THE BMJ'S MANUSCRIPT COMMITTEE MEETING

We thought your study addresses an important and interesting research question. We had the following concerns.

1. **Outcomes seem to match between the various documents (paper, protocol, registry and SAP). Having said that, there are a considerable number of outcomes at 4 post baseline time points (e.g., 56 comparisons in Table 2, 32 in Table 3, 24 in Appendix 2, 72 in Appendix 3). You state a-priori that no adjustment for multiple testing – and ultimately most findings are non-significant.**

We agree that we had a considerable number of outcomes and statistical tests, which could have been condensed in hindsight. However, most of these statistical tests tested a different hypothesis (i.e. a different pathway through which our program could affect fall risk), which would not warrant correction for repeat testing.

2. ***Abstract: "Both groups had a similar rate of falls and proportion of fallers at 12-months (p=0.071 and p=0.461 respectively)"; we would like to see this quantified and not just p-values – what was the rate of falls in both arms or the number of falls per arm?***

We now provide these rates and proportions in the abstract (line 106-109).

"The rate of falls (0.60 (SD 1.05) falls per year in IG and 0.76 (SD 1.25) in CG; incidence rate ratio, IRR:0.84; 95% confidence interval, 95%CI:0.62-1.13; p=0.071) and proportion of fallers (34.6% in IG and 40.2% in CG; relative risk, RR:0.90, 95%CI: 0.67-1.20, p=0.461) were not statistically different at 12-months."

We have also added the proportion of fallers in the intervention (IG) and control (CG) groups to the results (line 438-439).

"Overall, 188 participants (37.4% in total; 34.6% in IG and 40.2% in CG) fell at least once in the 12-month follow-up."

3. ***We would appreciate a comment on the switch of analysis from negative binomial (specified in the protocol) to a Poisson model (in the SAP and paper) - the switch conveniently changes the finding at 2 years. Furthermore, the results in Figure 2 are based on a mixture of Poisson and negative binomial (e.g., rate of falls at 12m 0.72; 95% CI 0.66 to 1.02 is NB, whilst rate of falls at 24m is 0.84; 95% CI 0.72 to 0.98 is Poisson).***

We have now added that the decision to perform Poisson regression was a-priori registered in our statistical analysis plan but not our protocol paper (line 379). The decision to use Poisson regression was due to new insights with respect to methodology of the CACE analysis and, as can be seen in our OSF registration dates, made prior to completion of data collection. We have updated figure 2 and corrected the errors in the results section (due to copying the wrong results). Thank you for identifying this.

"Poisson regression was selected over negative binomial regression (as a-priori registered in our statistical analysis plan, but not in our protocol paper) to allow for a direct comparison to our planned complier average causal effects analysis since the latter was based on a Poisson model."

4. ***The sample size – as reported in the paper – is not reproducible nor particularly informative. The protocol provides much more detail, including assumptions, effect size (based on the primary outcome IRR) etc, but still not fully reproducible. So this needs addressing and inserting into the manuscript. Please describe the sample size calculation using an outcome and effect that have intuitive meanings to a clinician. This will help put in context the formal result of failure to detect this difference. The present "33% effect" will mean little to most people.***

We have updated this section based on these comments (line 331-343).

*"Based on previous evidence, we carried out an a-priori sample size calculation (8) in Stata using a custom code with 5000 simulations, which showed that 500 participants were required to achieve 80% power to find a fall rate reduction of 33% (i.e. an incidence rate ratio of 0.67) in the intervention vs. control group that is statistically significant at $p<0.05$, considering an overdispersion of 1.2, 0.8 falls/person-year in the control group, and a follow up duration of 22-months to account for 20% dropout (8). We then ran power calculations in G*Power V3.1.7 for our secondary outcomes, considering an ANOVA design with 4 measurements and 20% dropout. These showed that we would have 90% power to detect a statistically significant ($p<0.05$) small reduction (i.e. effect size $f=0.15$) in concern about falling in the intervention vs. control group assuming a within-subject correlation of 0.75 (8). A subsample of 200 participants with repeat physical assessments, would provide us with a 95% power to detect a statistically significant ($p<0.05$) large reduction (i.e. effect size $f=0.38$) in postural sway in the intervention vs. control group, considering a within-subject correlation of 0.76 (8)."*

5. ***Might you provide more detail about the intervention, so that others could reproduce it if they wanted to.***

We have now added that the exercises focused on standing balance, targeted stepping and step-up box exercises. The intervention is detailed in the study protocol and provided via an app, which will be available on our website once released to the public. We have now added these details to the methods section.

"The exercise focuses on standing balance, targeted stepping and step-up (box) exercises. More information about the program can be found in the study protocol (8) and via https://www.standingtall.org.au/."

6. ***Most authors (including corresponding) are affiliated with Neuroscience Research Australia which commercially markets "the PPA (NeuRA FallScreen)". This seems to be an assessment tool, not the intervention app (https://www.neura.edu.au/research-clinic/fbrg/). We would like some clarification whether the organization has rights and plans to commercial the app itself. Along these lines, is "Standing Tall" a trademark for a commercial program? If so, we would like to ask you to remove it from the title of the paper.***

StandingTall is the name of the trial as well as that of the app that we have developed to provide the intervention through. It is not a trademark for a commercial program. NeuRA does have the rights to commercialise the app; however, to date there are no concrete plans for commercialisation. We have kept *StandingTall* in the title of the paper, similar to in the protocol paper, as this is how other researchers and participants are familiar with the trial.

REVIEWER 1

This article is very well written and it's subject is highly relevant and appropriate to community level strategies in the prevention of falls on older people. This is especially so in light of the pandemic and reduced social mobility within communities. Several areas could be highlighted and considered important information;

Thank you and we agree that this work can have even more impact due to the COVID19 pandemic.

7. ***The recruitment of participants was based on advertisement and word of mouth. Information on how widespread in terms of locality, areas covered and medium of advertisement would allude to whether there was a wider inclusion of different types of communities into the study.***

We have now added the average distance from NeuRA and specified that the advertisements were printed in local newspapers (line 202-204).

"We recruited community-living older people in the Sydney metropolitan area via flyers, printed advertisements in local newspapers, presentations at residential and community senior centres, and word of mouth. Study participants lived on average 12 km, with a range of 1.2 - 46.9 km, from NeuRA at Randwick NSW."

8. ***Was there a difference in sociodemographic background of subjects? This is raised as often respondents to such a trial often belong to a similar or homogeneous population.***

The descriptive characteristics of the participants are provided in Table 1. We did not obtain specific SES scales but given the locality and the relatively high number of years of education and percentage of computer owners, we assume that our sample was relatively affluent. We have now added this as a limitation (line 572-574).

"Fourthly, our cohort of community-dwelling older people was highly educated, had a high percentage of computer ownership and lived in more affluent areas of Sydney; our results may not generalise to usage in more rural or less affluent areas."

9. ***Were all the participants assessed for their ability or digital literacy to use the digital platform independently or otherwise? If so how was this assessed?***

Participants were not assessed for their digital literacy or ability to use technology. They were provided with manuals and training to use the iPad and program components. We have now added this to the methods section (line 235-237).

"Participants received a manual on how to use the tablet computer, and were guided to the basic features of the tablet computer and health promotion education programme after their baseline assessment."

10. Did participants also perform other types of exercises of outdoor activities?

Participants were asked to perform the *StandingTall* exercise in addition to their usual activities, which would comprise outdoor activities. Our IPEQ planned exercise outcome showed that both groups did about 3 hrs/wk of planned exercise, with a slight but not statistically significant 0.1-0.9 hrs/wk increase in the intervention compared to the control group (Table 2).

11. The usability and age appropriateness was tested on a group of older people; should include further details such as sample population numbers, mean age, gender etc.

We have now provided further details on this group (line 194-198).

"StandingTall was developed using consumer design principles. A group of older people were involved during the development of the StandingTall application. They were asked to evaluate an early version on its usability and age appropriateness as a means to engage in fall prevention exercises using tablet-based technology. A two-week feasibility study was conducted in 10 community-dwelling older people in November 2013. The average age of the participants was 77.5 years (range 67-82), and 6 participants were female. PPA scores ranged from 'mild' to 'marked' (median z score 1.68; range 0.79-2.94) and 7 participants had experienced falls in the previous year. Adherence was high with participants reporting the program was suitable for older people."

REVIEWER 2

This study examined an important area of fall prevention and fall management. Using e-health home-based programme is innovative. It has a large sample size. Please find below comments for authors to address.

Thank you for these comments.

12. Introduction: Please add a hypothesis of the study

We have now added the hypothesis to the introduction (line 174-176).

"We hypothesised that StandingTall would lead to a fall rate reduction, when compared to a control group with minimal intervention."

13. Methods: For randomization, elaborate which web-based randomization programme was used to implement random allocation sequence. It is indicated in Page 8, the allocation ratio is 1:1, why the subject number is not identical in IG and Con group (i.e. IG=254 vs CON=249)

Block randomisation, which we used in the current study, can lead to slightly unequal group sizes when the last block is not filled completely. We further randomised couples and singles separately to avoid contamination, which could have also contributed to the unequal group sizes. The web-based program is currently only available to NeuRA staff (a login in required), we have updated its description (line 220-222).

"Allocation was performed centrally using a ~~web-based~~ custom randomisation programme by an investigator not involved in participant assessments or delivery of the intervention."

14. Methods: Training protocol – when compared with control group, intervention group had an additional exercise time of 120 minutes/week and was given exercise equipment. The improvement in outcomes could simply be attributed to the great placebo effect and additional treatment time, regardless of which mode of exercise delivery. Please comment.

In the current study, we were interested in the effect of the *StandingTall* balance exercise programme on falls and fall risk. We used a health promotion education programme in both the intervention and control group to control for the use of technology and for the method of data collection (i.e. falls during the trial period) through a tablet computer. We agree that the addition of 120 minutes/wk of balance training is likely what resulted in the fall rate reduction. We have clarified this in the manuscript (line 239-244).

"The intervention group received the StandingTall programme, with exercise equipment (foam cushion, stepping box, exercise mat), in addition to the health promotion education programme and usual care, received by both groups. The StandingTall intervention consisted of balance exercises delivered through a tablet computer in the participants' homes with embedded behavioural change techniques, including a weekly calendar for scheduling exercises, goal setting and educational fact sheets."

15. Methods: How the participants registered their exercise duration?

Adherence was automatically obtained via the *StandingTall* balance exercise app, please see line 251.

16. ***Methods: There are many secondary outcomes. For instance, physical measurements included balance, functional mobility and gait assessment, step performance, SPPB – why each of them is needed.***

The secondary outcomes provide insight into different aspects of fall risk and were included to understand the pathways through which the programme would lead to a fall rate reduction. We acknowledge that we have quite a few secondary outcomes, but considering our trial registration and the complementary nature of these outcomes, we need to report on each of them. Please also refer to our answer to Question 1 by the BMJ manuscript committee and limitation three (line 569-572).

17. ***Methods: Justify the use of 3 tests to assess cognitive function - Montreal Cognitive Assessment(17), Trail-Making Tests (TMT)(18), and the Victoria Stroop task(19).***

These tests were used to assess global cognition, set shifting and response inhibition, we have now added this to the methods (line 299-301).

"Cognitive function was assessed with the Montreal Cognitive Assessment (18) for global cognition, Trail-Making Tests (TMT) (19) for set-shifting, and the Victoria Stroop task (20) for response inhibition."

We have also included an overarching justification at the start of this section:

"*Secondary outcome measures were assessed at baseline, at 6-months to examine acute effects, and at 12, 18, and 24-months to examine retention effects. These measures included common fall risk factors: (...)*."

18. ***Methods: Justify the use of 3 scales to measure Health-related quality of life - 12-item WHO Disability Assessment Schedule(23), 5-level EuroQol- 5 Dimension (EQ-5D-5L)(24), and 20-item Assessment of Quality of Life 6-Dimensions (AQoL-6D) questionnaires(25).***

These tests were included to allow for comprehensive cost-effectiveness analyses (to be reported in a future paper). We included the outcomes here because we registered their use.

19. ***Methods: Why lab re-asst was only performed in 226 participants? How to determine that 226 participants for the lab re-asst, who did the selection, blinded?***

This decision was based on both power to detect a difference and cost to undertake these assessments. We have added this information to the methods and revised the sample size section to address this comment.

Subheading - Randomisation and blinding (line 223-224):

"Only the first 226 participants were invited for repeated physical tests to reduce costs and participant' time."

Subheading – sample size calculation (line 339-343):

"A subsample of 200 participants with repeat physical assessments would provide us with a 95% power to detect a large reduction (i.e. effect size f=0.38) in postural sway in the intervention vs. control group that is statistically significant at $p<0.05$, considering a within-subject correlation of 0.76 (8)."

20. ***Results – are there any differences in baseline demographic between those who continued the training and follow-ups vs those who dropped out.***

We agree that this is an interesting analysis, which we have planned to perform in the future. We do feel that it is outside the scope of the current paper and would distract from the main results, and as such have not added it.

21. "We did find a small improvement of 0.03 (95% CI 0.01-0.06) on the EQ-5D-5L utility score at 6-months in IG compared to CG." Comment on clinical significance.

Unfortunately, there is no data for minimally important differences of the EQ-5D-5L in the population of older Australians. There is some simulation-based data, which suggests that the change we found may be clinically important. We have now included this interpretation in the discussion:

"Quality of life measured with the EQ-5D-5L utility index also showed a small, but potentially clinically relevant (37), significant improvement at 6-months, however no significant differences were found at 12 or 24-months."

22. The baseline value of TUG indicates that the participants are fit and have good mobility, why did the investigators target this group?

The control group experienced 0.76 falls/year with over one third falling at least once, indicating that falls are a major issue even in this relatively fit and healthy group.

23. Discussion: The study found no significant improvement in the primary outcomes at 12-month but there was reduction of fall rate and injurious fallers at 24-month. The discussion has to be strengthened by giving some insights on why there was reduction of fall rate and injurious fallers at 24-month despite no improvement in the physical, balance and mobility outcomes. The authors cited "their 20% reduction in the proportion of injurious fallers at 24-months, may be higher than the previously reported 12% reductions (3)". Which part of training programme contributes to the positive findings? Which particular aspect of e-Health makes it effective? Or is it just because of the additional exercise time of the IG when compared to the CG?

The lack of clear findings in our secondary outcomes indicates that we are unable to pinpoint the pathways through which this reduction is achieved. As per our response to Q14, in the current study we were interested in the effect of the eHealth *StandingTall* balance exercise program on falls and fall risk. We show that the program reduces falls but are unable to determine which aspects of the program contributed mostly. As mentioned in the introduction, balance exercise is key for effective fall prevention interventions. We suspect that the use of eHealth increased adherence, and thereby the dose that people participated in, but a different study design would be required to test such a hypothesis.

24. In the conclusion, the investigators claimed that this is a "low resources and low-cost intervention". However, each participant was given a tablet, exercise equipment and exercise mat etc. And human resources are needed to registered the entry regularly. It seems that the cost and resource are not that low. Please comment.

We agree with the reviewer and have changed the sentences accordingly (line 605-607).

"In conclusion, our results show that a tailored e-Health exercise programme is an effective~~, low-resources, and thus low-cost,~~ intervention towards the prevention of falls in older people."

25. More improvement for those with less concerns about falls, and lower physio risk, what is the significance of these findings?

The findings of these a-priori planned subgroup analyses seem to suggest that people with lower fall risk at baseline improve more on physiological fall risk than people with a higher fall risk at baseline, and that people with a higher concern about falling at baseline improve less on concern about falling than people with a lower concern about falling at baseline. While this could indicate that the program is of more benefit to a healthier cohort of older people, our

fall incidence analyses in the same subgroups were not able to confirm this interpretation. We have now added that a sentence to indicate that these are preliminary findings (line 533-537).

"Interestingly, our pre-registered subgroup analyses found no significant modification of falls but did find indications of significant modification of the assessment outcomes at 12-months, in people with lower physiological fall risk and lower concern about falling benefitting more. Further research will be required to confirm effectiveness of StandingTall in older people with an increased risk of falling."

REVIEWER 3

I have read and reviewed the manuscript. In my opinion, the article addresses an important topic and is interesting for readers. Falls matter to clinicians and patients and fall prevention has been addressed in many trials. However, few of them included unsupervised, home-based, e-Health balance exercises. This is intriguing in light of the COVID-19 pandemic. In this perspective, I think this study adds enough to existing knowledge and will be cited.

Strength of this study is the large sample size and the extensive clinical assessment. Also, methodology is fine and results are clearly presented. However, results are not very consistent and the message is less clear than expected. Authors found reduction in rate of falls only at second baselines (24 months) without balance improvement. Moreover, improvements seem to disappear using a different statistical model.

Weakness. Although participants characteristics are adequately described, study inclusion and exclusion criteria are problematic. It seems the study enrolled subjects without balance disorders and falls. Including subjects not in need of a fall prevention program may have weakened the results.

Overall, it seems that inclusion of healthy subjects having lower fall risk, an overoptimistic expected reduction in rate of falls and a lower than planned dosage may have caused lack of balance improvements and not firm results on falls reduction.

The present study tests the effects of a fall prevention treatment in preventing falls in older people. The intervention consists in balance exercises delivered via an App. Results show a reduction in rate of falls two years after baseline assessment without a firm, concomitant improvement in balance.

Thank you for these constructive comments; we believe that by addressing these comments we have improved the manuscript.

26. Introduction: The introduction nicely reports evidences from literature on fall prevention in this population and overall, the research questions are clearly defined. A brief overview on treatment related changes after balance rehabilitation would improve readers understanding of methods and results of the present paper.

In the introduction, we focus on the primary outcome of falls. We feel that the addition of a review on effectiveness of exercise on the secondary outcomes in the introduction does not fit with the primary aims of the study. In response to Q33, we did add further discussion around the effect of balance exercise on balance.

27. Methods: The design of the study is appropriate. Prospective RCT is a typical study design to assess fall prevention interventions. In this section authors should state whether this intervention incorporates international recommendations for fall prevention in this population.

We have now added this to the methods (line 247-248).

"Participants were asked to exercise for at least two hours per week for the duration of the trial, in line with the international recommendations for fall prevention at the time of the study (2)."

28. In addition, authors should add an extra appendix to better explain treatment provided. Is it possible to better explain how balance exercises were tailored? For example providing an algorithm to link subjects' impairments and functional deficits to treatments provided. Likewise, is it possible to explain how exercise difficulty increased over time? This would improve reproducibility.

These aspects were part of the inbuilt algorithms of the application and are not publicly available. More information about the *StandingTall* application can be found on our website (https://www.standingtall.org.au/). We now refer to the study protocol and website for more information (line 244-246).

"More information about the programme can be found in the study protocol (8) and via https://www.standingtall.org.au/."

29. Linear models using generalized least squares are adequate to assess the impact of interventions in longitudinal study since the errors are allowed to be correlated. Negative binomial and Poisson distributions have been used to this purpose. Appendix 1 reports results comparing negative binominal vs Poisson distribution. Results are inconsistent, this suggests results are not so conclusive. A larger sample size is probably needed to get firmer conclusions. Please, provide a comment.

We agree that although the effect sizes are very similar, the p-values for the negative binomial and Poisson regression are not. This decision to use a Poisson distribution was made prior to any analyses. We have included the negative binomial results for completeness but do not consider these our main results.

30. The sample size statement is unclear. What does 33% reduction in incidence rate mean? Is it a 33% reduction compared to the control group or an absolute change? Assuming it is the between group difference 33% seems a quite high percentage. Please comment.

We have now revised the sample size section (line 330-343) and hope to have clarified that it is a reduction in the intervention with respect to the control group. We agree that in hindsight the anticipated effect size was large and will take this into consideration for our future trials. We expect, as the reviewer also alluded to in their introductory statement, that a combination of imperfect adherence, potentially active control intervention (health education program), and an overly ambitious effect size estimation led to a slightly underpowered study, at least at 12 months where the total dose would have been lower than at 24 months.

*"Based on previous evidence, we carried out an a-priori sample size calculation in Stata using a custom code with 5000 simulations, which showed that 500 participants were required to achieve 80% power to find a fall rate reduction of 33% (i.e. incidence rate ratio of 0.67) in the intervention vs. control group that is statistically significant at $p<0.05$, considering an overdispersion of 1.2, 0.8 falls/person-year in the control group, and a follow up duration of 22 months to account for 20% dropout (8). We then ran power calculations in G*Power V3.1.7 for our secondary outcomes, considering an ANOVA design with 4 measurement and 20% dropout. These showed that we have 90% power to detect a small reduction (i.e. effect size $f=0.15$) in concern about falling in the intervention vs. control group that is statistically significant at $p<0.05$ assuming a within-subject correlation of 0.75 (8). A subsample of 200 participants with repeat physical assessments would provide us with a 95% power to detect a large reduction (i.e. effect size $f=0.38$) in postural sway in the intervention vs. control group that is statistically significant at $p<0.05$, considering a within-subject correlation of 0.76 (8)."*

31. Moreover, I do not understand why missing data were not imputed for primary outcome. In general this approach is not recommended (see for example Little et al. The Prevention and Treatment of Missing Data in Clinical Trials. doi:10.1056/NEJMsr1203730.) Authors stated data were missing at random. Chained equations are usually preferred to means single imputation to avoid biased standard errors. Please comment.

There was no need to impute the fall incidence as we used a Poisson regression which takes in to account exposure, i.e. duration of follow up. As stated in the methods section, for the proportion of fallers we assumed that faller status was maintained during censoring. This could have led to a slight underestimation of the number of fallers in both groups, which would be minor given our intensive follow up resulting in a low amount of missing data. We did impute our secondary outcomes using joint multivariate normal imputation (line 365), which is Markov chain Monte Carlo technique. We have now added the reference (HUQUE, M. H., CARLIN, J. B., SIMPSON, J. A. & LEE, K. J. 2018. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC medical research methodology, 18, 1-16.) for the used imputation technique (line 364).

32. Results: Table 1 reports baseline characteristics. Time to a first fall should be reported. In addition, please clearly report % of subjects who did not fill in the falls diary and proportion of injurious fallers in both groups at 24-month FU.

Table 1 reports baseline characteristics, while the requested variables are obtained during follow up. Time to first fall is not an outcome used in the current study and hence not included. The number of people with incomplete follow-up, and thus missing falls diaries, is provided in line 361; we have now added that these cases were distributed evenly between the IG and CG. We have also added the proportion of injurious fallers per group (line 438-439).

"The faller status of people with incomplete follow up (n=66 at 12-months and n=188 at 24-months, distributed evenly over IG and CG) was assumed to be maintained during censoring."

"210 participants (41.7%; 37.4% in IG and 46.2% in CG) experienced an injurious fall during the 24-month follow-up."

33. Discussion could be improved. In specific results do not clearly show the effects of intervention in reducing falls or improving balance. For example between groups differences in rate of falls at 24-months vanished simply using a negative binomial model. Several hypotheses could be explored and mentioned in this section. Is it possible that treatment was provided to subjects not in need of a balance prevention program? Is it possible that better results would be observed recruiting subjects having balance disorders and experiencing 1 or more falls before be beginning of the study? Additionally, is a two hour/week training enough to foster balance changes? This is of importance since results show that only 40% of participants achieved the prescribed dose at 6-months follow up. Only 34% of participants received the prescribed dose at 12 months. This does not reflect trends seen in the results where improvements more evident after 12 months. I think these three elements may be better addressed in the discussion section to suggest improvements for further studies.

A 2011 Cochrane review including 94 studies testing the effects of exercise interventions on balance in older people (Howe TE, Rochester L, Neil F, Skelton DA, Ballinger C. Exercise for improving balance in older people. Cochrane Database of Systematic Reviews 2011, Issue 11. Art. No.: CD004963) showed weak evidence that some types of balance exercise are moderately effective, immediately post intervention, in improving clinical balance outcomes in older people. While a number of research studies using both core and functional balance

outcome measures have shown an effect (on which we based our sample size calculation), the overall evidence is not strong. In addition, the evidence for technology-driven programs has not yet been confirmed through a systematic review. We have now addressed this in the discussion (line 515-518).

"*While a number of research studies testing the effects of exercise interventions on balance in older people have shown an effect), systematic review evidence has suggested that the evidence for a moderate effect is weak(36).*"

Our sensitivity analyses also do not suggest a stronger effect in individuals with an increased risk for falling due to physiological fall risk, fall history, concerns about falling or slower executive function. There is also no evidence from the literature that the effect of balance exercise on falls or balance would be greater in higher risk populations. We therefore assume that, as per our response to Q30, that the lack of significant findings on fall rates at 12 months is due to a lower intervention dose. We do not have any evidence for this assumption and have therefore refrained from including it into the discussion. We do note that we did not reassess balance at timepoints after 12 months, which has now been added as a limitation of this study (line 520-526).

"We did not repeat these assessments at 24-months, where we did see a significant reduction in falls."

34. Again, in the discussion authors stated that "Adherence was good [...]". However, 60% of participants did not receive adequate dose of the intervention. Additionally, the following statement "The high adherence and zero serious adverse events support the feasibility and safety [...] intervention to a population level" seems overreaching. The same is for "StandingTall is a scalable intervention and can be easily implemented into clinical practice". Feasibility is not the aim of this study.

We have rephrased these sentences in line with the reviewer's comments (line 541-546; 559-560). We have also added that the comparison with existing literature needs to be carefully made since our numbers are actual measured minutes of exercise, while most literature reports inherently unreliable self-reports or crude estimations.

"Adherence to the intervention was higher than reported for previous exercise trials, with 40% of participants being fully adherent over the first 6-months and 30% being fully adherent over the full 2-years compared to pooled estimates of 21% in previous trials(33). This is particularly encouraging as adherence was collected automatically and is therefore a true representation of the actual dosage of balance training people received, while often these figures have lower accuracy due to self-reports or estimations based on number of sessions attended(33)."

"The relatively high adherence and zero serious adverse events show promise for upscaling the intervention to a population level."

35. I am concerned on the lack balance improvements. This makes difficult to explain the main outcome and suggests that expectation bias may have distorted the results. Authors should compare their treatment protocol with already published protocols to highlight differences. They should also provide evidence to readers that this protocol complies with published recommendations.

We did not repeat the lab-based balance and neuropsychological assessments at 18 and 24 months. Moreover, we did not explore all aspects of balance, specifically not balance recovery ability, which could be an important pathway through which exercise training could reduce falls. As such, we do not consider the lack of significant balance improvement as an indication that the effects on falls are chance findings. Moreover, the significant effects of fall rates and

injurious fallers at 24 months, and the similarity of the effect sizes at 12 months, suggest that we merely did not have enough power to show the intervention effect at 12 months at $p<0.05$ (as reviewer 4 also points out). We hope to have resolved this issue by our changes based on Q33.

36. ***Finally, authors stated "This trial might have been underpowered for detecting differences in fall risk factors, as our sample had a lower fall risk than anticipated." It seems that lower fall risk, an overoptimistic expected reduction in rate of falls and a lower than expected dosage may have caused lack of balance improvements. Please, elaborate on this issue in the discussion.***

Our exercise program was designed on the best evidence to prevent falls. We hope that we have addressed this comment by our changes based on Q33.

REVIEWER 4

In this population-based randomized controlled trial, Delbaere and colleagues evaluate the effectiveness of an e-Health balance exercise program delivered via an App compared to health education program only. Participants were 503 individuals age 70 or above who were independent in activities of daily living and without cognitive impairment or any other major disease precluding exercise. The trial was assessor blinded and randomization done in blocks. The primary outcomes were the rate of falls and the number of fallers over 12-months. The authors further tested 18 secondary outcomes. The results indicated no statistically significant differences of the primary outcomes. About the secondary outcomes, the intervention group had a 16% lower rate of falls over 2-years compared to the controls (incidence rate ratio:0.84, 95% confidence interval, 95%CI:0.72-0.98) and a lower proportion of injuries after 2 years.

Trial on the population levels are very difficult to conduct and the authors are to be congratulated to have achieved to keep participants motivated. The trial is overall well conducted and analyzed, including the complier averaged causal effect and bootstrap analyzes. I have a few suggestions that the authors may want to consider.

Thank you.

37. The conclusions in the abstract and the discussion differ and just be similar. The authors should be clearer whether they call their study a null finding study or focus on some statistically significant findings of the secondary outcomes. The authors stated in the protocol that they did not correct for multiple testing as they viewed all their 18 secondary outcomes are plausible and a priori testable outcomes. I challenge this view and rather encourage the authors to look at the consistency of some effect estimates. The incidence rate ratio of the primary outcome (rate of falls after 12 months, 0.82, incident rate ratio of falls after 24 months, 0.84, 95% CI 0.72-0.98). This consistency can also be seen in Figure 2. Thus, one could speak of a consistent effect on falls in the trial. It is apparent that the effect in the trial was not as strong as a priori expected and as a result, the sample size was a little too small. Nevertheless, I think that the consistent effect on falls is more important than focusing on small differences of the P values, one statistically "significant" the other not.

We thank the reviewer for this observation and have revised the abstract and highlights to be consistent with the results and discussion sections.

Abstract (line 106-109; 120-122)*:*

"The rate of falls (0.60 (SD 1.05) falls per year in IG and 0.76 (SD 1.25) in CG; incidence rate ratio, IRR:0.84; 95% confidence interval, 95%CI:0.62-1.13; p=0.071) and proportion of fallers (34.6% in IG and 40.2% in CG; relative risk, RR:0.90, 95%CI: 0.67-1.20, p=0.461) were not statistically different at 12-months."

"StandingTall balance exercise did not significantly affect our primary outcomes. It did significantly reduce the rate of falls and number of injurious fallers over 2-years with similar, albeit not statistically significant at p=0.071, effects at 12-months."

Highlights (line 137-138)*:*

"The StandingTall programme ~~did not significantly affect rate of falls and proportion of fallers at 1-year; however StandingTall did~~ significantly reduce the rate of falls and number of injurious fallers over 2-years with a dose adherence of 30 to 40%."

38. The exercise program included a daily program of up to 20 min per day (provided on a tablet computer). People who did not at least exercised 100 min per week were contacted by phone and all participants in the intervention group received two home visits. I wonder how applicable this program will be outside the trial when people are only using the App also considering that, as expected, not all participants in the trial were adherent. Could the authors discuss this point in the paper?

The phone calls for adherence were only conducted during the first 6 months (line 264) but we agree that these activities could have led to higher adherence than in a completely uncontrolled environment. We are currently trialling implementation without follow up and using telehealth installs and hope to provide some insights within the next year. We acknowledge that adherence may be lower in a community setting and highlight that the comparison of our adherence rates with literature needs to be done carefully as the use of technology allows for more accurate insight into adherence (line 541-554).

"Adherence to the intervention was higher than reported for previous exercise trials, with 40% of participants being fully adherent over the first 6-months and 30% being fully adherent over the full 2-years compared to pooled estimates of 21% in previous trials(33). This is particularly encouraging as adherence was collected automatically and is therefore a true representation of the actual dosage of balance training people received, while often these figures have lower accuracy due to self-reports or estimations based on number of sessions attended(33). Eighty percent of IG participants had a median adherence of 105-minutes over 6-months, and over half sustained a median adherence of 120-minutes over 24-months, despite the low level of contact during the study (two home visits in the first month and incidental follow-up calls during the first 6 months). We do acknowledge that when rolled out to the community with potentially even less follow up, adherence might be lower."

39. I may have missed it, but I am missing a subgroup analysis by sex. I think that one cannot assume similar physiological effects of an exercise program according to sex.

We did not register a subgroup analysis by sex so refrain from including it in the current paper. We thank the reviewer for this suggestion and will take it on board for future analyses/studies.

40. Please report proper effect estimates in the abstract and not just P values. Please report P values to the 2 decimal when >0.10.

We now report effect sizes and p-values in the abstract. Please see our response to Q37 for the changes to the manuscript. We did keep the 3-decimal precision for p-values exceeding 0.10 to keep consistency with reporting of smaller p-values.

41. Please omit the statement that this is the first large study about this topic. While this may be correct, it is not a scientific statement and does not indicate the quality of a study.

We have now omitted this statement from the discussion as suggested.

42. The authors call the program "unsupervised." As there was a follow-up on how often and how long participants did the program plus there were calls and visits, I am not certain that "unsupervised" is the right label.

We agree that our intervention was not fully unsupervised and have now removed this statement as suggested.

43. While not directly related to this study, the authors may want to add a reference to the study by Stensvold https://doi.org/10.1136/bmj.m3485 showing some benefit of a high intensity interval training on all cause mortality (to underscore the importance of exercise programs in elderly people).

We thank the reviewer for the suggestion but, given that mortality is not part of our outcomes (while we agree that its important), have decided to not include this reference. We do now highlight that exercise programs may have benefits beyond fall prevention (line 593-597).

"*In their global action plan on physical activity 2018–2030, the World Health Organisation has advocated exercise as a protective factor in the development of non-communicable diseases such as diabetes, cardiovascular disease, stroke and certain cancers. Recent evidence also shows that exercise can delay the onset of dementia and improve mental health in older people.*"