# Statistical Analysis Plan

Moderate-Intensity Exercise Versus High-Intensity Interval Training to Recover Walking Post-Stroke: HIT-Stroke Trial 2

ClinicalTrials.gov ID: NCT06268041

Version 1

Finalized 2024-08-11

**Randomization performance**
To assess randomization performance, baseline participant characteristics will be summarized by treatment group. If a baseline prognostic factor is found to meaningfully differ between groups, a sensitivity analysis will assess whether adjusting for that covariate during hypothesis testing would alter the conclusions.

**Treatment fidelity analysis**
Training fidelity data (e.g. intensity, repetition and perceived exertion) will be compared between treatment groups to evaluate treatment fidelity. Each of these dependent variables will be tested in a separate statistical model, which will include fixed effects for treatment group, session number (modeled as a categorical effect), training bout (for variables collected at overground 1, treadmill and overground 2 each session) and all possible interactions. Repeated sessions within the same participant will be modeled with compound symmetry covariance and repeated bouts within the same session will be modeled with unconstrained covariance. Variances will not be constrained to be the same between sessions or between treatment groups. Missing data will be handled with the method of maximum likelihood.

**Adverse event (AE) analysis**
To assess the relative odds of harms, post-randomization AEs will be compared between treatment groups with logistic regression. Separate models will be tested for overall AEs and for each AE categorization, using the number of participants with AE(s) in that category as the dependent variable and fixed effects for [treatment group], [study site] and [baseline walking limitation severity (<0.4, 0.4-1.0 m/s)]. If there is a substantial sample size imbalance between sites, the [study site] term will be dropped from the model, to avoid giving excessive weight to a small number of participants at the lowest enrollment site(s). A 'substantial imbalance' will be considered present if the site with the highest enrollment has >20% more enrolled participants than the site with the smallest enrollment. If only one group has participant(s) with AE(s) in a particular category, continuity correction will add 0.5 participants with and without an AE to each group to permit odds ratio calculation. In this case, it is not possible to adjust for study site or baseline walking limitation severity.

## Hypothesis testing

The primary analyses for both Aims will use *intent-to-treat* methods.

**Aim 1: Determine the optimal locomotor training intensity for safely eliciting meaningful and sustained improvements in walking capacity among chronic stroke survivors.**

*Hypothesis 1a*: Compared with moderate-intensity aerobic training (MAT), high-intensity interval training (HIIT) will elicit meaningfully greater improvement in walking capacity, as measured by a 6-minute walk distance (6MWD) mean change difference significantly greater than +14 meters[1] from baseline (PRE) to post-treatment (12WK).

*Hypothesis 1b*: Differences in 6MWD improvement between the treatment groups will be sustained 3 months after treatment completion (3moPOST). In other words, HIIT will elicit significantly greater 6MWD improvement than MAT from PRE to 3moPOST.

*Hypothesis 1c*: There will be no significant differences in the relative odds of serious adverse events between HIIT and MAT.

The smallest between-group difference of interest for hypothesis 1a (+14 meters) is smaller than clinically meaningful change thresholds that are typically applied to this outcome, which usually range from +20 to +50 meters.[1,2] We opted for a smaller between-group difference of interest for hypothesis 1a because both groups are receiving active treatment, and even relatively small between-group differences have the potential to move a substantial proportion of the distribution of within-group changes across a meaningful change boundary. The selected 14-meter smallest effect of interest is the smallest among recommended 6MWD change thresholds.[1]

**Statistical modeling**
Analysis for hypothesis 1c is described above. Hypothesis testing for hypotheses 1a and 1b will use a general linear model with 6MWD as the dependent variable and fixed effects for [treatment group (HIIT, MAT)], [testing time point (PRE, 4WK, 8WK, 12WK, 3moPOST)], [group x time], [baseline walking limitation severity (<0.4 m/s, ≥0.4 m/s)], [baseline walking limitation severity x time], [study site] and [study site x time], with unconstrained variance & covariance between repeated testing time points within the same participant. Time will be modeled as a categorical factor (i.e. analysis of response profiles), to avoid making assumptions about the time course

of changes. As described above for the AE analysis, the [study site] terms will be dropped from the model if there is a substantial sample size imbalance between the sites.

Hypothesis 1a will be tested by the [group x time] effect from PRE to 12WK, against a null hypothesis of a +14-meter mean change difference for HIIT vs. MAT, using a one-sided minimal effects test.[3] If the null hypothesis is not rejected, a two-sided test against a null hypothesis of zero mean change difference will be performed to assess the possibility of statistically significant but not clinically meaningful between group differences, or significant and meaningful differences favoring MAT. Hypothesis 1b will be tested by the [group x time] effect from PRE to 3moPOST, against a null hypothesis of zero mean change difference, using a two-sided test. Both hypotheses will be tested at the .05 significance level since they are addressing distinct scientific questions. Missing data will be handled with the method of maximum likelihood, which assumes data are missing at random (MAR). Secondary outcomes will also be tested using the same modeling, with false discovery rate (FDR) correction across both time points and all measures, with the Benjamini-Hochberg procedure.[4]

Missing at random (MAR) sensitivity analysis

As done in the first HIT-Stroke Trial,[5] a sensitivity analysis which does not make the MAR assumption will also be performed to assess the extent to which the conclusions may be sensitive to this assumption. For example, it is possible that the MAR assumption could inflate treatment effect estimates from assuming that participants withdrawn due to AEs had similar outcomes to those not withdrawn, which seems implausible in many cases. Thus, this sensitivity analysis will handle missing data in a way that assumes poor outcomes when missingness is related to an AE.[6] For any outcome data point missing due to AE-related participant withdrawal, the true value will be assumed to be distributed around the minimum of the baseline and the last observation for that participant. Otherwise, if the data point missingness is not AE-related, the true will be assumed to be distributed around the last observation for that participant. These distributions for the true values will be assumed to be normal with a standard deviation of 15 m, matching the observed standard deviation of 6MWD changes after 4 weeks of no intervention in a similar population.[7] Using the general multiple imputation framework,[8,9] 50 datasets will be generated by random sampling to impute the missing values, each dataset will be analyzed, and the model estimates will be averaged across datasets.

Consideration of sex and other biological variables

To preliminarily assess for differences in optimal intensity between persons with mild/moderate vs severe baseline walking limitations,[7,10] we will also add a [treatment group x baseline walking limitation severity x time] interaction term to the primary model, acknowledging that this analysis will not be well powered. There is no strong evidence or compelling reason to suspect that the optimal locomotor training intensity depends on other baseline covariates (e.g. sex). However, we will still preliminarily assess this possibility by testing for [treatment group x covariate x time] interactions, using a separate model for each covariate (while also including the lower order terms involving the covariate in the model). If the interaction is not significant, differences in overall training responsiveness will be tested by the [covariate x time] interactions, after dropping the [treatment group x covariate x time] interaction term from the model. The significance level will be FDR corrected across cofactors, except when testing the prespecified stratification factor (baseline walking limitation severity) and the following baseline covariates, which are being prespecified based on prior analysis: Fugl-Meyer lower limb motor function score, Brief Balance Evaluation Systems Test score, Activities-specific Balance Confidence Scale score, age, participant report of pain-limited walking duration, and participant report of prior walking exercise >2 days/week.

**Sample size and power**

This study was initially powered based on a responder analysis for the primary aim. However, given increasing recognition of methodological issues with this statistical approach,[e.g.11] it was decided before the enrollment of the first participant to replace the responder analysis with the minimal effects test described above. Therefore, an updated sample size calculation was conducted based on the minimal effects test. For full transparency, we describe both the original and the revised sample size calculations below.

Original sample size analysis

This study was originally powered to detect a 25% difference in the proportion of meaningful 6MWD responders between HIIT and MAT. Using data from the first HIT-Stroke Trial (N=55),[5] we estimated that 27% of the MAT group will have meaningful improvement, by obtaining predicted/fitted outcomes under the assumption that 1/3 of participants will have severe baseline walking limitations (according to the planned stratified sampling / restricted enrollment). Sample size analysis thus assumed a 52% response proportion in the HIIT group (27% in MAT + 25% difference = 52% in HIIT), which is lower than the empirical value of 53%

from our preliminary trial (i.e. a slightly conservative estimate). Calculations were based on the logistic regression framework, using the closed-form solution of Demidenko[12] implemented in the R package 'WebPower',[13] for a two-sided test with 80% power at the 0.05 significance level, with no interim analysis. This calculation estimated a needed sample size of 124. To conservatively account for up to 20% missing outcome data, the target enrollment was set at 156 (78 per group). This missing data adjustment is conservative because our preliminary multicenter trial only had 10% missing outcomes (7% if not counting participants that had to be withdrawn due to mid-trial COVID-19 shutdown) and because the proposed analysis will impute missing values using reasonable assumptions.

Revised sample size analysis

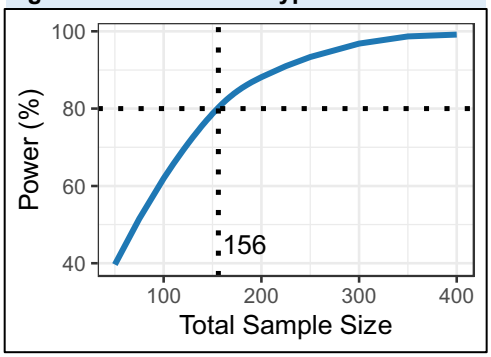
Subsequently, this study was powered to test whether the mean 6MWD change from PRE to 12WK is significantly more than 14 meters[1] greater with HIIT vs. MAT. A simulation was conducted using 6MWD population parameters that were estimated using data from the first HIT-Stroke Trial (N=55).[5] MAT means and residual (co)variances were taken directly from primary analysis results.[5] HIIT means were conservatively set at 80% of the observed mean change differences from MAT (i.e. 80% of +15, +29 and +44 meters at 4WK, 8WK and 12WK, respectively). Percentages of missing data were set at the observed values after excluding missingness due to COVID shutdown and conservatively rounding up to the nearest whole percent (Table 1).

**Table 1. Assumed 6MWD population parameters for sample size analysis.** Values in meters or %.

| Time point | Adjusted means | | Residual (co)variances | | | | Missing data % |
|---|---|---|---|---|---|---|---|
| | MAT | HIIT | PRE | 4WK | 8WK | 12WK | |
| **PRE** | 177 | 177 | 8,421 | 8,479 | 9,283 | 9,372 | 0% |
| **4WK** | 189 | 201 (+12) | - | 11,014 | 9,969 | 10,429 | 2% |
| **8WK** | 206 | 230 (+24) | - | - | 11,935 | 12,218 | 6% |
| **12WK** | 204 | 239 (+35) | - | - | - | 12,905 | 15% |

For each potential sample size, 1,000 random samples were drawn from a multivariate normal distribution with these population parameters (Table 1). The proposed minimal effects test for hypothesis 1a was performed in each sample, and the proportion of samples where the null hypothesis was rejected was taken as the power for each sample size. The power curve (Fig 1) was then smoothed using LOESS and the smallest sample size with ≥80% power was identified. This simulation estimated a total needed enrollment of 156 participants (78 per group; same as the original sample size analysis).

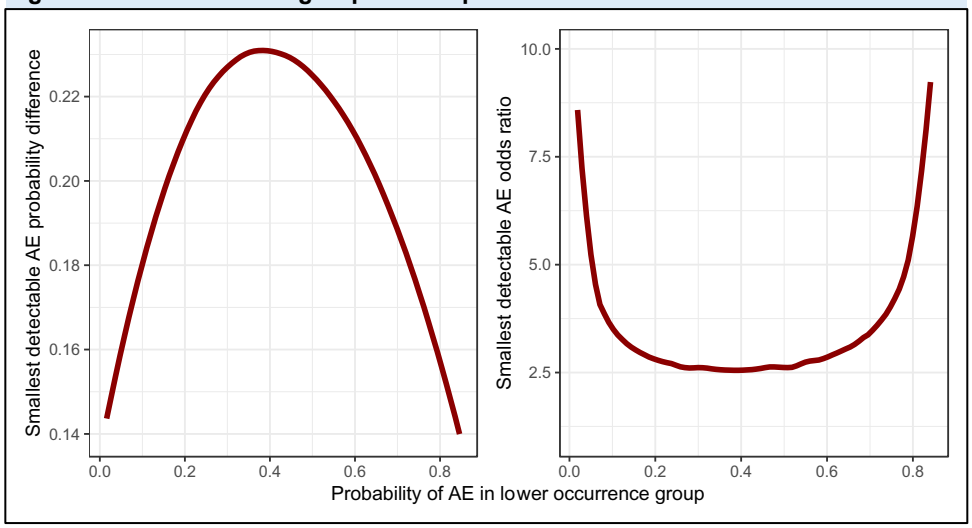
**Fig 1. Power curve for hypothesis 1a**




Power for AE analysis

For hypothesis 1c and other between-group AE comparisons, we also calculated the smallest probability differences and odds ratios (OR) that would be detectable with 80% power using logistic regression. Calculations used the closed-form solution of Demidenko[12] with the variance correction implemented in the R package 'pwrss',[14] for a two-sided test at the .05 significance level. This analysis was performed across the range of possible baseline AE probabilities in the lower occurrence group, and power curves were LOESS smoothed (Fig 2).

**Fig 2. Power for between-group AE comparisons with N=156**




For an AE type with 10% baseline probability, results indicated that we will be powered to detect a probability difference as small as 18%, equating to an OR of 3.50. For an AE type with 50% baseline probability, the smallest detectable probability difference is 22% (OR = 2.57). For an AE type with 75% baseline probability, the smallest detectable probability difference is 17% (OR = 3.83).

**Aim 2: Determine whether vigorous training intensity is particularly critical for patients with severe walking limitations.**

Based on our preliminary data, we do *not* hypothesize that the optimal training intensity differs for patients with severe versus mild/moderate walking limitations, nor that there will be any other significant interaction between baseline walking limitation severity and treatment group when assessing mean walking capacity changes. Instead, our preliminary data suggest that *HIIT will produce better outcomes than MAT regardless of baseline walking limitation severity*, and that *only HIIT will generate clinically meaningful improvement in the severe walking limitation subgroup*.

**Specific Hypotheses**

*Hypothesis 2a:* Compared with MAT, HIIT will elicit significantly greater improvements in walking capacity for persons with severe baseline walking limitations *(speed <0.4 m/s)* **and** for persons with mild/moderate baseline walking limitations (speed ≥0.4 m/s), as measured by mean 6MWD changes from PRE to 12WK.

Among participants with **severe** baseline walking limitations (speed <0.4 m/s) …

*Hypothesis 2b*: **MAT will not** elicit meaningful improvement in walking capacity from PRE to 12WK, as indicated by a mean 6MWD change (and upper 95% confidence limit) <20 meters.[2]

*Hypothesis 2c*: **HIIT will** elicit meaningful improvement in walking capacity from PRE to 12WK, as indicated by a mean 6MWD change (and lower 95% confidence limit) ≥20 meters.[2]

Among participants with **mild/moderate** baseline walking limitations (speed 0.4-1.0 m/s) …

*Hypothesis 2d*: **Both HIIT and MAT will** elicit meaningful improvement in walking capacity from PRE to 12WK, as indicated by mean 6MWD changes (and lower 95% confidence limits) ≥20 meters.[2]

A somewhat larger and more typical meaningful change threshold is being applied for the within-group 6MWD changes in hypotheses 2b-2d (+20 meters) versus the smallest between-group difference of interest for hypothesis 1a (+14 meters). The within-group change threshold of +20 meters is still on the lower end of typically used 6MWD change thresholds, which usually range from +20 to +50 meters.[1,2]

**Statistical modeling**

Hypothesis testing will use the same general linear model as hypotheses 1a-1b above, but with an added [baseline walking limitation severity x treatment group x time] interaction. Variances will also not be constrained to be the same between baseline walking limitation severity subgroups or treatment groups. Hypothesis 2a will be tested by two-sided time contrasts from PRE to 12WK between treatment groups, within each baseline walking limitation severity subgroup, at the .05 significance level. Hypotheses 2b-2d will be tested by one-sided time contrasts from PRE to 12WK against a null hypothesis of +20 meters within the relevant baseline walking limitation severity subgroups and treatment groups. These nested contrasts will also use the .05 significance level but will only be tested if the corresponding between-group difference in mean change from hypothesis 2a is statistically significant (i.e. hierarchical testing). Secondary outcomes will be tested with FDR control across all measures.[4] As in Aim 1, the primary analysis will handle any missing data with the method of maximum likelihood (under the MAR assumption), and the same MAR sensitivity analysis from Aim 1 will be performed. We will also test for baseline covariate effects using the same methods described above for Aim 1.

**Power analysis**

Aim 2 power analysis was based on the expected Aim 1 sample size of 156 and the assumption that 1/3 of participants will have severe baseline walking limitations, according to the planned stratified sampling.

To assess power for hypothesis 2a, we first obtained the standard errors for the mean change estimates within each treatment subgroup using data from the first HIT-Stroke trial[5] with the statistical model proposed above. These standard errors were then used to estimate standard deviations of change by multiplying by the square root of the available sample size in each treatment subgroup. These estimated standard deviations were then used to calculate the smallest between-group differences in change within each subgroup that would be detectable with ≥80% power from a two-sided test at the .05 significance level, using the R package 'pwrss'.[14]

These calculations estimated that we be powered to detect a difference in 12WK 6MWD change between HIIT and MAT as small as 18.9 meters within the severe walking limitation subgroup and as small as 32.2 meters

within the mild/moderate walking limitations subgroup. The estimated mean change differences from the first HIT-Stroke trial[5] using the proposed model were greater than these minimal detectable effects within each subgroup, suggesting we will be sufficiently powered for this hypothesis (Table 2).

**Table 2. Power analysis for 12-week 6MWD changes within treatment subgroups.** Values in meters or N.

| | Subgroup | Target N | Detectable mean change difference with ≥ 80% power | Mean change difference in prior HIT-Stroke Trial using proposed Aim 2 model |
|---|---|---|---|---|
| 2a | Severe limitations: HIIT-MAT | 54 | 18.9 | 39.3 |
| 2a | Mild/mod limitations: HIIT-MAT | 104 | 32.2 | 45.2 |

| H# | Treatment subgroup | Target N | Hypothesized change | Detectable change with ≥ 80% power | Mean ± SD change in prior HIT-Stroke Trial using proposed Aim 2 model |
|---|---|---|---|---|---|
| 2b | Severe limitations: MAT | 26 | < 20 | ≤ 14.3 | 11.8 ± 11.3 |
| 2c | Severe limitations: HIIT | 26 | ≥ 20 | ≥ 35.7 | 51.1 ± 31.2 |
| 2d | Mild/moderate limitations: MAT | 52 | ≥ 20 | ≥ 38.4 | 43.4 ± 52.6 |
| 2d | Mild/moderate limitations: HIIT | 52 | ≥ 20 | ≥ 42.1 | 88.6 ± 63.0 |

Hypotheses 2b-2d will each be tested by within-group change estimates from PRE to 12WK within each treatment subgroup, compared with the minimal clinically important group-change benchmark of 20 meters.[2] For power analysis, we estimated the largest within-group change that would be detectable as significantly less than 20 meters for hypothesis 2b, and the smallest within-group changes that would be detectable as significantly greater than 20 meters for hypotheses 2c-2d. To do this, we used the estimated standard deviations of change within each treatment subgroup (described above) to calculate the largest (hypothesis 2b) or smallest (hypotheses 2c-2d) detectable effect sizes with ≥80% power for a one-sided test at the .05 significance level (Table 2), using the R package 'pwrss'.[14]

For hypotheses 2b-2d, these calculations estimated that we will be powered to detect significant differences from +20 meters for mean changes ≤ 14.3 meters in the severe walking limitations MAT subgroup, ≥ 35.7 meters in the severe walking limitations HIIT subgroup, ≥ 38.4 meters in the mild/moderate walking limitations MAT subgroup and ≥ 42.1 meters in the mild/moderate walking limitations HIIT subgroup. Each of the observed mean change estimates from the first HIT-Stroke trial[5] was within the range of these detectable effects, suggesting we will be sufficiently powered for these hypotheses.

## References


1. Bohannon RW, Crouch R. Minimal clinically important difference for change in 6-minute walk test distance of adults with pathology: A systematic review. *J Eval Clin Pract*. 2017;23(2):377-381.
2. Perera S, Mody SH, Woodman RC, Studenski SA. Meaningful change and responsiveness in common physical performance measures in older adults. *J Am Geriatr Soc*. 2006;54(5):743-749.
3. Daniël Lakens, Scheel AM, Isager PM. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*. 1(2):259-269.
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Statist Soc B*. 1995;57(1):289-300.
5. Boyne P, Billinger SA, Reisman DS, Awosika OO, Buckley S, Burson J, Carl D, DeLange M, Doren S, Earnest M, Gerson M, Henry M, Horning A, Khoury JC, Kissela BM, Laughlin A, McCartney K, McQuaid T, Miller A, Moores A, Palmer JA, Sucharew H, Thompson ED, Wagner E, Ward J, Wasik EP, Whitaker AA, Wright H, Dunning K. Optimal intensity and duration of walking rehabilitation in patients with chronic stroke: A randomized clinical trial. *JAMA Neurol*. 2023;80(4):342-351. PMCID: PMC9951105.
6. Duncan PW, Sullivan KJ, Behrman AL, Azen SP, Wu SS, Nadeau SE, Dobkin BH, Rose DK, Tilson JK, LEAPS Investigative Team. Protocol for the locomotor experience applied post-stroke (LEAPS) trial: A randomized controlled trial. *BMC Neurol*. 2007;7:39. PMCID: PMC2222229.
7. Boyne P, Doren S, Scholl V, Staggs E, Whitesel D, Carl D, Shatz R, Sawyer R, Awosika OO, Reisman DS, Billinger SA, Kissela B, Vannest J, Dunning K. Preliminary outcomes of combined treadmill and overground high-intensity interval training in ambulatory chronic stroke. *Frontiers in Neurology*. 2022;13:812875.
8. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis.* 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2011.
9. Higgins J, Savović J, Page MJ, Elbers RG, Sterne J. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane handbook for systematic reviews of interventions version 6.3.* www.training.cochrane.org/handbook: Cochrane; 2022.
10. Dean CM, Ada L, Lindley RI. Treadmill training provides greater benefit to the subgroup of community-dwelling people after stroke who walk faster than 0.4m/s: A randomised trial. *J Physiother*. 2014;60(2):97-101.
11. Abugov R, Clark J, Higginbotham L, Li F, Nie L, Reasner D, Rothmann M, Yuan X, Sharretts J. Should responder analyses be conducted on continuous outcomes? *Pharm Stat*. 2023;22(2):312-327.
12. Demidenko E. Sample size determination for logistic regression revisited. *Stat Med*. 2007;26(18):3385-3397.
13. Zhang Z, Yuan KH. *Practical statistical power analysis using webpower and R.* Granger, IN: ISDSA Press; 2018.
14. Bulus M. pwrss: Statistical Power and Sample Size Calculation Tools. R package version 0.3.1. 2023.
15. Guo Y, Logan HL, Glueck DH, Muller KE. Selecting a sample size for studies with repeated measures. *BMC Med Res Methodol*. 2013;13:100-100. PMCID: PMC3734029.