 18 and 38 years old (inclusive), body mass index (BMI) up to 40 kg/m^2, and the ability to independently give consent electronically. Electronic consent required that a subject speak native or fluent English and have a high school degree or equivalent. The study protocol authorized the Principal Investigator (PI) to use their professional judgment if there were a question about patient ability to give informed consent virtually, but in practice, all patients meeting study criteria were invited to consent.

***Line 139: 64 patients were consented by only 56 ultimately participated in the study. What happened to the other 8 participants?**

Response: Thank you for this question and happy to clarify! All declined to schedule their exams for various personal reasons. The breakdown is provided here and has been added to the paper. Four patients stopped responding to trial coordinator emails. One patient had a family emergency requiring a departure from the trial catchment area at short (three days) notice, for the duration of the trial. One patient had scheduling conflicts preventing completing participation in a timely manner, despite trial efforts to accommodate. One requested withdrawal for unspecified reasons. One disclosed material new information to the trial coordinator subsequent to signing her informed consent form that made her no longer an appropriate candidate for the trial (a mid-cycle birth control change of uncertain impact on risk of subclinical pregnancy; given the PI's recommendation for an additional cycle to reach steady state and exclude an incidental pregnancy, the trial would have needed to significantly delay the last patient last visit to accommodate, which was not feasible). We have updated this section accordingly in the paper (see Lines 157-160):

> Of those consented, 56 participants were ultimately enrolled. Of the other eight, four stopped responding, one left the trial catchment area, one withdrew for unspecified reasons, one was unable to schedule, and one disclosed disqualifying medical information after consent.

***Line 142: If subjects had oligomenorrhea or amenorrhea were visits still scheduled on consecutive days, or just whenever the subject had availability? How did this affect your results if the visits were not on consecutive days for this group of patients?**

Response: Thank you for this question. Every patient was scheduled on the same or consecutive days. Oligo/amenorrhea affected only the scheduling of the first (virtual) visit, allowing for that to occur at any time; the in-person was always within 1 day of the first visit, whenever that occurred. This was further clarified in Lines 160-163.

> In order to reduce the risk of inadvertent fetal exposure, a regulatory concern to the FDA[25], women with regular menses were scheduled between days 3 and 10 of their menstrual cycle. Women with oligomenorrhea or amenorrhea were scheduled on consecutive days as well but at any time in their cycle.

***Line 162: How often were there discrepant ratings on clinical quality requiring a tie-breaker?**

Response: In total, 15 of 334 (4%) organs required a tie-breaker. These included 8 organs captured at-home and 7 captured in-clinic. This information was included in Supplementary Table 2 and is mentioned in the body of the text in the 'Quality Rating Procedure' section (see Lines 221-223).

> Each primary rater provided ratings for all 334 such folders. The third rater ultimately rated 15 of 334 (4%) folders (8 at-home, 7 in-clinic) when there was a disagreement on quality between the primary raters.

***Line 164: How much were participants compensated for their participation?**

Response: Participants were compensated $100, which was discussed with IRB as a reasonable Fair Market Value (FMV) for the in-person comparator, inclusive of time, travel to site, PPE, parking costs, etc. (See Lines 186-187).

> All participants who completed the study received their results and were compensated for compensation at fair market value.

***Lines 177-179: This is confusing. Was a rater given all the cine clips from one patient for each location to view back to back? Ie, did the rater view, for one patient, all the cine clips from the at home TVUS immediately followed by all the cine clips from the office TVUS? Or did the rater view one cine clip from one patient then another random cine clip from another patient**?

Response: We agree that this section benefits from clarification. The last sentence of your comment most accurately captures the rating procedure, with the caveat that each rater viewed and rated *two* cine clips at a time covering *one* organ, namely sagittal and transverse clips of the same organ from the same exam. These clips were kept together so that the raters could view the same organ from two dimensions before scoring it. The raters viewed these clips of one organ from one patient, then another randomized organ from another patient. With each woman contributing six randomization events (left ovary; right ovary, uterus; from home and in-clinic) and a minimum of six women's organs randomized for the independent raters from both home and in-clinic, raters saw a random sequence of, at minimum, 36 home and in-clinic ovaries and uteri. This has been clarified in the text (see Lines 193-204):

> Cine clips for each organ captured at-home and in-clinic were assigned to folders named based on an Excel-generated list of 500 4-digit random integers via 500 instances of the formula ROUND(10000*RAND(),0). Each folder contained two views (sagittal and transverse) of a single organ (left ovary, right ovary, or uterus) obtained from a single participant in a single setting (at-home virtual or in-person). There were 334 such folders in total (56 subjects x 3 organs x 2 modalities, less 2 folders due to subject with only one ovary). Folders were sorted based on their file number from smallest to largest within each batch of videos provided to the raters, randomizing the order they were viewed by the raters. Video folders were provided to raters in batches that included, at minimum, all cine clips (three organs in two different settings) from six different subjects, totaling a minimum of 36 folders per batch. Raters were blinded to the organ, patient, and setting, and viewed a random sequence of organs from multiple subjects and settings in each batch.

***Lines 186-188: This is confusing. Perhaps the grammar makes this sentence difficult to follow. What are you trying to describe?**

Response: As per your feedback, the structure of the sentence has been reworked for clarity and to address additional comments from Reviewer 1. Please see Lines 214-220

> For the purposes of assessing the primary endpoint, a “clinical quality” scan of an organ required two raters to provide a score of either “3” or “4” *and* that the same raters identified the target organ correctly. (i.e., complete visualization of the ovary [cortex, stroma, antral follicles] in perpendicular planes [sagittal, transverse]). If the rater did not accurately identify the target organ, these scans were considered beneath clinical quality and the folder was rated as less than the threshold for clinical quality (scored as “1”).

***Lines 193-194: At this point, it seems clear that each cine has its own folder. However, this contradicts with lines 173-175 as well as line 198. Please clarify what was actually in each folder. Is there an ovary cine in each folder? Or do some folders only have uterus cines.**

Response: Each folder contained two views (sagittal and transverse) of a single organ (left ovary, right ovary, or uterus) obtained from a single participant in a single setting (at-home virtual or in-person). There were 334 such folders in total (56 subjects x 3 organs x 2 modalities; subtract 2 folders, one per modality, for a subject with only one ovary). The relevant sections have been clarified (see Lines 195-198).

> Each folder contained two views (sagittal and transverse) of a single organ (left ovary, right ovary, or uterus) obtained from a single participant in a single setting (at-home virtual or in-person). There were 334 such folders in total (56 subjects x 3 organs x 2 modalities, less 2 folders due to subject with only one ovary).

***Lines 224-228: Confusing, run on sentence, please reword**

Response: This paragraph has been largely re-written in response to this and other reviewer comments.

***Line 229: Do you mean "rating" rather than "rate"?**

Response: Thank you for this clarifying question. This is correct as written. The endpoint is the difference in the rate at which organs obtained by each modality were determined to be clinical quality, as described in the methods section.

***Lines 233-235: How often did this happen? Was it with the at home or clinical scan more often?**

Response: Removing organs that were acquired close together in time occurred 9 times (5% of organs imaged at home). All excluded images were rated as 'clinical quality', so their exclusion did not impact the conclusion of the study. The text was updated to note that this exclusion only applied to virtual scans, where such correlation was of primary concern (see Lines 259-266).

> Additionally, organs obtained at-home were excluded if acquired in less time than was deemed to be a credible independent process from the previous cine clip (<30 seconds to locate the subsequent organ). This process excluded nine organs (5% of images obtained at-home), which had all been rated “clinical quality.” After adjustment, each subject contributed, on average, the statistical equivalent of 2.05 independent observations to the primary endpoint across the three organs imaged. Despite their complexity, these adjustments had little impact on the study’s conclusions as each organ met the pre-defined noninferiority endpoint when assessed separately

***Line 236: Did you look at time to image for at home vs clinical scans?**

Response: The imaging time reported was active imaging only, from insertion into the vagina to removal of the probe from the vagina, and included both cine clip capture as well as time navigating to and between organs.

In response to this question as well as questions received from clinicians following the trial, a post hoc analysis of time from the scheduled beginning of the exam until the insertion of the probe was conducted, and showed no meaningful difference in pre-scanning (i.e., patient setup and preparation) time between modalities, with both home and in-clinic averaging around 10 minutes (specifically, means of 10 minutes 4 second at home and 10 minutes 29 seconds in clinic).

For the original scanning time analysis, see lines 312-315.

> The mean duration of the virtual exam, accounting for the entirety of the time the probe was inserted into the vagina, was 13 minutes 40 seconds (SD 3.93 minutes). Duration of the in-clinic exam was 8 minutes 30 seconds (SD 2.21 minutes). The mean difference was 5 minute 10 seconds, 95% CI (3.93, 6.37).

***Line 246: What do you mean by fractional?**

Response: We appreciate your comment. This Statistical Methods section was largely re-written in response to the other reviewer's comments. Fractional references the fact that each organ image rating was not treated as an entire observation for the purpose of calculating the test statistic for the primary image quality endpoint. Rather, a weighting factor was applied based on how many images were contributed to the analysis by that patient, reducing the impact of each image contributed for every additional image contributed. This was done to address concerns about test statistic inflation due to intra-patient correlation of image ratings. Ultimately, the conclusions of the paper are not sensitive to the use of this adjustment, which was designed *a priori* in response to specific comments from FDA.

***Lines 252-257: Please make discussion of the null hypothesis more concise**

Response: Thank you for this comment. This was made more concise (see Lines 268-271)

> The null hypothesis for AFC estimates was that the difference between measured AFCs at-home and in-clinic for each organ would be outside the prespecified margin of equivalence of 2.65 follicles. The 2.65 follicles represents the average inter-cycle variation reported in literature[39–41].

***Line 268-269: "in the its ability" please fix grammar**

Response: Thank you for catching this error—this was fixed in the manuscript.

***Lines 272-273: What about patients who did not have previous in-clinic TVUS experience? How was their view of at home TVUS experience evaluated if they didn't have anything to compare it to?**

Response: Patients without previous TVUS experience who provided an NPS for virtual at-home scanning were coded as a neutral, or '0' NPS for recalled in-clinic scans. As recalled in-clinic NPS was overall negative, this had the effect of improving the average recalled in-clinic NPS and reducing the estimated difference between virtual and recalled in-clinic NPS. Without this adjustment, which is favorable to in-clinic scans, the at-home NPS would have been 75.9 points higher than in-clinic, instead of 58.1 points higher. The text has been updated for clarity (see Lines 285-289)

> NPS for virtual scans was expected to be significantly higher than the NPS of recalled in-clinic scans (for patients with previous TVUS experience). Patients without previous TVUS experience who provided an NPS for virtual at-home scanning were coded as a neutral, or '0' NPS for recalled in-clinic scans. As recalled in-clinic NPS was overall negative, this had the effect of improving the average recalled in-clinic NPS.

***Overall: Please shorten the methods section and make it more concise**

Response: We have reduced the length of this section. A significant portion of the length of this section was driven by the content describing the primary endpoint computations, due to CONSORT guidelines. We have shortened this description significantly, and provided this detail up above **in blue**. We hope the abbreviated version in the paper is adequate for CONSORT guidelines. If it would be helpful to publish the materials contained in the **above response in blue**, we would be pleased to do so (whether in the Methods section - though we realize that may significantly affect word count - or as a supplement, together with the patient instructions which were requested in another comment).

Results
***Line 283: What are the standard deviations or ranges for age and BMI?**

Response: Thank you for this comment. Participant characteristics, including age and BMI ranges, were included in the Results sectionin Lines 3004-306 and in Table 1.

> The mean age of participants was 27.6 (19-35) years old and mean BMI was 23.6 kg/m^2 (19.5-33.9 kg/m^2), 19 (34%) of whom had a BMI $\geq$ 25 kg/m^2.

***Line 287: "higher than the original 25% target for the trial". What does this mean? You were targeting patients with a BMI around 25? Does this introduce bias?**

Response: The original inclusion criteria specified at least 25% of the trial subjects should have BMI over 25 to ensure sufficient generalizability, given the sometimes greater difficulty of ultrasound in higher BMI populations. In practice, this target was achieved with no specific efforts to recruit this population, so bias is unlikely. The 25% target was shared with FDA in the draft protocol, as inclusion of higher-BMI patients was a regulatory request with a view to generalizability, but is not of particular relevance in this context. As such, this line has been removed from the paper.

***Line 291: OCPs stands for oral contraceptive pills**

Response: Thank you for noting this error—this was corrected.

***Line 338: What do the numbers "10" and "26" correspond to?**

Response: These numbers correspond to the number of respondents in each category of neutral response. These have been replaced by percentages.

***Line 348-349: While this is technically true because there were no false positives, there were no true positives either. This statement is a bit of a stretch.**

Response: We appreciate this comment and can understand the reaction to the use of specificity in this context. As the reviewer points out, this specificity calculation is technically correct, as there were no False Positives in the results, however it is unusual to discuss specificity in the absence of True Positives. This is because specificity is only noteworthy if there was a strong *a priori* reason to expect potential False Positives might occur. This expectation rarely exists in the absence of positives in a dataset. Given this feedback, we have scaled back and de-emphasized this statement, making the below statement instead (see Lines 361-362).

> No patient was identified as a having a submucosal fibroid, despite the raters being prompted on their survey.

Just to give some context for our original stance, our trial design did uniquely present a strong *a priori* expectation due to the presence of a specific question regarding submucosal fibroids on the rater form as well as the statements by the independent raters (during unblinding interviews) that they were actively looking for submucosal fibroids during their rating process. Notably, these statements included a question in the unblinding meeting about whether the rater had missed any fibroids in their rating process. An imaging device with low-specificity for submucosal fibroids should be expected to produce False Positives under such a design - potentially, even, at a higher rate than if submucosal images had been included, potentially allowing the raters to calibrate their assessments of clear examples. Instead, the study found that two heavily primed raters did not note a single submucosal fibroid in any of the 224 uterine cine clips they evaluated over the course of the study.

Discussion
***Line 375: "The NPS for…" This is confusing. Please specify that this is the NPS for at-home TVUS**

Response: This sentence was re-written for clarity (see Lines 395-396).

> The NPS of recalled in-clinic TVUS experience was, in fact, negative, potentially representing patients who would hesitate to seek care in-person.

***Lines 376-377: "likely representing patients who would hesitate to seek care in-person." This is speculation. Are patients really not going to seek care in person? Unfortunately, infertility treatment involves in-person care, in-person TVUS, in-person blood work, and in-person procedures. While consults can be done remotely, it would be impossible to treat a patient for infertility without seeing them in person.**

Response: Thank you for this comment. We have softened our language from 'likely' to 'potentially' (see Lines 395-396). In our clinical experience, some patients may postpone testing that they find possibly unpleasant or intimidating. We agree that many components of infertility care require in-person visits. However, for logistical or personal reasons, some patients may not prefer to have their initial evaluations on site and may find initial testing at-home more favorable, allowing their care to progress if appropriate or needed. As an example, one trial participant volunteered that she never would have gone to an REI clinic for care without participation in this trial. The thought of the clinic was too intimidating, but the privacy of an at-home setting felt much more manageable and allowed her to realize that she needed to subsequently access in-person care for diminished ovarian reserve.

> The NPS of recalled in-clinic TVUS experience was, in fact, negative, potentially representing patients who would hesitate to seek care in-person.

***Line 380: "characteristics" in pleural. What besides BMI was physically diverse?**

Response: Besides BMI (19.5-33.9 kg/m2), the patient population was diverse in age (19-35 years), professional occupation, contraceptive use, as well as range of physical disabilities. See Lines 399-405, as well as Participant Characteristics under the Results section (Lines 303-311) for more detail.

> The primary strength of our study was that the population was diverse in physical characteristics (BMI), range of professions, disabilities, and contraceptive use. Hence, our results are likely generalizable to the general population of reproductive-aged women. Notably, the trial population reflected individuals with disabilities including a visually impaired subject and another with a history of vaginismus and prior challenges with in-clinic care. The diversity of our trial population demonstrates the broad, proactive interest of reproductive-aged persons to pursue self-evaluation of ovarian reserve.
>
> Fifty-six participants were enrolled from December 2020 to May 2021. The trial population was diverse along a number of metrics (Table 1). The mean age of participants was 27.6 (19-35) years old and mean BMI was 23.6 kg/m2 (19.5-33.9 kg/m2), 19 (34%) of whom had a BMI > 25 kg/m2. Occupations ranged from undergraduate to graduate students, to a wide range of professional careers. Of the 5 (9%) of subjects who were verified HCPs (e.g., residents, nurses, therapists), none had prior ultrasound expertise or professional experience administering ultrasounds. Seventy-three percent reported using contraception: oral contraceptive pills (27%), intrauterine devices (41%), Nuvaring (4%), and progestin-only pills (2%). Two participants reported physical disabilities including visual impairment and vaginismus.

***Line 381: "persons" Persons or females? Sex assigned at birth females were included, not men, correct?**

Response: Thank you for your comment. We have adjusted this to say females (see Line 400-401).

> Hence, our results are likely generalizable to the general population of reproductive-aged females.

***Lines 388-389: Specificity was not actually assessed in your study so it is hard to say the specificity is 100%.**

Response: Thank you for this comment**.** As mentioned, it is admittedly unusual to discuss specificity in the absence of True Positives: specificity is only noteworthy if there was a strong *a priori* reason to expect potential False Positives. This expectation rarely exists in the absence of positives in the dataset. Given this feedback, we have scaled back and de-emphasized this statement, making the below statement instead (see Lines 361-362).

> No patient was identified as a having a submucosal fibroid, despite the raters being prompted on their survey.

Just to give some context for our original stance, our trial design did uniquely present a strong *a priori* expectation due to the presence of a specific question regarding submucosal fibroids on the rater form as well as the statements by the independent raters (during unblinding interviews) that they were actively looking for submucosal fibroids during their rating process. Notably, these statements included a question in the unblinding meeting about whether the rater had missed any fibroids in their rating process. An imaging device with low-specificity for submucosal fibroids should be expected to produce False Positives under such a design - potentially, even, at a higher rate than if submucosal images had been included, potentially allowing the raters to calibrate their assessments of clear examples. Instead, the study found that two heavily primed raters did not note a single submucosal fibroid in any of the 224 uterine cine clips they evaluated over the course of the study.

***Line 405: Write what "FP" stands for**

Response: Fertility Preservation has been spelled out in its entirety everywhere it is referenced in the paper.

***Line 406: Write what "RE" stands for**

Response: Reproductive Endocrinologist has been spelled out in its entirety the first time it appears in the paper, but given this is the last paragraph, we are happy to spell this out here.

Figures
***Table 2: Remove line regarding specificity as this was not actually assessed**

Response: Thank you for this comment**.** In response to earlier feedback, we have greatly de-emphasized the point on submucosal fibroid specificity particularly in the Results and Discussion to reflect that there were actually no True Positives in our population.

Given our trial design presented such a strong *a priori* expectation due to the presence of a specific question regarding submucosal fibroids on the rater form as well as the statements by the independent raters (during unblinding interviews) that they were actively looking for submucosal fibroids during their rating process. Notably, these statements included one nervous question in the unblinding meeting about whether the rater had missed any fibroids in their rating process. An imaging device with low-specificity for submucosal fibroids should be expected to produce False Positives under such a design - potentially, even, at a higher rate than if submucosal images had been included, potentially allowing the raters to calibrate their assessments of clear examples. Instead, the study found that two heavily primed raters did not note a single submucosal fibroid in any of the 224 uterine cine clips they evaluated over the course of the study. Given this context, we find the absence of any False Positives potentially clinically significant and important to mention. We have kept this line in Table 2 since this endpoint was assessed and for mathematical completeness.

Reviewer #3:

This study examined the use of virtual transvaginal ultrasound performed by patient compared to in clinic transvaginal ultrasound performed by the UT with the primary outcome of image quality

and AFC. This study is well designed to understand the acceptability of this innovative approach to the patient and physician.

Response: Thank you for your feedback on our manuscript. We appreciate your thoughtful engagement and comments and hope we address them adequately below.

Methods
Page 8 - the authors should provide more information on the vaginal probe including MHz, dimensions etc. Were any modification made to the probe to improve ease of use for the patient?

Response: The Clarius EC7 HD scanned operates at a frequency of 3-10 MHz with a maximum depth of 15cm though the screen base setting was set to 8cm. Its size (328 x 78 x 38mm) and weight (410g) allows for single-handed use, and it was used without any modifications to the stock device. This was included per your feedback in the Methods section (see Lines 127-135).

> Clarius Mobile Health is a healthcare company based in British Columbia. The Clarius Ultrasound Scanner (K192107) is a cordless, wireless app-based portable transvaginal ultrasound system for medical professionals with an existing 510(k) FDA clearance for in-clinic use[23]. It operates at a frequency of 3-10 MHz. The trial used a base setting of 8 cm for depth. The probe size (328 x 78 x 38mm) and weight (410g) allowed for single-handed use. Each probe (model EC7 HD) was paired with an iPhone X® with the Clarius Ultrasound App installed. The Clarius Live feature was enabled, allowing for "tele-ultrasound imaging." The probe was used under the "IVF" application setting without modification and as described by the manufacturer's instructions, with the exception of the novel user (the patient) and setting (home).

Page 9 - exclusion criteria - I am surprised that vaginismus was not an exclusion criterion.

Response: Thank you for this comment. Vaginismus was not an exclusion criteria, but subjects were counseled by the PI/Sub-I that participation was entirely voluntary and had the potential to be unpleasant. One subject had vaginismus and reported that the at-home setting was preferable for her. In general, the goal was to restrict the population as little as possible to maintain real-world generalizability (as, for instance, women with vaginismus may desire fertility information as well). See Lines 146-148, 401-405.

> Criteria were kept intentionally broad to reflect the scope of women who might be interested in participating, excluding only for specific medical or regulatory concerns.
>
> Notably, the trial population reflected individuals with disabilities including a visually impaired subject and another with a history of vaginismus and prior challenges with in-clinic care. The diversity of our trial population demonstrates the broad, proactive interest of reproductive-aged persons to pursue self-evaluation of ovarian reserve.

Also, absence of an ovary should have been an exclusion criterion when each organ is being evaluated in the analysis.

You bring out an interesting point to definitely consider. However, as women with prior unilateral oophorectomy are evaluated for ovarian reserve prior to intended egg freezing or embryo banking, our investigators felt it was appropriate to include the subject in the trial. The trial goal was to maintain real-world generalizability, and including a range of patients likely to be interested in ovarian reserve assessment was important to that end. Practically, she simply contributed four randomization events rather than six, leading to a slightly different number of total organs in the trial but no other statistical impact.

8 subjects did not participate after consenting and the authors note that this was for personal reasons. Please elaborate these reasons.

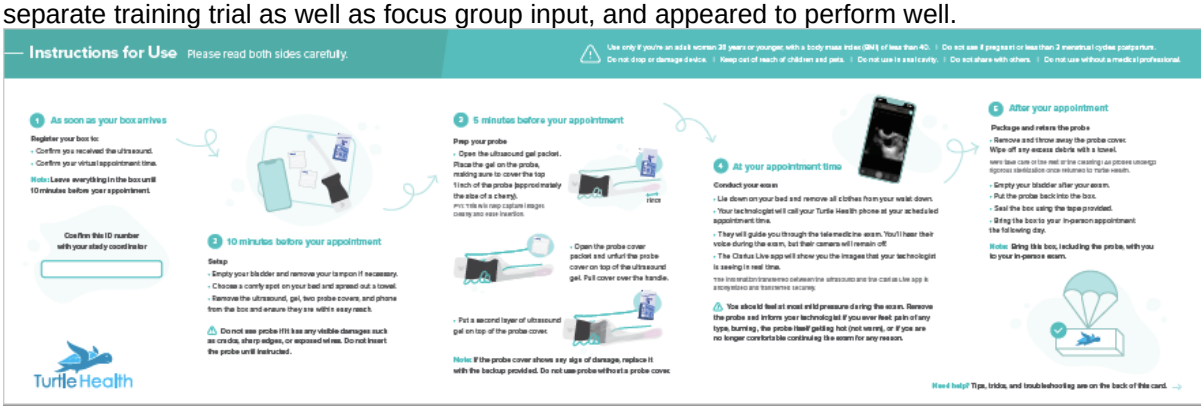
As for the 8 subjects who did not end up participating after consent, the breakdown behind this is provided here: Four patients stopped responding to trial coordinator emails. One patient had a family

emergency requiring a departure from the trial catchment area at short (three days) notice, for the duration of the trial. One patient had scheduling conflicts preventing completing participation in a timely manner, despite trial efforts to accommodate. One requested withdrawal for unspecified reasons. One disclosed material new information to the trial coordinator subsequent to signing her informed consent form that made her no longer an appropriate candidate for the trial (a mid-cycle birth control change of uncertain impact on risk of subclinical pregnancy; given the PI's recommendation for an additional cycle to reach steady state and exclude an incidental pregnancy, the trial would have needed to significantly delay the last patient last visit to accommodate, which was not feasible). We have updated this section accordingly in the paper (see Lines 158-160):

> Of the other eight, four stopped responding, one left the trial catchment area, one withdrew for unspecified reasons, one was unable to schedule, and one disclosed disqualifying medical information after consent.

Page 9 - what instructions were provided to the patients on how to perform the ultrasound? This can be added as a supplement.

Response: Patient instructions were appended as a supplementary PDF, as requested (see representative image below). Full size is 5 x 16 inches. These instructions were developed through a separate training trial as well as focus group input, and appeared to perform well.



On what day of the cycle were the scans performed?

Response: Those with IUDs or oligo/amenorrhea were scheduled at any time of their cycle. Those with regular menstrual intervals were scheduled between days 3 and 10 for risk mitigation against potential early fetal exposure (due to the fact that from a regulatory perspective, evaluation of home fetal exposure was beyond the scope of this trial). Studies, including Petersen *et al* evaluated AFC at any point during the cycle as it is less cycle-dependent than other measures. We took this general principle, but restricted only as needed for regulatory risk mitigation purposes. This was further clarified in the manuscript (see lines 160-163)

> In order to reduce the risk of inadvertent fetal exposure, a regulatory concern to the FDA[25], women with regular menses were scheduled between days 3 and 10 of their menstrual cycle. Women with oligomenorrhea or amenorrhea were scheduled on consecutive days as well but at any time in their cycle.

What materials were supplied in the package i.e. gel, probe cover etc

Response: The package contained: two gel packets and two probe covers (including a backup of each), ultrasound, iPhone, chargers for both (only if needed), instructions, and a Turtle logo sticker. This was included in the manuscript (see lines 164-166)

> Prior to the exam, participants were shipped a Turtle Health package that included a Clarius TVUS probe, an iPhone, two lubricant packets, two probe covers, chargers for both probe and iPhone, instructions and a Turtle logo sticker.

What training was provided to the UT to guide the patients and acquire optimal images.

Response: The UTs underwent a proprietary training program, including study materials, simulated exercises, and a separate precursor trial with live patient exams. Extensive training, guided by both the PI on technical feedback as well as Turtle Health on patient experience aspects, is a critical success factor in remote scanning. This was included in the manuscript (see lines 168-170)

> Prior to the study, these UTs underwent a proprietary training program, including study materials, simulated exercises, and a separate precursor trial with live patient exams.

Did any patient have a retroverted uterus?

Response: We did not specifically ask our raters to identify the uterine position as anteverted or retroverted but on review of the scans, there were certainly a small number of retroverted uteruses. The majority was anteverted as expected.

Were any ovaries difficult to visualize due to bowel gas?

Response: A small number of ovaries were difficult to visualize due to bowel gas. This is reflected in the scoring, with 3 (3)% of ovaries rated not clinical quality in-clinic and 6 (5%) rated not-clinical quality at-home, with some, but not all of these rating failures due to bowel gas. This is included within the numbers cited in the primary endpoint results discussion. Frequently, suboptimal visualization due to bowel gas can be sufficiently improved by the patient applying downward abdominal pressure in the relevant location under the guidance of the UT (see Line 172-174).

> The UTs used key commands to guide the participant through the exam, including: “insert the probe, “go slower,” “go faster,” “stop,” “rotate the probe 90 degrees to your left,” “move your hand to your left,” “press down [if bowel gas seen],” “move your hand to your right,” and “move the probe up and down, floor to ceiling.”

Page 10 - US in the office was performed using the same probe. Was it attached to a regular office US machine or were the images sent wirelessly?

Response: The same configuration was used in-clinic as at home to avoid any source of bias. Clips were recorded and stored on the paired iPhone until they were uploaded and randomized. The probe was not attached - nor does it have the capability to be attached - to a regular office ultrasound machine. It is a cordless, hand-held device.

STATISTICS EDITOR COMMENTS:

Abstract: Need to format the primary outcome and its evaluation in Results in terms of non-inferiority. That is, the reader cannot tell from the Abstract what non-inferiority difference was posited and therefore how that difference related to the difference in clinical quality of the two cohorts. The conclusion needs a re-write: the non-inferiority test was only regarding the quality of the U/S images, not specifically for assessing a metric of ovarian reserve.

Response: We have added the non-inferiority and equivalence margins to the results section of the abstract, as suggested (see Line 39-41). We have also removed the reference to ovarian reserve in factor of specifically citing study endpoints in the abstract’s conclusion (see Line 45-46).

> Ninety-six percent of virtual and 98% of in-clinic images met “clinical quality.” The difference of -2.4% (97.5% CI lower bound: -5.5%), was within the noninferiority margin (18%).
>
> Virtual TVUS remotely guided by a UT is non-inferior to in-clinic TVUS for producing clinical quality images and is equivalent for estimating AFC.

Lines 218-222 and reference # 18: The reference cited does suggest using alpha = 0.025 for inference testing, but does not stipulate or even mention 18% as the margin, so the statement is incomplete. What was the basis for asserting that a difference of as much as 18% in clinical quality would be acceptable? Seems that the margin is wide, even by the Authors' admission on lines 224-228, hence the study would require larger samples. There is a potential downside to having a higher proportion of unacceptable quality studies.

Response: This Statistical Methods section has been largely re-written in response to this and related questions regarding expected study power. The primary justification for the 18% margin is the clinical judgment of physicians about the minimum clinical quality rate for such a novel device to have clinical utility. FDA's guidance suggests that: "There can be flexibility in selecting the M2 margin, choosing a wider margin, for example, when: 1. The primary endpoint does not involve an irreversible outcome such as death (in general, the M2 margin will be more stringent when treatment failure results in an irreversible outcome) 2. The test product is associated with fewer serious adverse effects or better tolerability than other therapies already available 3. The test product has another advantage over available therapies that warrants use of a less stringent margin (M2)." Given the lack of irreversible outcome, significantly better tolerability (as measured by NPS), and the unique advantage remote imaging has over in-clinic in terms of access, this device meets all three cases justifying a wider non-inferiority margin. Of note, the study was 85% powered to demonstrate non-inferiority with a much more conservative 10% margin under reasonable assumptions. The authors, however, did not feel such conservatism was required under the circumstances. See Lines 242-254.

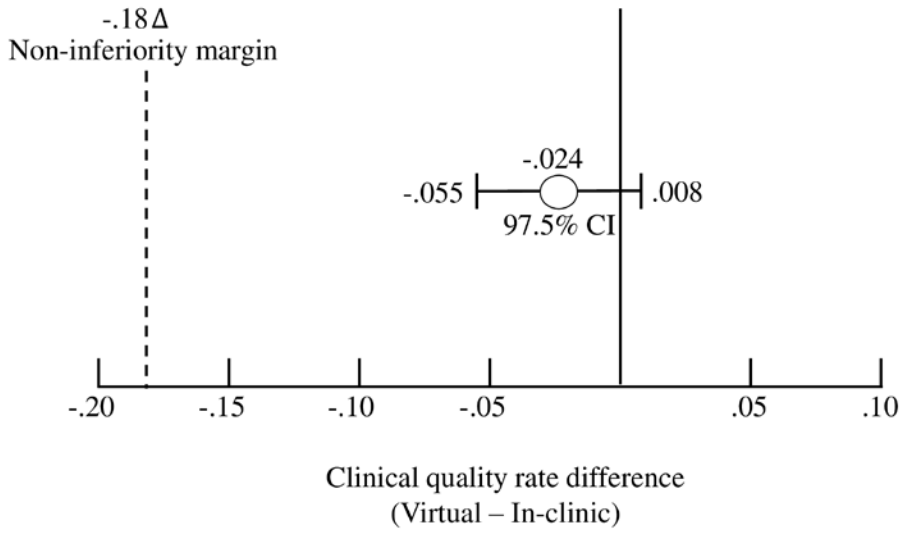
> Sample size was selected to ensure >90% power to demonstrate non-inferiority of the clinical quality image rate, given a non-inferiority margin of 18%, a true margin of up to 7%, a 95% clinical quality image rate in-clinic, an adjustment for up to a 60% intra-patient correlation between organ images, and an alpha of 2.5%. This alpha was chosen based on convention for one-sided tests, utilizing FDA guidance[19]. AFC equivalency was well-powered (>90%) under reasonable assumptions.
>
> Power calculations for sample size and confidence intervals were based on a non-inferiority test for paired binary data using a RMLE-based test-statistic[19,22]. The non-inferiority margin was chosen utilizing FDA guidance that the maximum acceptable margin (18%) is justified in cases of major improvements in tolerability, access to care, and lack of irreversible outcomes[19,36]. In practice, such a generous margin proved unnecessary—the trial was still >85% powered to demonstrate non-inferiority at a 10% non-inferiority margin, given the observed delta of 2.4%

Table 1: Need units for age.

Response: Thank you—this has been added to Table 1.

Table 2: Need to clearly separate the primary from the secondary outcomes. Only the primary was factored into the sample size calculation, the secondaries were not. Regardless of whether the secondary outcomes were statistically significant, they are of interest, but second tier compared to the primary outcome. It would also be preferable to format the primary outcome in the usual graphical display of a non-inferiority study, so that the reader can easily see how the difference between the cohorts compares with the non-inferiority margin.

Response: Thank you for this comment. Primary and secondary endpoint sections have been added to the table as recommended. The recommended graphical display of the noninferiority study was also created and attached as Figure 2 (see representative image below). We have added into the text a reference to Figure 2 for the readers.



EDITORIAL OFFICE COMMENTS:

1. The Editors of Obstetrics & Gynecology have increased transparency around its peer-review process, in line with efforts to do so in international biomedical peer review publishing. If your article is accepted, we will be posting this revision letter as supplemental digital content to the published article online. Additionally, unless you choose to opt out, we will also be including your point-by-point response to the revision letter. If you opt out of including your response, only the revision letter will be posted. Please reply to this letter with one of two responses:

Response: A. OPT-IN: Yes, please publish my point-by-point response letter.

2. When you submit your revised manuscript, please make the following edits to ensure your submission contains the required information that was previously omitted for the initial double-blind peer review:

*** Include your title page information in the main manuscript file. The title page should appear as the first page of the document. Add any previously omitted Acknowledgements (ie, meeting presentations, preprint DOIs, assistance from non-byline authors).**

*** Funding information (ie, grant numbers or industry support statements) should be disclosed on the title page and in the body text. For industry-sponsored studies, the Role of the Funding Source section should be included in the body text of the manuscript.**

*** Include clinical trial registration numbers, PROSPERO registration numbers, or URLs at the end of the abstract (if applicable).**

*** Name the IRB or Ethics Committee institution in the Methods section (if applicable).**

*** Add any information about the specific location of the study (ie, city, state, or country), if necessary for context.**

Response: These edits were completed as instructed.

3. Obstetrics & Gynecology uses an "electronic Copyright Transfer Agreement" (eCTA), which must be completed by all authors. When you uploaded your manuscript, each co-author received an email with the subject, "Please verify your authorship for a submission to Obstetrics & Gynecology." Please check with your coauthors to confirm that they received and completed this

form, and that the disclosures listed in their eCTA are included on the manuscript's title page.

Response: We have confirmed with our coauthors that they have received and verified their authorship and disclosures.

4. For studies that report on the topic of race or include it as a variable, authors must provide an explanation in the manuscript of who classified individuals' race, ethnicity, or both, the classifications used, and whether the options were defined by the investigator or the participant. In addition, the reasons that race/ethnicity were assessed in the study also should be described (eg, in the Methods section and/or in table footnotes). Race/ethnicity must have been collected in a formal or validated way. If it was not, it should be omitted. Authors must enumerate all missing data regarding race and ethnicity as in some cases, missing data may comprise a high enough proportion that it compromises statistical precision and bias of analyses by race.

Use "Black" and "White" (capitalized) when used to refer to racial categories. The nonspecific category of "Other" is a convenience grouping/label that should be avoided, unless it was a prespecified formal category in a database or research instrument. If you use "Other" in your study, please add detail to the manuscript to describe which patients were included in that category.

Response: Race/ethnicity was not collected in a validated way and will thus be omitted from the text of the manuscript and from Table 1.

5. Clinical trials submitted to the journal as of July 1, 2018, must include a data sharing statement. The statement should indicate 1) whether individual deidentified participant data (including data dictionaries) will be shared; 2) what data in particular will be shared; 3) whether additional, related documents will be available (eg, study protocol, statistical analysis plan, etc.); 4) when the data will become available and for how long; and 5) by what access criteria data will be shared (including with whom, for what types of analyses, and by what mechanism). Responses to the five bullet points should be provided in a box at the end of the article (after the References section).

Response: This was completed and all 5 responses were included at the end of the article after the References in a Text Box.

6. Standard obstetric and gynecology data definitions have been developed through the reVITALize initiative, which was convened by the American College of Obstetricians and Gynecologists and the members of the Women's Health Registry Alliance. Obstetrics & Gynecology has adopted the use of the reVITALize definitions. Please access the obstetric data definitions at https://urldefense.com/v3/__https://www.acog.org/practice-management/health-it-and-clinical-informatics/revitalize-obstetrics-data-definitions__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQVrZMEB_e$ and the gynecology data definitions at https://urldefense.com/v3/__https://www.acog.org/practice-management/health-it-and-clinical-informatics/revitalize-gynecology-data-definitions__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQVu9NySEr$. If use of the reVITALize definitions is problematic, please discuss this in your point-by-point response to this letter.

Response: None of the reVITALize definitions are problematic for our paper.

7. Because of space limitations, it is important that your revised manuscript adhere to the following length restrictions by manuscript type: Original Research should not exceed 5,500 words. Stated word limits include the title page, précis, abstract, text, tables, boxes, and figure legends, but exclude references.

Response: Our paper adheres to the 5500-word count limit. It is below that limit at this time at 5444

words, excluding the Data Sharing Statement and the References.

8. Specific rules govern the use of acknowledgments in the journal. Please note the following guidelines:

*** All financial support of the study must be acknowledged.**
*** Any and all manuscript preparation assistance, including but not limited to topic development, data collection, analysis, writing, or editorial assistance, must be disclosed in the acknowledgments. Such acknowledgments must identify the entities that provided and paid for this assistance, whether directly or indirectly.**
*** All persons who contributed to the work reported in the manuscript, but not sufficiently to be authors, must be acknowledged. Written permission must be obtained from all individuals named in the acknowledgments, as readers may infer their endorsement of the data and conclusions. Please note that your response in the journal's electronic author form verifies that permission has been obtained from all named persons.**
*** If all or part of the paper was presented at the Annual Clinical and Scientific Meeting of the American College of Obstetricians and Gynecologists or at any other organizational meeting, that presentation should be noted (include the exact dates and location of the meeting).**
*** If your manuscript was uploaded to a preprint server prior to submitting your manuscript to Obstetrics & Gynecology, add the following statement to your title page: "Before submission to Obstetrics & Gynecology, this article was posted to a preprint server at: [URL]."**

Response: All acknowledgements were recognized and received written permission, and our oral presentation of this abstract version at a national meeting was disclosed.

9. The most common deficiency in revised manuscripts involves the abstract. Be sure there are no inconsistencies between the Abstract and the manuscript, and that the Abstract has a clear conclusion statement based on the results found in the paper. Make sure that the abstract does not contain information that does not appear in the body text. If you submit a revision, please check the abstract carefully.

In addition, the abstract length should follow journal guidelines. The word limit for Original Research articles is 300 words; Reviews is 300 words; Case Reports is 125 words; Current Commentary articles is 250 words; Executive Summaries, Consensus Statements, and Guidelines are 250 words; Clinical Practice and Quality is 300 words; Procedures and Instruments is 200 words. Please provide a word count.

Response: These edits were completed as instructed. The abstract word count is 263 words including headers.

10. Abstracts for all randomized, controlled trials should be structured according to the journal's standard format. The Methods section should include the primary outcome and sample size justification. The Results section should begin with the dates of enrollment to the study, a description of demographics, and the primary outcome analysis. Please review the sample abstract that is located online here: https://urldefense.com/v3/__http://edmgr.ovid.com/ong/accounts/sampleabstract_RCT.pdf__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQVvNIk3Kw$. Please edit your abstract as needed.

Response: These edits were completed as instructed.

11. Only standard abbreviations and acronyms are allowed. A selected list is available online at https://urldefense.com/v3/__http://edmgr.ovid.com/ong/accounts/abbreviations.pdf__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQVhTv0rP5$. Abbreviations and acronyms cannot be used in the title or précis. Abbreviations and acronyms

must be spelled out the first time they are used in the abstract and again in the body of the manuscript.

Response: This was double-checked.

12. The journal does not use the virgule symbol (/) in sentences with words. Please rephrase your text to avoid using "and/or," or similar constructions throughout the text. You may retain this symbol if you are using it to express data or a measurement.

Response: All (/) were removed from the manuscript.

13. ACOG avoids using "provider." Please replace "provider" throughout your paper with either a specific term that defines the group to which are referring (for example, "physicians," "nurses," etc.), or use "health care professional" if a specific term is not applicable.

Response: We have replaced throughout, with the one exception of its use in the title of the registered trial: "SELF-HELP Validation Study: Sonograms Enable Looking Forward – Home Examinations Led by Providers."

14. In your Abstract, manuscript Results sections, and tables, the preferred citation should be in terms of an effect size, such as odds ratio or relative risk or the mean difference of a variable between two groups, expressed with appropriate confidence intervals. When such syntax is used, the P value has only secondary importance and often can be omitted or noted as footnotes in a Table format. Putting the results in the form of an effect size makes the result of the statistical test more clinically relevant and gives better context than citing P values alone.

If appropriate, please include number needed to treat for benefits (NNTb) or harm (NNTh). When comparing two procedures, please express the outcome of the comparison in U.S. dollar amounts.

Please standardize the presentation of your data throughout the manuscript submission. For P values, do not exceed three decimal places (for example, "P = .001"). For percentages, do not exceed one decimal place (for example, 11.1%").

Response: We have endeavored to follow this guidance, but have retained p-values where requested in the reviewer comments above, or where they proved critical for context. NNT does not seem applicable to our study. All *P* values were standardized throughout the submission.

15. Your manuscript contains a priority claim. We discourage claims of first reports since they are often difficult to prove. How do you know this is the first report? If this is based on a systematic search of the literature, that search should be described in the text (search engine, search terms, date range of search, and languages encompassed by the search). If it is not based on a systematic search but only on your level of awareness, it is not a claim we permit.

Response: This point is well taken. We do believe we have a priority claim, but we have softened the language significantly to address this feedback. Our understanding of this claim is based on the following points:

1. We have conducted several literature searches over the last two years, most recently on 11/18/2021. While we have identified a number of other home transvaginal studies largely funded by Sonaura/Fertihome, none of them were both (a) statistically powered and (b) focused on anything other than the assessment of women undergoing IVF cycle monitoring (i.e., provided extensive training for the monitoring of potentially enlarged ovaries, a very different use case, population, and process). Thus, we did not locate any articles that stated or implied a first claim of ability to show – on a statistical, powered basis – non-inferiority of home scanning regarding clinical image quality and AFC estimation.

2. In the Sponsor's interactions with FDA over the past one and a half years, FDA has consistently indicated that they are not aware of others who have demonstrated a similar type of claim. This was evidenced by FDA recommending the Sponsor file for De Novo (i.e., first-in-class) status, and then subsequently accepting De Novo filing 210021 from Turtle Health without sending the filing to internal 510(k) vs. De Novo arbitration. The Sponsor is able to share documentation of the De Novo filing acceptance with *Obstetrics and Gynecology*, if helpful to verify this point. Thus, unless there is unpublished research, which also has not been submitted to FDA, we are fairly confident to express that we believe we likely have a priority claim, while leaving some flexibility in the language in case we are incorrect. We have modified the language in the article (see Lines 369-377) to reflect our rationale for this understanding. If this justification is insufficient, we understand this priority claim may be struck.

> To our knowledge, this is one of the first studies to demonstrate that virtual TVUS evaluation of antral follicle ovarian reserve assessment performed at-home by a patient is clinically non-inferior to in-clinic examination. A literature search as of November 11, 2021 of comprehensive databases such as Science Direct, PubMed and Google Scholar using the search terms “self-administered ultrasound”, “self-performed ultrasound”, “at-home ultrasound”, or “patient performed ultrasound” AND “vaginal”, along with “antral follicle” and “ovarian reserve” yielded no relevant results. No such studies have been completed for healthy patients not actively undergoing infertility treatment. Additionally, FDA accepted a De Novo filing for the device in June of 2021, indicating no existing predicated device with this intended use.

16. Please review the journal's Table Checklist to make sure that your tables conform to journal style. The Table Checklist is available online here: https://urldefense.com/v3/__http://edmgr.ovid.com/ong/accounts/table_checklist.pdf__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQVu1ntAIa$.

Response: The table checklist was reviewed and edits made appropriately to our tables.

17. Please review examples of our current reference style at https://urldefense.com/v3/__http://ong.editorialmanager.com__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQViXMmsIE$ (click on the Home button in the Menu bar and then "Reference Formatting Instructions" document under "Files and Resources). Include the digital object identifier (DOI) with any journal article references and an accessed date with website references. Unpublished data, in-press items, personal communications, letters to the editor, theses, package inserts, submissions, meeting presentations, and abstracts may be included in the text but not in the reference list.

In addition, the American College of Obstetricians and Gynecologists' (ACOG) documents are frequently updated. These documents may be withdrawn and replaced with newer, revised versions. If you cite ACOG documents in your manuscript, be sure the references you are citing are still current and available. Check the Clinical Guidance page at https://urldefense.com/v3/__https://www.acog.org/clinical__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQVjw9q7m8$ (click on "Clinical Guidance" at the top). If the reference is still available on the site and isn't listed as "Withdrawn," it's still a current document.

If the reference you are citing has been updated and replaced by a newer version, please ensure that the new version supports whatever statement you are making in your manuscript and then update your reference list accordingly (exceptions could include manuscripts that address items of historical interest). If the reference you are citing has been withdrawn with no clear replacement, please contact the editorial office for assistance (obgyn@greenjournal.org). In most cases, if an ACOG document has been withdrawn, it should not be referenced in your manuscript.

Response: References have been checked and are consistent with the guidelines.

18. When you submit your revision, art saved in a digital format should accompany it. If your figure was created in Microsoft Word, Microsoft Excel, or Microsoft PowerPoint formats, please

submit your original source file. Image files should not be copied and pasted into Microsoft Word or Microsoft PowerPoint.

When you submit your revision, art saved in a digital format should accompany it. Please upload each figure as a separate file to Editorial Manager (do not embed the figure in your manuscript file).

If the figures were created using a statistical program (eg, STATA, SPSS, SAS), please submit PDF or EPS files generated directly from the statistical program.

Figures should be saved as high-resolution TIFF files. The minimum requirements for resolution are 300 dpi for color or black and white photographs, and 600 dpi for images containing a photograph with text labeling or thin lines.

Art that is low resolution, digitized, adapted from slides, or downloaded from the Internet may not reproduce.

Response: Both figures are high-resolution TIFF files.

19. Authors whose manuscripts have been accepted for publication have the option to pay an article processing charge and publish open access. With this choice, articles are made freely available online immediately upon publication. An information sheet is available at https://urldefense.com/v3/__http://links.lww.com/LWW-ES/A48__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQVlfvBGTn$. The cost for publishing an article as open access can be found at https://urldefense.com/v3/__https://wkauthorservices.editage.com/open-access/hybrid.html__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQVo5Yi1ag$.

If your article is accepted, you will receive an email from the editorial office asking you to choose a publication route (traditional or open access). Please keep an eye out for that future email and be sure to respond to it promptly.

If you choose open access, you will receive an Open Access Publication Charge letter from the Journal's Publisher, Wolters Kluwer, and instructions on how to submit any open access charges. The email will be from publicationservices@copyright.com with the subject line, "Please Submit Your Open Access Article Publication Charge(s)." Please complete payment of the Open Access charges within 48 hours of receipt.

Response: Thank you for this advice—the authors would be pleased to choose open access to widely share this manuscript.

If you choose to revise your manuscript, please submit your revision through Editorial Manager at https://urldefense.com/v3/__http://ong.editorialmanager.com__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQViXMmsIE$. Your manuscript should be uploaded as a Microsoft Word document. Your revision's cover letter should include the following:

* A confirmation that you have read the Instructions for Authors (https://urldefense.com/v3/__http://edmgr.ovid.com/ong/accounts/authors.pdf__;!!OToaGQ!-i-TZPczVmjhlN8ndcEdmqSCNI7M9DXEAngHv1kMZWhMtFJWXiti2Pb3PYLQViFriJhl$), and
* A point-by-point response to each of the received comments in this letter. Do not omit your responses to the Editorial Office or Editors' comments.