Supplement 2

Statistical Analysis Plan for CAP

This document is based on section 15 of the protocol but it provides additional details so that the analysis plan is completely specified. In the process of adding details, the DCC decided to change some of the analysis plans as described in the next section.

Summary of changes from the original plan that was in the protocol.

The table below provides a summary of the differences between this detailed statistical analysis plan and the plan that was in the protocol.

Analysis Question	**Original Plan**	**Additional details and/or changes in the plan**
Primary Outcome (Community Ambulation)		
Primary analysis of Primary outcome: comparison of treatment groups with respect to CAP (Section 1.1.1)	Chi-Square test or Fisher's Exact Test if data are sparse. Cutoffs for significance based on interim monitoring plan. 95% confidence interval for difference between groups in the proportion of community ambulators	No change in statistical tests. Details added regarding construction of confidence intervals for differences in proportions: If the p-value is based on Chi-Square, we will use standard asymptotic intervals. If it is based on Fisher's Exact test, we will use exact intervals.
Sensitivity analysis of conclusions about primary analysis due to nonrandom missingness of the outcome. (Section 1.1.2)	Not in original plan	Published method developed by Magder [1]
Secondary analysis of primary outcome: slightly revised outcome definition, and adjustment for rates of indeterminate outcome data. (Section 1.1.2)	Weighted estimating equations to account for rates of indeterminate outcomes in the two treatment groups	We propose modifying the strategy of choosing weights. Rather than basing the weights on risk of missing outcomes, we will base the weights on imbalances between the groups, whatever the cause. Candidate variables for weights are listed. We will rank the potential weighting variables by a published method[2], and choose a number to include based on a rule of thumb regarding the number of covariates a model can have.

Analysis Question	Original Plan	Additional details and/or changes in the plan
Assessment of Effect Modification of site, period (pre and post decreased visit frequency) and patient characteristics. (Section 1.1.3)	No details provided	Additional Details: Effect modification will be tested on the risk-difference scale based on a binary regression model with an identity link. Also, baseline physical performance as measured by the SPPB will be tested as a potential effect modifier.
Model-based estimates of Delayed and Sustained Effects on Primary Outcome (Section 1.2.1)	Longitudinal logistic regression for two time points (baseline, 16-week, 40-week) model fit with GEE to account for repeated measures from same individual	Changed to a longitudinal binary regression model with identity link and random effect for subject to account for repeated measures. Rationale: Estimates interpretable as rate differences.
Subject-based estimates of delayed and sustained effects on Primary Outcome (Section 1.2.1)	Not in original plan	3x3 table in each treatment group with rows indicating community ambulation at 16 weeks (yes, no, indeterminate) and columns indicating community ambulation at 40 weeks (yes, no, indeterminate).
Secondary and Tertiary Outcomes		
Longitudinal Data Analysis of quantitative secondary and tertiary outcomes. (Section 1.2.2)	Longitudinal Analysis based on three time points (baseline, 16-weeks, and 40-weeks) fit by GEE.	Changed from GEE to maximum likelihood with an unstructured variance. Rationale: A little more flexible than GEE and may adjust for baseline values better.
Sensitivity analysis to account for missing outcome data (Section 1.4)	Weighted Estimating Equations	Adjustment for covariates in the longitudinal regression model. Covariates chosen by a published method
Sensitivity analysis for non-adherence to protocol (removed)	Weighted analysis with weights based on inverse probability of treatment received as a function of covariates and treatment	Not included in new plan due in part to difficulty defining adherence.
Sensitivity analyses to variation between physical therapists or sites (Section 1.5)	Not in original plan	Rerun the analyses including a random effect for physical therapist or site.

Section 1. Detailed Statistical Analysis Plan

Analyses for all aims will be performed according to the intention-to-treat (ITT) paradigm. Prior to confirmatory analysis, exploratory data analyses will be performed. These exploratory analyses will consist of histograms for continuous outcomes to examine whether a transformation is needed to meet modeling assumptions and frequencies for categorical data to assess whether the data are sparse. With the exception of the statistical test of the primary hypothesis, all statistical tests will be two-sided and will not be adjusted for multiple comparisons. As described in greater detail below, the test of the primary hypothesis (comparing the groups with respect to proportion who are community ambulators at 16 weeks) will be based on a one-sided 0.025-level hypothesis test procedure performed at five time points throughout the trial.

1.1 Primary Outcome

1.1.1 Primary Analysis of Primary Outcome.

The primary aim is to determine if a 16-week intervention based on aerobic conditioning, specificity of training, and muscle overload for strengthening (the PUSH intervention) is more successful in producing community ambulation at 16 weeks post-randomization than an intervention of transcutaneous electrical nerve stimulation, flexibility, and AROM exercises (the PULSE intervention). Therefore, the one-sided null hypothesis that the PUSH intervention does not result in a higher proportion of community ambulators 16 weeks post-randomization will be tested. This hypothesis will be tested at five time points based on a Z-statistic (which is equivalent to the square root of the Pearson chi-square statistic). The critical values for each time point were chosen to preserve an overall type-1 error rate of 0.025. Table 1 shows the critical values:

Table 1. Critical Values for Each Planned Analysis

Analysis	Expected % of information available	Critical z-value for upper-bound alpha spending (for efficacy)	Critical z-value for lower bound alpha spending (for inefficacy/harm)
First interim analysis	28.6%	3.09	-1.76
Second interim analysis	40%	3.03	-1.57
Third interim analysis	60%	2.69	-0.97
Fourth interim analysis	80%	2.37	-0.21
Final analysis	100%	2.03	2.03

If the exploratory analyses reveal data sparseness (expected frequency less than 5 for at least one combination of treatment group and community ambulation status), Fisher's exact test will be performed instead of using the chi-square statistic. A 95% confidence interval will be constructed for the probability of community ambulation in each group by inverting two one-sided tests based on the binomial distribution. If the primary analysis p-value is based on the Z-statistic, then a confidence interval for the difference in community ambulation between the two groups will be constructed using standard asymptotic methods. If the primary analysis is based on Fisher's exact test, then an exact 95% confidence interval will be constructed for the difference in probabilities using the method of Chan and Zhang as implemented in SAS. [3]

The binary outcome variable will be ability to walk at least 300 meters in six minutes (yes/no). This variable will take the value of 'yes' if the participant was tested with the SMWT and walked 300 m or more in six minutes. This variable will take the value of 'no' if 1) the participant was

tested with the SMWT and walked less than 300 m in six minutes or 2) the participant was not tested with the SMWT (or the participant was tested but not according to protocol) and adjudication resulted in the participant's being classified as a treatment failure (see 11.5 for description of the adjudication procedure). Participants whose adjudication result is 'indeterminate' will be excluded from the primary analysis.

1.1.2 Secondary Analyses of Primary Outcome

In a secondary analysis, an alternative outcome variable will be created to represent the participant's community ambulation status at 16-week follow-up. This secondary variable will take the value of 'yes' if the participant was tested with the SMWT and walked 300 m or more in six minutes. The secondary variable will take the value of 'no' if 1) the participant was tested with the SMWT and walked less than 300 m in six minutes, 2) the participant was not tested with the SMWT and adjudication resulted in the participant's being classified as a treatment failure because of death, sickness, or gait speed < 0.6 m/s, or 3) the participant was tested with the SMWT but not according to protocol and adjudication resulted in the participant's being classified as a treatment failure, whatever the reason. The secondary variable will take the value of 'missing' if 1) the participant was not tested with the SMWT and adjudication resulted in the participant's being classified as a treatment failure based only on self- or proxy-reported walking limitation or 2) the participant was not tested with the SMWT (or the participant was tested but not according to protocol) and adjudication resulted in the participant's being classified as 'indeterminate'. All participants, including those with a missing value for the alternative outcome variable, will be included in the secondary analysis. Weights will be used to adjust for covariate imbalances between the groups due to chance or differential rates of indeterminate outcomes. For the PUSH group, the weight for each observation will be defined as the inverse probability that an observation would be assigned to the PUSH group given the covariates for that observation. This will effectively up-weight observations from groups with lower probability of being in PUSH. Similarly, for the PULSE group, the weight for each observation will be defined as the inverse probability that an observation would be assigned to the PULSE group given the covariates for that observation. The probability of being assigned to PUSH (or PULSE) will be estimated as a function of covariates based on a logistic regression model. Candidate variables for creating the weights will include age, sex, and type of hip fracture; baseline BMI, MNA®-SF score, CES-D score, 3MS score, distance walked in six minutes on SMWT, SPPB score, mPPT score, NHATS balance score, gait speed on 50-ft fast walk, and gait speed on 4-m usual walk; and the presence of cardiac disease, pulmonary disease, diabetes, and history of stroke or TIA at baseline. Variables will be ranked for inclusion based on the Beach-Meier approach.[2] A common rule of thumb for logistic regression models is that there should be at least 10 events and 10 non-events for every covariate in the model. We will follow that rule of thumb (based on the number of community ambulators observed in our sample) in deciding how many covariates to include in the model used to estimate weights.

Finally, we will perform a sensitivity analysis to assess the degree to which the conclusions of the analysis could be affected by biases due to data missing not at random. To do so we will use the methods described in Magder, 2003. [1]

1.1.3 Assessment of effect modification by site, period, and patient characteristics

Study site will be investigated as a modifier of the effect of the intervention by testing a site-by-intervention interaction term on the risk difference scale based on a binary regression model with an identity link. If there is evidence that study site is an effect modifier (i.e., $p<0.1$ for the interaction term), we will report site-specific treatment effects.

In the fall of 2014, a decision was made to modify the frequency of intervention visits during the first 8 weeks from 3 per week to 2 per week. We will assess the impact of this protocol change on the intervention outcomes in a secondary analysis. To do so, assuming no significant interaction between site and intervention group, we will assess the statistical significance of the interaction between the period during which the participant was randomized (before vs after the protocol change) and treatment group. If there is a significant interaction, then we will make separate treatment group comparisons in the two periods (one comparison for participants randomized during the period when there were three sessions per week and another for participants randomized during the period when there were two sessions per week).

In addition, a series of analyses will be performed to examine the differential impact of the PUSH intervention relative to the PULSE intervention in subgroups defined by participant characteristics. To do this, a variable-by-intervention interaction term will be tested for each of the following variables:

1. Gender
2. Age at baseline (≥85 years versus 60-84 years)
3. Baseline depression (CES-D score ≥ 16 versus CES-D score < 16)
4. Baseline balance confidence (with median Activities-Specific Balance Confidence scale score as the cutpoint to define the subgroups)
5. Baseline nutritional status (MNA®-SF score <8 versus MNA®-SF score ≥8)
6. Baseline cognitive status (3MS score <91 versus 3MS score ≥91)
7. Baseline physical performance (SPPB score <7 versus SPPB score ≥7)

If the interaction term for any of these subgroup variables is significant, results will be presented separately in strata of the subgroup variable.

1.2 Secondary Objectives

1.2.1 Delayed and Sustained Effects

We will use binary regression models with random intercepts fit by maximum likelihood to examine whether the proportion of community ambulators differs between the PUSH and PULSE interventions at 40 weeks post-randomization. Also we will assess whether the difference in proportions at 40 weeks changed from the difference in proportions at 16 weeks. This approach will implicitly take into consideration the 16-week outcome in estimating the 40-week outcome and thereby provides some protection against biases due to missing data. The longitudinal model for this aim is expressed by the equation:

$$p_{ij} = a_i + b1\ X_i + b2\ t40_{ij} + b3\ X_i t40_{ij}. \qquad \text{(Eq. 1)}$$

where p_{ij} is the probability of a community ambulation for participant i at time j, a_i is the random intercept for participant i, X_i is the intervention indicator (1/0) variable; b1 is the treatment effect at 16 weeks; $t40_{ij}$ is the 40-week follow-up time post-randomization indicator (0=16 weeks; 1=40 weeks); and $X_i t_{ij}40$ is the intervention-by-time interaction variable. In fitting this model, 16-week data points with indeterminate outcomes will not be included. Also 40-week data points with indeterminate or missing (by design) outcomes will not be included. The fitted coefficients in Eq.1 provide estimates of the proportion of community ambulators in the PUSH vs PULSE interventions at 16 and 40 weeks post-randomization. To test the null hypothesis of equal proportion of community ambulators in both groups at 40 weeks post-randomization, we will test the null hypothesis H_0: (b1+b3)=0 using a two-sided test with type I error of 0.05. This test will be performed regardless of results from the primary aim. However, the interpretation of results from this test will depend on those from the primary aim. If there is evidence for a difference in

proportion of community ambulators at 16 weeks, then rejecting this hypothesis can be interpreted as evidence for a sustained effect of the PUSH intervention on community ambulation at 40 weeks; if there is no evidence for a difference in proportion of community ambulators at 16 weeks, then rejecting this hypothesis can be interpreted as evidence for a delayed effect of the PUSH intervention on community ambulation at 40 weeks. We are also interested in describing the trends in community ambulation from 16 to 40 weeks post-randomization in both groups. To assess the null hypothesis of no change in the between-group difference of proportion of community ambulators between 16 and 40 weeks post-randomization, we will test the null hypothesis H_0: b3=0 using a Wald chi-square test with type I error of 0.05. All treatment effects will be reported with their respective 95% confidence intervals.

To estimate delayed response and sustainability at an individual level, we will analyze the 40-week community ambulation results in strata defined by the 16-week results. Among those who were not community ambulators at 16 weeks, we will determine the proportion who were community ambulators at 40 weeks. Similarly, of those who were community ambulators at 16 weeks, we will estimate the proportion who were community ambulators at 40 weeks. This information can be summarized in each treatment group using a 3 by 3 table with rows and columns indicating community ambulation at 16 and 40 weeks, respectively (yes, no, indeterminate).

1.2.2 Secondary and Tertiary Outcomes

Secondary outcomes include five variables (endurance, dynamic balance, walking speed, quadriceps strength, and lower extremity function) that are hypothesized to be precursors to community ambulation. In addition, we will examine the difference between the treatments with respect to the following tertiary outcomes: ADLs, balance confidence, quality of life, physical activity, lower extremity physical performance, depressive symptoms, increase of ≥ 50 meters in distance walked in six minutes, cognitive status, and nutritional status.

With the exception of “increase of 50 meters in distance walked in six minutes”, these variables are quantitative. Prior to analyzing the relationship between treatment and the quantitative variables we will use histograms and side-by-side box plots to examine the shape of the distributions and identify potential outliers. If the data strongly depart from normality, we will consider transforming the data or using rank-based methods for the analyses described below.

Mixed effects models fit by restricted maximum likelihood will be used to compare the PUSH and PULSE interventions at 16 and 40 weeks post-randomization with respect to each outcome. Increase of ≥ 50 m in distance walked, a dichotomous outcome, will be analyzed using the same method as the primary outcome. All of the other secondary and tertiary outcomes are continuous; therefore, a normal model will be used to estimate the parameters in the following equation:

$$\mu_{ij} = b0 + b1\ t16_{ij} + b2\ t40_{ij} + b3\ X_i\ t16_{ij} + b4\ X_i\ t40_{ij}, \qquad \text{(Eq. 2)}$$

where μ_{ij} is the mean of the jth outcome from the ith participant , X_i is the intervention indicator (1/0) variable; $t16_{ij}$ and $t40_{ij}$ are the 16- and 40-week follow-up time post-randomization indicators, respectively; and $X_it_{ij}16$ and $X_it_{ij}40$ are the intervention-by-time interaction variables. To account for the correlation between repeated measures from the same person, we will fit an unstructured variance/covariance matrix to the three time points. This model accounts for outcomes at three time points: baseline, 16 weeks, and 40 weeks post-randomization. By treating the baseline value as an outcome, we quantify mean changes in the outcome relative to

baseline levels. Differences between the two groups in changes from baseline to 16 and 40 weeks post-randomization will be compared using Eq. 2 by testing H_0: b3=0 and H0: b4=0 respectively using a Wald chi-square test with type-I error of 0.05. All treatment effects will be reported with their respective 95% confidence intervals.

1.3 Cost-Effectiveness of Interventions

To assess the cost-effectiveness of study interventions, the EEC will conduct analyses of within-trial comparisons for the economic endpoints (resource utilization/costs and SF-6D/QALYs) and will also undertake a model-based analysis that allows the economic value of both study interventions to be assessed relative to usual care. Longitudinal modeling appropriate for repeated measures data will be used to make inferences on the overall differences in cost and QALYs associated with the study interventions. The basic statistical analyses of cost and QALYs will be similar to the approach described for other study endpoints, but will be undertaken in the EEC in close collaboration with the DCC.

Statistical analyses of SF-6D will produce an estimate of the incremental QALYs associated with the PUSH intervention at each time point where SF-36 is measured. The estimated difference in QALYs attributable to the PUSH intervention will be estimated by taking a time-weighted average of the time-specific intervention effects. Statistical analyses of cost data, which will be adjusted to a constant dollar year (e.g., 2012 US dollars), will produce an estimate of the incremental costs associated with the PUSH vs.PULSE intervention.

The incremental cost-effectiveness ratio (ICER), which is defined as the net change in cost divided by the net change in effectiveness (QALYs) when interventions are ranked in order of increasing cost, is the focus of the economic analysis. When estimated as added cost per QALY gained, the ICER allows the value of interventions in hip fracture to be compared with interventions in other diseases. ICERs will be estimated using both the statistical analysis of cost and QALY data (i.e., trial-based ICER) and a model-based analysis that combines trial results with other existing data (i.e., model-based ICER). The trial-based ICER addresses the economic value of the PUSH intervention relative to the PULSE, while the model-based ICER estimates the economic value of the study interventions relative to usual care (as described below).

The second objective of the cost-effectiveness analysis is to develop and implement a decision-analytic modeling framework that will incorporate within-trial findings regarding costs and QALYs for the purpose of evaluating the cost-effectiveness of the PUSH intervention and PULSE relative to usual care. Model-based analyses are commonly employed to extend or augment clinical trials because the cost of trials precludes study of all interventions of interest and because it is often desirable to consider the value of interventions over a longer time horizon than what is observed in the trial. To make inferences about the economic value of the study interventions relative to usual care, a Markov state-transition modeling framework that incorporates trial results will be developed and utilized. Estimates of the cost of the study interventions will be derived from time estimates recorded in the field over the course of the study. Estimates of the QALY impact of usual care will be derived from existing hip fracture cohorts (control arms of other trials). Estimates of changes in SF-6D for similar patient groups are available from the control arms of BHS RCTs. Because these studies have tracked resource utilization with few questions and follow up measures, extensive sensitivity analyses that vary the impact of the interventions on resource utilization will be undertaken to characterize the magnitude of change in costs that would be required to qualitatively affect the conclusions of the economic analysis.

The model-based ICER for the study interventions relative to usual care will be compared qualitatively with the costs per additional QALY estimates of other commonly accepted medical interventions. Uncertainty in the model-based analysis will include estimation of cost-effectiveness acceptability curves, which represent the probability that a particular cost-effectiveness threshold (e.g., $100,000 per QALY gained) is achieved when variability in cost and QALYs is considered in probabilistic sensitivity analyses.

The above cost-effectiveness analyses will not be performed unless at least one of the two following conditions is satisfied:

- There is a statistically significant difference in the primary outcome between groups, OR
- There is a statistically significant and clinically meaningful between group difference in any of the secondary outcomes listed below.

Clinically meaningful differences for the secondary outcomes at 16 and 40-week follow-up:

	Clinically Meaningful Difference
Other Secondary Outcomes: Distance walked in 6 minutes (m)	50 m [1]
Short Physical Performance Battery score	1.0 [1]
Gait speed, 50-ft fast walk (m/s)	0.10 m/s [2]
Gait speed, 4-m usual walk (m/s)	0.10m/s [1]

[1]Perara et al, 2006
[2]Palombaro et al, 2006

1.4 Missing Data

By design, there will be no missing data at baseline because only participants with complete baseline data will be randomized. At follow-up, scores for scales that have published rules for handling missing scale items (e.g., the CES-D and the SF-36) will be calculated using those rules. If necessary, the PAT-D and each of its subscales will be prorated based on nonmissing items. However, the whole scale will be considered missing if more than 30% of its items are missing. Similarly, each of the subscales will be missing if more than 30% of the subscale items are missing. If necessary, the mPPT will be prorated based on nonmissing items. However, the scale will be considered missing if three or more of the tasks were not done because of technical issues (e.g., no stairs available for stair climbing task). All other scales will be considered missing if any part of the scale is missing. As a secondary analysis, to correct for imbalances due to chance or differential missing data, we will adjust for covariates in our models. Candidate variables that will be adjusted for are age, sex, and type of hip fracture; baseline BMI, MNA®-SF score, CES-D score, 3MS score, distance walked in six minutes on SMWT, SPPB score, mPPT score, NHATS balance score, gait speed on 50-ft fast walk, and gait speed on 4-m usual walk; and the presence of cardiac disease, pulmonary disease, diabetes, and history of stroke or TIA at baseline. As described above, these variables will be ranked for inclusion as adjustment variables using the Beach-Meier approach.[2] We will adjust for the same number of covariates in this analysis as in the analysis described in above for covariate adjustment for the analysis of our primary outcome. This approach will result in unbiased estimates if the missingness is at random, conditional on the covariates.

1.5 Accounting for Variability

To address the possibility that variation between physical therapists could affect the findings, we will rerun the analyses including a random effect for physical therapist. Also, to account for variability between clinical sites, we will run analyses including a random effect for site.

References

1. Magder LS. Simple approaches to assess the possible impact of missing outcome information on estimates of risk ratios, odds ratios, and risk differences. *Controlled clinical trials.* 2003;24(4):411-421.
2. Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. *Controlled clinical trials.* 1989;10(4 Suppl):161s-175s.
3. Chan IS, Zhang Z. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics.* 1999;55(4):1202-1209.