APPENDIX

Supplementary Statistical Information

The purpose of the analysis was to evaluate the variability in post-stroke trajectories and identify impairments that contributed to variability in post-stroke recovery rates. Additional interest was in comparing functional status at the final 6-month test session according to baseline stroke severity status.

Given the study's repeated measurements, the linear mixed model approach was naturally suitable for finding significant factors that contributed to variation in patterns of response.[1,2] We estimated the linear trend over time (sessions) for each of the three repeated outcomes [paretic lower extremity loading (PLEL) during sit-to-stand, gait speed (GS), Physical Functioning Index (PFI)] including main effects for session (coded as 1, 2,…, 6 months), gender (1=Male; 0=Female), and three baseline factors centered at their means: age, score on star cancellation, and Fugl-Meyer lower extremity motor scale (FM-leg) score. We excluded race and letter cancellation score because preliminary analyses showed that these factors had little effect on the outcomes in the presence of the other factors in the model. The interactions of session and star cancellation score as well as session and FM-leg score were also evaluated to examine whether baseline neglect and stroke severity affected the rate of recovery. Main effects and interactions were retained in each of the three models regardless of statistical significance ($p < 0.05$). Including interactions in each model allowed for an assessment of functional status at six months without undue constraints from a parallel slopes assumption that would be imposed in the absence of interactions. Due to limited sample size, trends based on session were assumed to be linear. To account for intra-subject correlation as well as to provide subject-specific inferences, we included random intercept and slopes in the model.

Restricted maximum likelihood estimation (REML) via PROC MIXED in SAS 9.2 was used to perform the analysis. Because of the small sample size, the Kenward-Roger degrees-of-freedom correction was used in hypothesis testing.[3] Statistical details of the model are provided below.

The linear mixed model for response at the t-th session for the i-th individual can be written as

$$Y_{it} = \beta_0 + \beta_1 Session_{it} + \beta_2 Gender_i + \beta_3 AgeCen_i + \beta_4 baseFMCen_i + \beta_5 StarCen_i + \beta_{14} Session_{it} * baseFMCen_i + \beta_{15} Session_{it} * StarCen_i + \gamma_{i0} + \gamma_{i1} * Session_{it} + \varepsilon_{it} \qquad i = 1, \dots, 33, \quad t = 1, \dots, 6$$

where,

- $\boldsymbol{Y_{it}}$ – the observed responses PLEL, GS, or PFI
- $\boldsymbol{Session_i}$ – the session number that patients were attending, 1, 2, …, 6
- $\boldsymbol{Gender_i}$ – the participant's gender (1=Male, 0=Female)
- $\boldsymbol{AgeCen_i}$ – the participant's age in years (centered)
- $\boldsymbol{baseFMCen_i}$ – the participant's baseline FM-leg score (centered)
- $\boldsymbol{StarCen_i}$ – the participant's star cancellation score (centered)
- $\boldsymbol{Session_{it} * baseFMCen_i}$ – the interaction term between session and baseline FM-leg
- $\boldsymbol{Session_{it} * StarCen_i}$ – the interaction term between session and star cancellation
- $\boldsymbol{\beta_0}$ – the fixed effects parameter of intercept
- $\boldsymbol{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_{14}, \beta_{15}}$ – the fixed parameters of main effects and interactions
- $\boldsymbol{\gamma} = (\boldsymbol{\gamma_{i0}}, \boldsymbol{\gamma_{i1}})'$ are the random intercept and slope, respectively, for the i-th individual such that $\boldsymbol{\gamma} \sim \boldsymbol{BIVN}(\mathbf{0}, \boldsymbol{G})$ has bivariate normal distribution with 2 *x* 2 covariance matrix $\boldsymbol{G}$
- $\boldsymbol{\epsilon_{it}} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2})$ – the measurement error

For example, the model for PLEL during the sit-to-stand task has the following estimated variance parameters:

$$\boldsymbol{G}_{2\times 2} = \begin{pmatrix} 0.03856 & -0.00314 \\ -0.00314 & 0.00037 \end{pmatrix}$$

and

$$\sigma^2 = 0.00497$$

The negative covariance of random intercept and slope indicates that patients with higher response value of PLEL at Session = 1 tended to have lower recovery rate (smaller slope).

Table 4 in the main paper reports the estimates of fixed effects, noting their levels of significance. For all three models, we confirmed that the main effect of race and the interactions of session with each of age and gender were not statistically significant (not shown). Due to the centering of continuous covariates, the statistically significant regression coefficient for session for each of the three outcomes indicated that subjects with average baseline stroke severity (as indicated by FM-leg and star cancellation scores) had (as expected) statistically significant functional improvement over time, for any age and gender. The fact that none of the two-way interactions was statistically significant for the GS and PFI outcomes indicated that the rate of recovery determined by these outcomes did not differ significantly over time by baseline stroke severity scores or demographic covariates. Also, baseline stroke severity did not significantly impact the PFI score immediately post-stroke (extrapolating to session=0), although there was a consistent trend toward a positive association between baseline FM-leg score (with higher scores indicating less impairment) and PFI.

For PLEL and GS, baseline FM-leg score was statistically significantly positively associated with initial (immediate post-stroke) performance score, whereas there were no significant differences ($p < 0.05$) in the adjusted initial performance scores according to neglect as measured by star cancellation (where lower scores indicate more neglect). As seen by the significant interactions, Table 4 in the main paper shows that both baseline FM-leg score and baseline star cancellation score were significantly associated with the rate of recovery measured by PLEL during the sit-to-stand task. First, holding star cancellation fixed, the mean recovery rate for PLEL score decreased by -0.002 for every unit increase in FM-leg score ($p=0.012$). Participants with worse baseline stroke severity (lower FM-leg score) had significantly higher recovery rate. Second, holding baseline FM-leg score fixed, the mean recovery rate for PLEL increased

by 0.001 for every unit increase in star cancellation score, suggesting that those participants with less baseline neglect (higher star cancellation scores) may have higher recovery rates, on average (p=0.063).

Information related to small sample size. We performed a post-hoc power calculation for the effects of baseline FM-leg on PFI, which was non-significant in the reported analysis results (Table 4). PFI ranges from 0 to 100, with 100 indicating full independence. An estimate for baseline FM-leg of 1.014 as reported in Table 4 is considered very small. Supplementary Table I below reports the power to detect a relationship of baseline FM-leg and PFI immediately post-stroke as provided by our sample size (n=33 subjects), study design (6 follow-up time points) and the linear mixed model for various values of the true regression coefficient. The power was determined with SAS software by simulation using the values for the regression coefficients reported in Table 4 to generate PFI assumed to have a multivariate normal distribution. The same linear mixed model analysis (including the same covariate design X) was used as reported in the main paper.

Supplementary Table I. Power provided by 33 participants with six visits for detecting baseline FM-leg effect in model for PFI using two-sided α=0.05 Wald t-test with Kenward-Roger denominator degrees-of-freedom correction in the final linear mixed model

True coefficient	0 (Type I error)	1.014 (observed)	1.5	2	2.5	3
Power	0.060	0.42	0.57	0.81	0.93	0.99

Supplementary Table II reports power for the interaction of session and baseline FM-leg which was estimated to be 0.053 in Table 4 in the main paper.

Supplementary Table II. Power provided by 33 participants with six visits for detecting interaction effect of session and baseline FM-leg in the model for PFI using two-sided α=0.05 Wald t-test with Kenward-Roger denominator degrees-of-freedom correction in the final linear mixed model

True coefficient	0 (Type I error)	0.053 (observed)	.20	.30	.40	.50
Power	0.054	0.08	0.42	0.75	0.94	0.99

These results show that our sample size had approximately 80% power to detect a clinically meaningful difference in baseline FM-leg score of 2.0, and an increase of 0.31 in slope for time trend (session) for every one unit increase in baseline FM-leg score. The two tables also show that the simulated Type I error for both effects (i.e., setting the coefficient to zero in the data generation model) is near the nominal 0.05 level, therefore the linear mixed model approach with Kenward-Roger degrees-of-freedom correction is valid for this data. Without the Kenward-Roger correction, the simulated Type I errors for testing the effects of baseline FM-leg and session-by-FM-leg interaction are 0.062 and 0.058, respectively. Therefore, the Kenward-Roger correction improves upon the validity of our approach, which would have very good performance in our setting even without the correction.

We conducted a small simulation study to verify the validity of the approach for the available data. The simulations showed that the model was valid with bias of the estimators being very small and coverage probabilities for confidence intervals near the nominal 95% confidence level. Supplementary Table III shows simulation results for PFI based on 500 simulations generating data from the final linear mixed model with values for true regression parameters set to their estimated values from Table 4 of the paper:

Supplementary Table III. Bias and coverage of 95% confidence intervals for baseline FM-leg and its interaction with session in the final linear mixed model for PFI based on 500 simulations

Effect	*True beta*	*Simulated average beta estimate*	*Nominal coverage probability*	*Simulated coverage probability*
Baseline FM-leg	*1.014*	*1.045*	*0.950*	*0.938*
*Session*Baseline FM-leg*	*0.053*	*0.055*	*0.950*	*0.942*

Information related to amount of therapy. The study recorded, at each monthly test session, the number of hours per week of physical therapy (PT) that each participant had been receiving during the month (by self-report). In a sensitivity analysis, we added this variable for amount of PT (and its square) to the model (results shown in Supplementary Tables IV, V, and VI). Although the amount of PT was inversely correlated with performance measures, the results for PLEL and GS from Table 4 were fairly robust to the influence of amount of PT. While results for physical function score were imprecise in the sensitivity analysis (with large standard errors for many factors), results for factors of interest from Table 4 remained non-significant.

Supplementary Table IV. Selected fixed effects estimates (SE) for PLEL under three different models

Effect	**Model A (Table 4)**	**Model B**	**Model C**
Session	0.0214 (0.0047)***	0.0189 (0.0053)***	0.0164 (0.0063)**
Session*Baseline FM-leg	-0.0021 (0.0008)**	-0.0019 (0.0008)**	-0.0018 (0.0008)**
Session*Star Cancellation	0.0010 (0.0005)*	0.0011 (0.0005)**	0.0011 (0.0005)**
Baseline FM-leg	0.0240 (0.0064)***	0.0230 (0.0063)***	0.0226 (0.0063)***
Star Cancellation	0.0006 (0.0040)	0.0002 (0.0039)	0.0001 (0.0039)
Weekly number PT sessions		-0.0025 (0.0084)	-0.0091 (0.0130)
Weekly number PT sessions squared		-0.0001 (0.0008)	0.0003 (0.0010)
Session* Weekly number PT sessions			0.0019 (0.0028)

SE = standard error. PLEL = paretic lower extremity loading. *p<0.10; **p<0.05; ***p< 0.01. All models adjusted for age and gender.

Supplementary Table V. Selected fixed effects estimates (SE) for GS under three different models

Effect	Model A(Table 4)	Model B	Model C
Session	0.0610 (0.0077)***	0.0541 (0.0086)***	0.0469 (0.0100)***
Session*Baseline FM-leg	0.0002 (0.0013)	0.0006 (0.0013)	0.0011 (0.0014)
Session*Star Cancellation	-0.0005 (0.0008)	-0.0002 (0.0009)	-0.0001 (0.0009)
Baseline FM-leg	0.0345 (0.0091)***	0.0320 (0.0086)***	0.0311 (0.0086)***
Star Cancellation	0.0093 (0.0053)*	0.0079 (0.0050)	0.0075 (0.0050)
Weekly number PT sessions		0.0010 (0.0112)	-0.0163 (0.0163)
Weekly number PT sessions squared		-0.0010 (0.0011)	-0.0002 (0.0012)
Session* Weekly number PT sessions			0.0058 (0.0038)

SE = standard error. GS = gait speed. *p<0.10; **p<0.05; ***p< 0.01. All models adjusted for age and gender.

Supplementary Table VI. Selected fixed effects estimates (SE) for PFI under three different models

Effect	Model A (Table 4)	Model B	Model C
Session	2.979 (0.687)***	2.457 (0.796)***	1.693 (1.019)
Session*Baseline FM-leg	0.053 (0.115)	0.135 (0.117)	0.176 (0.124)
Session*Star Cancellation	0.006 (0.073)	0.061 (0.075)	0.069 (0.077)
Baseline FM-leg	1.014 (0.716)	0.685 (0.668)	0.606 (0.672)
Star Cancellation	-0.038 (0.425)	-0.302 (0.396)	-0.328 (0.397)
Weekly number PT sessions		2.472 (1.437)*	0.505 (2.154)
Weekly number PT sessions squared		-0.355 (0.138)**	-0.249 (0.164)
Session* Weekly number PT sessions			0.629 (0.487)

SE = standard error. PFI = Physical Functioning Index. *p<0.10; **p<0.05; ***p< 0.01. All models adjusted for age and gender.

Information about comparisons at the 6-month testing session. In this section of the Appendix, we describe how we compared mean functional status at session 6 for two different groups. First, if star_cen is fixed at its mean, the difference in mean task at session 6 corresponding to a

change in baseline FM-leg score equal to the difference in the first (R_1) and third (R_3) quartiles of FM-leg (see results in Table 5 of the main paper) is:

$$\begin{aligned} E(Y|X = x, session = 6, FM_{cen} = R_3, star_{cen} = mean(star_{cen})) \\ - E(Y|X = x, session = 6, FM_{cen} = R_1, star_{cen} = mean(star_{cen})) \\ = (\beta_4 + 6\beta_{14})(R_3 - R_1) \end{aligned}$$

Next, if baseFM_cen is fixed at its mean, the difference in mean task performance at session 6 for a change in baseline star cancellation score equal to the difference in the first (Q_1) and third (Q_3) quartiles of star cancelation (see Table 5 in the main paper) is:

$$\begin{aligned} E(Y|X = x, session = 6, baseFM_{cen} = mean(baseFM_{cen}), star_{cen} = Q3) \\ - E(Y|X = x, session = 6, baseFM_{cen} = mean(baseFM_{cen}), star_{cen} = Q1) \\ = (\beta_5 + 6\beta_{15})(Q_3 - Q_1) \end{aligned}$$

A test of any difference at session 6 is given by $H_0: \beta_4 + 6\beta_{14} = 0$ and $H_0: \beta_5 + 6\beta_{15} = 0$ for individuals with any difference in FM-leg score and star cancellation score, respectively.

Based on the fitted model results, mean differences in task scores at the 6-month session were evaluated according to changes from the first to third quartiles in FM-leg and star cancellation scores, respectively. The results of hypothesis tests are shown in Supplementary Table VII, whereas Table 5 of the main paper shows 95% confidence intervals. For example, an increase in baseline FM-leg score of 10 (the difference in first and third quartiles) corresponded to an increase in PLEL during sit-to-stand at six months of 0.0115 (std. err.= 0.0044; p-value =0.014). Similarly, holding all other factors fixed, an increase in baseline star cancellation score of 9 (the difference in first and third quartiles) corresponded to an increase in PLEL during sit-to-stand at

six months of 0.0067 (std. err.= 0.0025; p-value = 0.013). For GS (during 10-m walk), an increase in baseline FM-leg score of 10 (the difference in first and third quartiles) corresponded to an increase in 10-m walk performance at six months of 0.0355 (std. err.= 0.0098; p-value =0.001).

Supplementary Table VII. Contrast testing mean differences in task performance at 6-month session

PLEL (during Sit-to-Stand)	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Test: baseFM + 6* Session * baseFM = 0	1	28.1	6.87	0.0140
Test: star + 6*Session * star = 0	1	28.4	7.03	0.0129
GS (during 10-m walk)				
Test: baseFM + 6* Session * baseFM = 0	1	29.0	13.24	0.0011
Test: star + 6*Session * star = 0	1	26.3	1.17	0.2900
Physical Function Index				
Test: baseFM + 6* Session * baseFM = 0	1	31.1	3.35	0.0767
Test: star + 6*Session * star = 0	1	28.9	0.00	0.9948

In combination with the estimated mean differences in Table 5 from the main paper, we see that for PLEL during the sit-to-stand task, a higher baseline FM-leg score (higher baseline star cancellation score) was significantly positively associated with higher PLEL, so that those with worse baseline FM-leg scores did not catch up by session 6. A similar conclusion is reached for GS during the 10m-walk with respect to baseline FM-leg score, but not in regards to star cancellation, for which estimates were in the same direction but non-significant. Finally, mean differences in PFI at session 6 for individuals with different FM-leg or star cancellation scores were not statistically significant; in other words, there is insufficient evidence from the data to conclude that individuals with worse baseline FM-leg or star cancellation scores were still lagging behind others on the PFI at session 6.

References

1. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics.* 1982;38(4):963-974.

2. Holditch-Davis D, Edwards LJ. Modeling development of sleep-wake behaviors. II. results of two cohorts of preterms. *Physiol Behav*. 1998;63(3):319-328.

3. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 1997;53(3):983-997.