Statistical Analysis Plan

rSTAND: Remote Digital Health Intervention to Improve Balance and Reduce Fall Risk

Protocol Number: PSC-0820-21

NCT05022589

Document Date: November 22, 2022

(1) ANALYSIS POPULATIONS

There are three a priori defined analysis populations, including a primary analysis population (i), a secondary analysis population designed to compare effect sizes in populations with no missing data (ii), and a population who completed all training visits (iii).

i. Intent to Treat (ITT) population: This is the a priori primary analysis population, defined as including all enrolled participants who completed at least one training session.

ii. Intent to Treat (ITT) Fully-Evaluable (FE) population: This is a secondary analysis population, defined as including all members of the ITT population that complete a post-intervention visit. Note that a participant may complete a specific visit but have missing data for a test in which case the participant is in the overall FE population but does not contribute data to the FE population for that visit, e.g., the number of evaluable cases for a specific test on a specific visit may be smaller than the FE population for that visit because of missing
data.

iii. Intent to Treat (ITT) Completers (C) population: This is a secondary analysis population, defined as including all members of the ITT-FE population who complete all intervention sessions. Note that the C populations are strict subsets of the FE populations; a person who completes the target number of training sessions but does not complete the evaluation visit is not a member of the C population.

(2) STATISTICAL PLAN

Feasibility

The Phase I data analysis plan a priori defines a primary intent-to-treat (ITT) population, a set of secondary evaluation populations, a primary outcome measure, a set of secondary outcome measures, a single primary evaluation time point, a criterion for statistical significance, and a statistical analysis methodology for secondary outcomes.

- The primary ITT population is defined as all participants who complete one training session. Note that this includes all enrolled participants except those who drop/withdraw post-enrollment and pre-training.
- As the main goal of this Phase I trial is to evaluate the feasibility, acceptability and usability of the rSTAND app.

To this purpose, we will conduct an analysis of the following primary outcome measure in participants enrolled in the study (single arm, open label).

1. Primary outcome of this study is the Net Promoter Score, which is calculated based on responses to a single question: How likely is it that you would recommend the Fall Prevention Program to a friend or colleague? The scoring for this answer is based on a 0 to 10 scale. Those who respond with a score of 9 to 10 are called *Promoters*, and are considered likely to exhibit value-creating behaviors, such as making more positive referrals to other potential program

participants. Those who respond with a score of 0 to 6 are labeled *Detractors*, and they are believed to be less likely to exhibit the value-creating behaviors. Responses of 7 and 8 are labeled *Passives*, and their behavior falls between Promoters and Detractors. The Net Promoter Score is calculated by subtracting the percentage of customers who are Detractors from the percentage of customers who are Promoters. For purposes of calculating a Net Promoter Score, Passives count toward the total number of respondents, thus decreasing the percentage of detractors and promoters and pushing the net score toward 0. Net promoter score of >20 would lead us to conclude that the usability/feasibility study had been successful.

The secondary outcome measures were collected by trained personnel at baseline and after the intervention:

1. Program adoption rate.
2. Usability.
    a. Usability ratings were obtained from all study participants post-intervention via a 7-point Likert-scale exit survey. This is a brief and embedded post-study questionnaire on program navigation, perceived benefits, and whether the program helped participants to make a positive change in their activities of daily living. Participants rate each sentence on the following 7-point Likert scale: 1 =Completely Disagree; 2 = Mostly Disagree; 3 = Somewhat Disagree; 4 = Undecided; 5 = Somewhat Agree; 6 = Mostly Agree; 7 = Completely Agree. Clinical trials run by Posit Science software using this survey have shown high internal consistency with a Chronbach's alpha of .917.
3. Engagement with the rSTAND program (number of training levels completed)
4. Activities and Balance Confidence scale (ABC), a 16-item self-report questionnaire to assess *confidence* in performing various activities without losing balance or experiencing a sense of unsteadiness (Total score).
5. Activity Measure for Post-Acute Care (AM-PAC), a validated patient-reported outcome measure designed to assess functional abilities across multiple domains, including basic mobility, daily activities, and applied cognitive functioning (T score).
6. Brief Pain Inventory (BPI), a 11-item self-report questionnaire comprised of self-reported current pain intensity and degree that pain interferes with daily life on a 10-point scale to assess the severity of pain and its impact on functioning (Total score).
7. Falls Calendar, to account for the frequency of falls during the trial period (Total number of falls and loss of balance).
8. Falls Detection, a spontaneous fall detection utility embedded in the Apple watch (Total number of falls).
9. Gait speed via the Gait App.

The criterion for statistical significance is $p < 0.05$. Results with $p < 0.1$ will be described as trends.

The primary analysis time point is the post-intervention assessment.

For feasibility, we relied on descriptive statistics and a single primary outcome. For preliminary efficacy, in smaller sample sizes like ours, there is a greater likelihood that there will be larger

differences at baseline between groups simply due to sampling error. Groups were compared on all baseline measures using Mann-Whitney U tests (for ordinal and continuous variables) or Pearson chi- square/Fisher's exact test (for categorical variables). To test whether groups differed in change over time, data were analyzed via linear mixed-effects models. Mixed-effects models are ideal for the analysis of longitudinal data, due to the fact that it tolerates missing observations–assuming that it is missing at random (MAR). First, simple growth models (change from baseline to post) were conducted to examine change over time (for all participants). Because complete case analyses have been shown to provide biased estimates in the presence of missing data (Enders, 2010), data were then modeled using the intent-to-treat (ITT) framework, in which all randomized subjects are included in the analysis, and missing data were handled via Full Information Maximum Likelihood (FIML). This method for handling missing data is considered a gold standard approach, as it gives relatively unbiased estimates (Enders, 2010), and is the standard for the analysis of randomized trials. For all models, fixed effects of time, group, and group x time interaction were included; a random intercept of subject was used. Kenward-Roger degrees of freedom were used due to the small sample size; statistical significance was evaluated at a two-sided alpha of p = 0.05; however, null hypothesis testing was de-emphasized given the small sample size of the control group (n = 8). We emphasize the effect size (Cohen's d) instead. Although various estimates of effect size exist for LMM, we chose a simplified version, defined as:

$$(\Delta(dCBT{+}WASABI) - \Delta(dCBT)) \,/\, \sigma pre$$

Where Δ refers to slope (i.e., change from baseline to post-intervention) and σpre refers to the pooled baseline standard deviation. Thus, values of d refer to change in standard deviation units from baseline.