STATISTICAL ANALYSIS PLAN

Official Title: Intervention Effectiveness Towards Improving Physical and Mental Health for Post-stroke Patients.

ClinicalTrials.gov Identifier: NCT04941482.

Document Type: Study Protocol.

Version: 2.0 – 25 December 2021.

Principal Investigator: Nguyen Thi Phuong Thao, MD, MSc

Affiliation: Institute for Preventive Medicine and Public Health, Hanoi Medical University, Vietnam.

1. Introduction and Study Overview

This Statistical Analysis Plan (SAP) describes in detail the planned analytical approaches for the above-mentioned randomized controlled trial. The study's objective is to evaluate the efficacy of a 6-month intervention combining Motivational Interviewing (MI) and a home-based rehabilitation program versus standard care in improving post-stroke outcomes. A total of 92 stroke patients were randomized (1:1) to the intervention or control group. Outcomes are measured at baseline, 1 month, 3 months, and 6 months post-intervention. The primary outcomes include depression (PHQ-9 score), fatigue (FSS score), cognitive function (MMSE score), and physical function (Barthel Index). Secondary outcomes include quality of life (Stroke Impact Scale) and exploratory fNIRS-based biomarkers of cortical activity. No interim analyses were conducted; all analyses described here are post-finalization of data collection. This SAP is finalized prior to unlocking the database to avoid data-driven decisions.

2. Analysis Populations

Intent-to-Treat (ITT) Population: The ITT population consists of all participants who were randomized, regardless of whether they completed the intervention or follow-up assessments. Each participant will be analyzed in the group to which they were originally assigned. The ITT population is the primary analysis set for efficacy outcomes, preserving the benefits of randomization.

Per-Protocol (PP) Population: The PP population is a subset of ITT including only those participants who completed the study with no major protocol deviations. For the intervention group, "completed the study" is defined as attending the majority of intervention sessions (at least 6 out of 8 MI sessions and at least 5 out of 6 months of rehab visits) and completing the 6-month follow-up assessment. For the control group, it entails completing the 6-month assessment and not adopting outside interventions similar to the study intervention. Major protocol deviations (if any occur, such as ineligibility discovered after randomization or failure to follow key procedures) will be listed and those cases may be excluded from PP analysis. The PP analysis will serve as a sensitivity analysis to see if results are consistent with the ITT analysis.

Safety Population: As this study involves minimal-risk behavioral interventions, no separate safety analysis population is defined; any adverse events will be descriptively reported for all enrolled participants.

3. Outcomes and Endpoint Definitions

Primary Efficacy Outcomes:

- ***Depression Severity:*** Change in Patient Health Questionnaire-9 (PHQ-9) score from baseline to follow-ups. The PHQ-9 is a 9-item scale (0–27 range); higher scores indicate worse depressive symptoms. The primary comparison will be the difference in mean PHQ-9 change from baseline to 6 months between groups. Additionally, trajectories over all time points will be considered. We will also examine the proportion of participants in each group achieving remission of depression (PHQ-9 < 5, for example) at 6 months as a supportive outcome.

- ***Fatigue Level:*** Change in Fatigue Severity Scale (FSS) score (range 9–63, higher = more fatigue). Same approach as PHQ-9: continuous score change and possibly categorization (e.g., FSS ≥36 indicating significant fatigue) for supplementary analysis.

- ***Cognitive Function*:** Change in Mini-Mental State Examination (MMSE) score (0–30, higher = better cognitive function). We will consider the mean change and group differences; additionally, the proportion of participants with cognitive impairment (MMSE ≤24) at each time point will be noted.

- ***Physical Functional Independence:*** Change in Barthel Index (BI) score (0–100, higher = more independent). Primary focus on mean score change at 6 months. Also, we may categorize outcomes (e.g., BI ≥ 90 as independent vs <90) for descriptive comparison.

Each of these primary outcomes will be analyzed as a continuous variable using appropriate parametric or nonparametric methods as outlined in Section 4. The primary time horizon for efficacy is the 6-month post-intervention assessment, though intermediate time points will be analyzed to observe trends.

Secondary Outcomes:

- ***Quality of Life:*** Stroke Impact Scale (SIS) 3.0 overall score (0–100, higher = better). Change in SIS from baseline to 6 months will be compared between groups. SIS subdomains (physical, emotion, communication, etc.) might also be explored for additional insights, but the overall score is the main focus.

- ***Functional Near-Infrared Spectroscopy (fNIRS) Biomarkers:*** Several exploratory endpoints derive from fNIRS measurements of cortical hemodynamics during a cognitive task:

  - Mean or peak change in oxyhemoglobin (ΔHbO_2, measured in mmol·mm or arbitrary units) in predefined regions of interest (e.g., left orbitofrontal cortex, right orbitofrontal cortex, dorsolateral prefrontal cortex) at baseline, 3 months, and 6 months.
  - The change in ΔHbO_2 from baseline to 6 months for those regions.

- We will specifically evaluate if the intervention group shows a greater increase in ΔHbO_2 in the left orbitofrontal cortex over time compared to controls, as prior evidence suggested a link between orbitofrontal activation and depression improvement.

- ***Derived binary outcomes for ROC analysis:*** e.g., "6-month depression status" (yes/no based on PHQ-9 threshold as noted) and using fNIRS metrics (such as 6-month ΔHbO_2) as predictors.

- ***Other Clinical Measures:*** If collected, the Modified Rankin Scale (mRS) for disability and NIH Stroke Scale (NIHSS) at baseline will be described but not tested as outcomes. The trial registration did not include these as outcomes, but baseline differences will be accounted for if necessary.

Safety Outcomes: The study will note any adverse events (AEs) or serious adverse events (SAEs) reported during the 6-month period, such as falls, medical complications, or psychological distress. Given the nature of the intervention, we do not anticipate intervention-related SAEs. AEs will be coded and summarized by group (e.g., number of falls during follow-up, any hospitalizations, etc.). No formal hypothesis testing on safety outcomes is planned due to expected low incidence and lack of power for such events.

4. Statistical Methods

4.1 General Analysis Principles

All statistical tests will be two-sided with a significance level of $\alpha = 0.05$. Outcomes will be analyzed primarily in the ITT population. We will use **descriptive statistics** to summarize baseline characteristics and outcomes at each time point by group. For continuous variables, we will report means and standard deviations (or medians and interquartile ranges if non-normal). Categorical variables will be summarized as counts and percentages. No formal hypothesis testing will be performed on baseline characteristics (any baseline differences observed will be noted but understood as chance differences due to randomization). Where relevant, 95% confidence intervals (CIs) will be provided for estimated effects (e.g., mean differences, odds ratios). All analyses will be conducted using Stata (v16.0) and R statistical software. fNIRS signal processing will utilize MATLAB with the NIRSIT toolbox, with processed outputs then analyzed in Stata/R.

4.2 Baseline Comparisons and Participant Flow

We will construct a CONSORT flow diagram to document participant flow (screened, randomized, completed, lost to follow-up) by group. Baseline demographic and clinical variables (age, sex, stroke type, time since stroke, baseline PHQ-9, etc.) will be presented by group to check for any notable imbalances. Although randomization should create comparable groups, if any substantial differences are apparent (e.g., a significant age difference or stroke severity difference between groups), these variables may be considered for inclusion as covariates in sensitivity analyses of outcomes. Categorical baseline differences will be assessed with chi-square tests and continuous differences with t-tests or Wilcoxon rank-sum tests (nonparametric) as appropriate. These tests are exploratory and will not affect the primary analysis model unless a clear need for adjustment is identified.

4.3 Primary Outcome Analysis

Longitudinal Analysis (Main Approach):

We will utilize **repeated measures analysis** to assess treatment effects on the continuous primary outcomes (PHQ-9, FSS, MMSE, BI) over time. A linear mixed-effects model (LMM) will be fitted for each outcome with fixed effects for treatment group (intervention vs control), time (categorical: baseline, 1m, 3m, 6m), and the group × time interaction. Each model will include a random intercept for each participant to account for within-subject correlations over time (assuming an unstructured covariance or other appropriate structure based on model fit). The primary parameter of interest is the group×time interaction term at 6 months, which tests whether the change from baseline to 6 months differs between groups. If this interaction is significant ($p<0.05$), it indicates a differential treatment effect over time. We will report the estimated mean change from baseline to 6 months in each group and the difference in change between groups (with 95% CI and p-value). For reference, we will also report group differences at intermediate time points (1 and 3 months) from the same model, though these are secondary.

Assumptions of the LMM (normality of residuals, etc.) will be checked. If necessary (e.g., for skewed distributions like FSS), we may apply transformations (e.g., log transform) or use a nonparametric approach. However, given our sample size, the LMM with robust standard errors should be adequate even if mild deviations exist.

As an alternative verification, we may perform a **GEE analysis** for each outcome. The GEE will treat time as a repeated factor and use an exchangeable working correlation to account for repeated measures. GEE provides population-averaged estimates and is robust to misspecification of correlation. The results from GEE (in terms of intervention effect over time) will be compared to the LMM results. We anticipate they should converge, but any discrepancies will be examined (e.g., if dropout patterns influence results differently).

Endpoint Analysis (Supportive):

We will also analyze the **6-month endpoint directly** using analysis of covariance (ANCOVA) or t-tests. For each outcome, an ANCOVA will be conducted with the 6-month score as the dependent variable, treatment group as factor, and baseline score of that outcome as a covariate. This is a common approach to improve precision. The estimated group effect from ANCOVA (difference in 6-month mean outcomes adjusted for baseline) will be reported with 95% CI. Alternatively, we may compare the change scores (6-month minus baseline) between groups via an independent-samples t-test (equivalent to ANCOVA under equal baseline means). Both approaches yield the treatment effect estimate; ANCOVA is preferred if there were any baseline imbalances in that outcome.

Additionally, within-group changes from baseline will be assessed using paired t-tests for descriptive context (e.g., PHQ-9 reduction in intervention group, p-value; in control group, p-value). These within-group tests will not be used to infer treatment efficacy but to illustrate magnitude of change.

Effect Size:

For each primary outcome at 6 months, we will compute **Cohen's d** effect size for the difference between intervention and control. This can be calculated as the difference in mean change scores (intervention minus control) divided by the pooled standard deviation of change. If using ANCOVA, we can compute an approximate effect size using the baseline-adjusted means. We will interpret effect sizes with standard benchmarks (d≈0.2 small, 0.5 medium, 0.8 large). In addition, effect sizes for within-group change (e.g., baseline vs 6-month in intervention group) may be reported to gauge the magnitude of improvement in absolute terms. To provide confidence intervals for effect sizes, we will use a bootstrap technique: the dataset will be resampled (with replacement) 1000 times, the effect size computed for each resample, and the 2.5th and 97.5th percentiles will form the CI. This bootstrap CI is especially useful if normality assumptions for effect size are uncertain.

Missing Data Handling:

Missing data is expected primarily from dropouts (participants who miss the 6-month assessment, for example). The primary mixed-model analysis (LMM or GEE) can accommodate missing-at-random (MAR) data by using all available observations per participant without imputation. This is our main strategy for ITT analysis – using mixed models inherently includes participants with partial data. However, we will examine the pattern of missingness: if dropout is related to observed characteristics (e.g., those with worse baseline depression are more likely to drop out), the mixed model (assuming MAR conditional on those characteristics) should handle it; if we suspect data are not missing at random, we may perform a sensitivity analysis with imputation.

For sensitivity, we will perform a multiple imputation for missing 6-month outcomes. We will use a multiple imputation by chained equations (MICE) approach including baseline values, group, and available follow-up values to impute missing outcomes. Perhaps 20 imputations will be generated. The primary outcomes will then be analyzed on each imputed dataset and pooled results reported. We may also use a simpler Last Observation Carried Forward (LOCF) as a very conservative scenario (assuming no change after dropout) and re-run primary comparisons to see if conclusions change. The LOCF analysis will not be primary but provides a worst-case bound (especially if the control group has higher dropout, LOCF would bias toward no difference).

Participants who withdraw consent entirely or are lost to follow-up will not contribute post-dropout data; they remain in ITT (with whatever data available up to withdrawal). If a participant dies during the study (which is possible given stroke population), that is a competing risk; we will include them in analysis up until the point of death. If many deaths occur (expected to be low), we could consider a composite outcome or survival analysis for mortality, but given sample size we will likely just report any deaths separately.

4.4 Secondary Outcome Analysis

Quality of Life (SIS): The SIS total score will be analyzed in a similar manner to primary continuous outcomes. A mixed model with group, time, and group×time will test for differences in trajectory. We anticipate improvements in SIS especially in the intervention group if physical and mental health improve. If SIS was only measured at baseline and 6 months (depending on study logistics), then a simple ANCOVA on 6-month SIS (adjusting for baseline SIS) will be done. SIS domain scores can be analyzed descriptively. Because SIS is a secondary endpoint, p-values will be interpreted cautiously.

Categorical Outcomes: If we derive any categorical outcomes (e.g., depression remission yes/no at 6 months, or proportion with any improvement in BI by >10 points), these will be compared using chi-square tests or Fisher's exact test between groups. For example, we might define "clinically significant improvement" in PHQ-9 as a drop of ≥5 points or final PHQ-9 < 5; we would then compare the fraction meeting that in each group (report risk difference or odds ratio with CI). These analyses are exploratory and will be clearly indicated as such.

Subgroup Analyses: We do not have pre-specified subgroup hypotheses, but exploratory subgroup analyses may include looking at whether the treatment effect on depression differs by gender, age group (<65 vs ≥65), or baseline depression severity (PHQ-9 median split). This would involve adding an interaction term between treatment and subgroup in the model. Given limited power, these will be interpreted with caution.

4.5 fNIRS Data Analysis

The fNIRS data will undergo preprocessing (performed in MATLAB) including filtering (to remove noise and drift), motion artifact correction, and segmentation of the task period. For each participant and each session (baseline, 3m, 6m), we will derive summary metrics such as the average oxyhemoglobin concentration change ($\Delta[HbO_2]$) in specific regions of interest during the cognitive task relative to rest baseline. We might average signals over channels covering the left orbitofrontal cortex (OFC), right OFC, ventrolateral PFC, etc., based on the sensor layout. These ROI-level ΔHbO_2 will be the variables used for analysis.

Longitudinal ROI Analysis: Using GEE, we will model ΔHbO_2 as the outcome with group, time (0, 3, 6 months), and group×time as factors, similar to other outcomes. We anticipate, for instance, a significant group×time effect in left OFC HbO_2, indicating the intervention group's HbO_2 increases over time relative to control (as observed in preliminary analyses). If GEE shows significance, we will follow up with post-hoc tests of group differences at 3 and 6 months, and time differences within each group. If data are normally distributed and complete, an LMM could also be used for confirmation. If some fNIRS measurements are missing (due to technical issues or dropouts), GEE can handle missing under MAR assumptions.

Correlation with Clinical Outcomes: We will compute Pearson or Spearman correlations between changes in fNIRS metrics (e.g., ΔHbO_2 from baseline to 6m in a region) and changes in clinical scores (e.g., PHQ-9 drop) across individuals. A positive correlation would indicate that participants with larger brain activation increases tended to have greater depression improvement, supporting fNIRS as a biomarker. A small p-value would suggest significance of this correlation.

ROC Analysis: One key exploratory analysis is to evaluate whether fNIRS measures can classify depression status. We will define a binary variable for depression at 6 months (for example, PHQ-9 ≥10 indicating clinically significant depressive symptoms vs <10 indicating minimal symptoms). We will then take certain fNIRS measures (like the 6-month left OFC HbO_2 level or the change in that level from baseline) and perform ROC curve analysis. The area under the ROC curve (AUC) will be calculated to assess discrimination ability. If AUC is significantly above 0.5, it indicates some predictive value. We will identify the optimal cutoff on the fNIRS measure that maximizes Youden's J (sensitivity + specificity – 1). For that cutoff, we will report sensitivity, specificity, positive predictive value, and negative predictive value. To account for sample variability, we will use a bootstrap approach (e.g., 1000 bootstrap samples of the data) to derive a confidence interval for the AUC and for the sensitivity/specificity at the chosen cutoff. For example, in preliminary findings, a cutoff ΔHbO_2 of approximately –0.47 mmol·mm in left OFC was identified for the experimental group depression prediction; we will verify if similar thresholds emerge and their stability via bootstrapping. This analysis is exploratory and aimed at generating hypotheses for future larger studies on using fNIRS as a screening tool.

4.6 Handling of Missing Data and Outliers

As noted, the mixed models and GEE inherently handle missing longitudinal data under MAR. We will examine the extent of missingness at each follow-up. If >10% of data are missing for key outcomes, we will do multiple imputation as described (Section 4.3). We will also conduct a per-protocol analysis (excluding dropouts) for comparison. If ITT and PP results differ meaningfully, we will attempt to understand why (e.g., maybe dropouts in control differ in outcome).

Outliers: Before final analysis, data will be checked for any outlier values (e.g., a BI score that is impossible, or an extreme fNIRS value beyond physiological range). Any obvious data entry errors will be corrected (by checking source documents). Legitimate outliers will be kept in the analysis, but we may perform robustness checks by running analyses with and without outliers to see if results change. If a particular participant's data unduly influence a result (Cook's distance or residual diagnostics in regression), this will be reported.

For questionnaire scales, if an individual item is missing, we will handle per instrument guidelines (often prorating or mean substitution if ≤1 item missing on PHQ-9 or FSS, etc.), otherwise the score is set to missing.

4.7 Assumption Checks

Normality and Homogeneity: We will use Shapiro-Wilk tests and Q-Q plot inspections for normality of residuals for key models. If PHQ-9 or other scores show significant skew, we may confirm results with nonparametric tests (e.g., Wilcoxon rank-sum for endpoint differences, or a rank-based ANOVA). Similarly, Levene's test will be used to assess equality of variances between groups for t-test/ANOVA assumptions. The repeated measures ANOVA requires sphericity; we will apply Mauchly's test for sphericity. If sphericity is violated, the Greenhouse-Geisser correction will be applied to F-tests (as already anticipated). The linear mixed model approach does not require sphericity, which is one reason we favor it primarily.

Independence: Given randomization, independence of observations across participants is expected. We will ensure no clustering effects (since single center, no cluster-randomization, except repeated measures within subjects which is handled by paired models).

Model Fit: For mixed models, we will check that model residuals vs fitted plots show no major deviations (e.g., heteroscedasticity). If necessary, we might try a transformation of an outcome (e.g., log-transform FSS if very skewed). For binary outcomes (if any logistic regression is used in subgroup or supplementary analyses), we will check for small sample issues (Firth's correction if cell counts are small) and overfitting.

Multiplicity: We acknowledge multiple outcomes are being tested (four primary scales). Our focus is on the collective evidence across these outcomes rather than any single hypothesis. Thus, we are not formally adjusting α for multiple comparisons in the primary domain; however, we will interpret the pattern of results (e.g., if all four move in the favorable direction with several significant, that strengthens conclusions). Secondary outcomes and subgroup analyses will be considered exploratory.

4.8 Software and Code Management

All analyses will be performed using licensed statistical software. The primary tools are: - Stata 16.0: for data cleaning, descriptive analyses, mixed models, t-tests, chi-square tests, etc. - R (RStudio): for additional graphics, possibly GEE modeling (using geepack or similar), and bootstrapping procedures. R may also be used for ROC analysis (using packages like pROC) to easily compute AUC and CIs. - MATLAB (R2023) with NIRSIT Toolbox: for fNIRS signal preprocessing and first-level analysis (filtering, baseline correction, GLM to derive HbO_2 values). The output (HbO_2 values per participant per time) will then be analyzed in Stata/R as described. - All code used for analysis will be saved with version control. Upon database lock, a copy of the raw dataset will be archived. The analysis scripts (do-files for Stata, R scripts, MATLAB scripts) will be annotated and preserved to ensure reproducibility. Any random processes (e.g., bootstrapping) will use set random seeds for reproducibility in reports.

5. Statistical Reporting

The results will be reported in accordance with CONSORT guidelines for RCTs. Continuous outcomes will be presented as mean (SD) at each time by group, and mean differences with 95% CIs. Primary analysis results will be summarized in tables showing group-by-time interaction p-values and effect estimates. Graphical representation: we will include longitudinal plots (e.g., mean PHQ-9 over time by group with error bars) to visualize trends. The ROC analysis results might be depicted with ROC curves for key fNIRS metrics. All figures will have appropriate labels and confidence bands as applicable.

We will produce separate tables for baseline characteristics, for primary outcomes at each time point, for primary analysis results, and for secondary outcomes. AEs will be listed in a table with counts by group.

Any deviations from this SAP (should they occur) will be described in the final study publication. The SAP will be appended to the publication as supplementary material if required by the journal.