PTJ: Physical Therapy & Rehabilitation Journal | *Physical Therapy*, 2024;104:1–11
https://doi.org/10.1093/ptj/pzad183
Advance access publication date January 3, 2024
Original Research

Early Prediction of Poststroke Rehabilitation Outcomes Using Wearable Sensors

Megan K. O'Brien, PhD[1,2,†], Francesco Lanotte, PhD[1,2,†], Rushmin Khazanchi, BA[3,†], Sung Yul Shin, PhD[1,2], Richard L. Lieber, PhD[2,4,5], Roozbeh Ghaffari, PhD[4,6], John A. Rogers, PhD[4,6,7,8], Arun Jayaraman, PT, PhD[1,2,*]

[1]Max Nader Lab for Rehabilitation Technologies and Outcomes Research, Shirley Ryan AbilityLab, Chicago, Illinois, USA
[2]Department of Physical Medicine and Rehabilitation, Northwestern University, Chicago, Illinois, USA
[3]Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA
[4]Department of Biomedical Engineering, Northwestern University, Evanston, Illinois, USA
[5]Shirley Ryan AbilityLab, Chicago, Illinois, USA
[6]Querrey Simpson Institute for Bioelectronics, Northwestern University, Evanston, Illinois, USA
[7]Departments of Materials Science and Engineering, Chemistry, Mechanical Engineering, Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois, USA
[8]Department of Neurological Surgery, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

*Address all correspondence to Dr Jayaraman at: ajayaraman@sralab.org

[†]Megan K. O'Brien, Francesco Lanotte, and Rushmin Khazanchi are cofirst authors.

Abstract

Objective. Inpatient rehabilitation represents a critical setting for stroke treatment, providing intensive, targeted therapy and task-specific practice to minimize a patient's functional deficits and facilitate their reintegration into the community. However, impairment and recovery vary greatly after stroke, making it difficult to predict a patient's future outcomes or response to treatment. In this study, the authors examined the value of early-stage wearable sensor data to predict 3 functional outcomes (ambulation, independence, and risk of falling) at rehabilitation discharge.

Methods. Fifty-five individuals undergoing inpatient stroke rehabilitation participated in this study. Supervised machine learning classifiers were retrospectively trained to predict discharge outcomes using data collected at hospital admission, including patient information, functional assessment scores, and inertial sensor data from the lower limbs during gait and/or balance tasks. Model performance was compared across different data combinations and was benchmarked against a traditional model trained without sensor data.

Results. For patients who were ambulatory at admission, sensor data improved the predictions of ambulation and risk of falling (with weighted F1 scores increasing by 19.6% and 23.4%, respectively) and maintained similar performance for predictions of independence, compared to a benchmark model without sensor data. The best-performing sensor-based models predicted discharge ambulation (community vs household), independence (high vs low), and risk of falling (normal vs high) with accuracies of 84.4%, 68.8%, and 65.9%, respectively. Most misclassifications occurred with admission or discharge scores near the classification boundary. For patients who were nonambulatory at admission, sensor data recorded during simple balance tasks did not offer predictive value over the benchmark models.

Conclusion. These findings support the continued investigation of wearable sensors as an accessible, easy-to-use tool to predict the functional recovery after stroke.

Impact. Accurate, early prediction of poststroke rehabilitation outcomes from wearable sensors would improve our ability to deliver personalized, effective care and discharge planning in the inpatient setting and beyond.

Keywords: Balance, Biomedical Engineering, Decision Making: Computer-Assisted, Gait, Inpatients, Outcome Assessment (Health Care), Patient Care Planning, Prognosis, Rehabilitation, Technology Assessment: Biomedical

Received: January 15, 2023. **Revised:** November 13, 2023. **Accepted:** December 3, 2023

Introduction

Stroke is a leading cause of disability worldwide.[1] Following initial treatment, many stroke survivors are admitted to an inpatient rehabilitation facility (IRF) for ongoing medical care and targeted, intensive, multidisciplinary therapy in the early stages after stroke. A primary goal of IRF rehabilitation is to maximize the neural and functional recovery to help patients reintegrate into the community upon discharge.[2] However, not all individuals have the same potential for recovery. Patients achieve widely varying levels of function after initial treatment, with some returning to premorbid function and others retaining severe deficits that require additional short- or long-term care.[3]

Starting at IRF admission, clinicians must plan *when* the patient will be discharged from the hospital, *where* they can be safely discharged (ie, to their home with or without caregiver assistance, or to a skilled nursing facility for ongoing rehabilitative care), and how to structure therapy to optimize a patient's overall discharge disposition. In the USA, the average IRF length of stay has decreased to 12.9 days for patients with Medicare,[4] giving clinicians, patients, and families only a brief window to design short-term care strategies and postdischarge plans suited to the patient's needs (eg, seeking and training caregivers, making home modifications or alternative living arrangements, and ordering assistive devices). Early, objective, and accurate predictions of a patient's functional recovery would help clinicians, patients, and families plan appropriate treatment and reintegration strategies based on the expected discharge disposition.

Numerous research models have been proposed to predict the stroke recovery.[5,6] Many of these models use exclusively information available from the electronic medical records (EMRs), including patient demographics and clinical information.[7–9] While such models lend themselves to simple and relatively undemanding clinical implementation, their resolution may not detect subtle differences between patients, leading more often to rules of thumb about recovery rather than predicting specific patient outcomes. Conversely, high-resolution metrics, such as from transcranial magnetic stimulation or brain imaging, could improve prediction resolution and accuracy,[10–13] but these measures are costly and are not often available in rehabilitation settings, posing barriers to clinical uptake.

Noninvasive wearable sensors show promise for capturing biomarkers of disease and recovery by mining patterns from continuous, high-resolution physiological or behavioral data.[14,15] We previously demonstrated that data from inertial measurement units (IMUs), recorded during a brief walking bout within a week of IRF admission, improved the prediction of ambulation ability at discharge compared to traditional functional assessments (FAs) and other patient descriptors.[16] However, a patient's discharge disposition depends on different abilities, such as navigating their home environment and performing activities of daily living safely and independently. Therefore, we propose 3 functional outcomes for prediction models which may be considered broadly representative of these attributes: the 10-Meter Walk Test (10MWT; ambulation), Functional Independence Measure score (FIM; independence, specifically related to motor tasks), and the Berg Balance Scale (BBS; risk of falling). To enhance the clinical value of model predictions, we used clinically significant cut-off scores to classify outcomes as signifying none-to-mild and moderate-to-severe impairment. Finally, while in our previous work we used sensor data solely from walking tasks, here, the recorded activities also encompassed simple balance tasks. Consequently, incorporating a nonambulatory population into our approach expands our insights into the potential of sensor-based prediction models for a broader range of patients and IMU data.

The objectives of the present study were to expand our early-stage prognostic models to predict 3 poststroke functional outcomes (ambulation, independence, and risk of falling) at IRF discharge for both patients who are ambulatory and patients who are nonambulatory using data recorded at admission and to evaluate the ability of IMU data to predict each of these 3 outcomes. We hypothesized that incorporating lower-limb IMU data would improve the prediction of discharge outcomes relative to models trained on clinician-scored FAs and demographic and clinical patient information (PI) alone.

Methods

Participants

Fifty-five patients were recruited from the inpatient rehabilitation unit of the Shirley Ryan AbilityLab (Chicago, IL, USA). Inclusion criteria were: having a primary diagnosis of stroke, being aged at least 18 years, and able and willing to give consent and follow study directions. Exclusion criteria were: having a known neurodegenerative pathology; pregnant or nursing; or utilizing a powered, implanted cardiac device for monitoring or supporting heart function. Medical clearance was obtained from the primary physician prior to participation. All individuals (or a proxy) provided written informed consent, and the study was approved by the Institutional Review Board of Northwestern University (Chicago, IL; STU00205532).

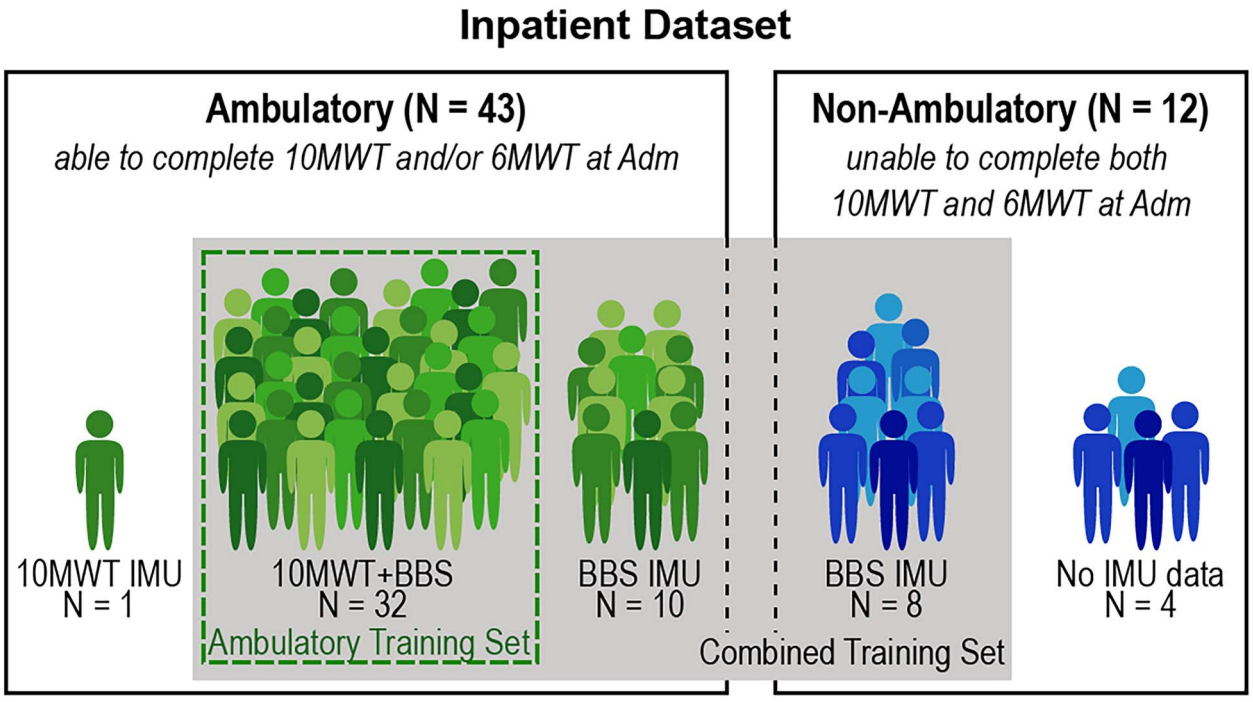
Experimental Protocol

Data were collected from patients at 2 timepoints: within 1 week of IRF admission and within 1 week prior to discharge. At each timepoint, participants completed a series of standardized FAs, including the 10MWT, BBS, 6-Minute Walk Test (6MWT), and Timed "Up & Go" test (TUG). FIM scores were extracted from the patient's EMR at each timepoint. All assessments were administered and scored by a licensed physical therapist. Assessments that could not be completed were scored as 0. PI—including demographics, premorbid activity level, and stroke characteristics—were obtained from the EMR and a study intake form.

Sensor data were collected from 3 flexible, wireless IMUs (BioStampRC; MC10 Inc, Cambridge, MA) during the FAs. These devices were attached to the lumbar region (L4–L5 level) and each ankle (proximal to the lateral malleolus, along the mid-sagittal line) by using an adhesive film (Tegaderm; 3M, St. Paul, MN, USA). They recorded triaxial signals from an accelerometer (sensitivity ±4 g) and a gyroscope (sensitivity ±2000 deg/s) sampled at 31.25 Hz.

Selection of Sensor Data for Model Training

We divided participants into 2 groups based on their walking status during IRF admission. Patients who were ambulatory ($N = 43$) were individuals who could complete at least 1



		Ambulatory Training Set	Combined Training Set (Amb,Non-Amb)
Ambulation (10MWT)	Community: > 0.4 m/s	26	33 (31,2)
	Household: ≤ 0.4 m/s	6	17 (11,6)
Independence (FIM Motor)	High Ind.: > 61	17	21 (21,0)
	Low Ind.: ≤ 61	15	29 (21,8)
Risk of falling (BBS)	Normal Risk: > 45	13	14 (14,0)
	High Risk: ≤ 45	19	36 (28,8)

Figure 1. Inpatient dataset available for model training and testing. Data were collected from 55 individuals undergoing poststroke inpatient rehabilitation at admission and discharge. Training sets for prediction models were determined based on ambulatory status during admission and the availability of IMU data from gait and balance tasks. For patients who were ambulatory at admission, we utilized their IMU data recorded during the 10MWT and BBS ($N = 32$). For patients who were nonambulatory at admission, we combined IMU BBS data for both patients who were ambulatory and nonambulatory ($N = 50$) and tested only on those who were nonambulatory ($N = 8$). All models were tested using a leave-1-subject-out approach. 6MWT = 6-Minute Walk Test; 10MWT = 10-Meter Walk Test; Adm = admission; Amb = ambulatory; BBS = Berg Balance Scale; FIM = Functional Independence Measure; IMU = inertial measurement unit; Ind = independence; Non-Amb = nonambulatory.

walking assessment at admission (10MWT, 6MWT, or TUG) with no more than moderate assistance from a physical therapist. Patients who were unable to complete all the walking assessments at admission were considered nonambulatory ($N = 12$).

To establish a simple yet inclusive set of physical activities to capture potential biomarkers of recovery across these 2 groups, we narrowed the sensor analysis to a single walking task that could be completed by most participants who are ambulatory and a series of nonambulatory tasks that could be completed by most participants regardless of ambulatory status.

For the walking task, we selected a single trial of the 10MWT at self-selected velocity, which we previously found to be predictive of ambulation discharge outcomes among individuals who are ambulatory.[16] In our present dataset, 33 patients who were ambulatory had IMU data during the 10MWT (Fig. 1).

For the nonambulatory tasks, we selected the first 4 items of the BBS (standing unsupported for up to 2 minutes, sitting unsupported for up to 2 minutes, stand-to-sit transition, and sit-to-stand transition), which are among the least demanding and had a high completion rate among all patients (Suppl. Fig. 1). In our dataset, 8 patients who were nonambulatory and 42 patients who were ambulatory had IMU data during these 4 tasks (Fig. 1).

Annotated sensor data for each task were cleaned by removing duplicate timestamps and resampling to the expected sampling frequency (31.25 Hz) using spline interpolation. Data processing, filtering, and subsequent feature extraction were completed in MATLAB R2017b (Mathworks Inc, Natick, MA, USA).

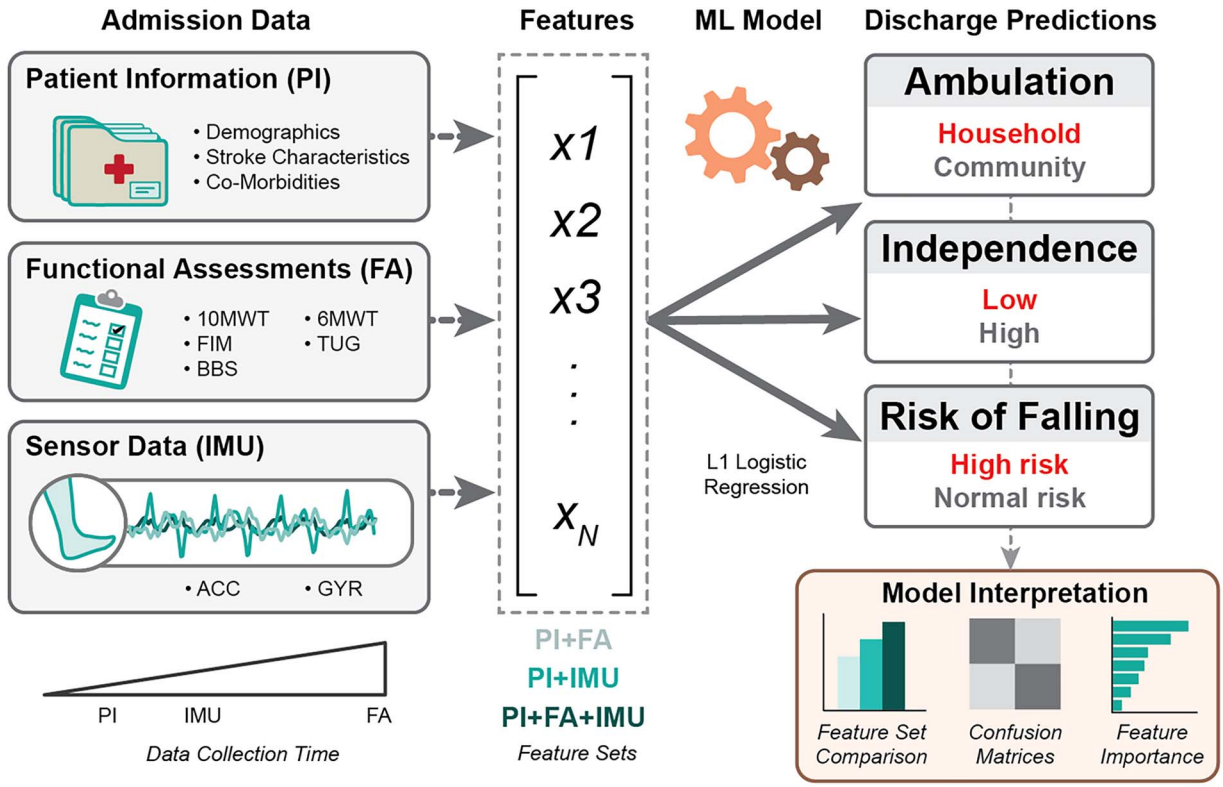


Figure 2. Data pipeline for prediction models. Data collected at inpatient rehabilitation facility (IRF) admission (PI, FA, and IMU signals) were combined in different feature sets and input into an L1-penalized logistic regression model. The model was trained to predict functional outcomes at IRF discharge, related to the classification of ambulation, independence, and risk of falling. 6MWT = 6-Minute Walk Test; 10MWT = 10-Meter Walk Test; Acc = accelerometer; BBS = Berg Balance Scale; FA = functional assessments; FIM = Functional Independence Measure; Gyr = gyroscope; IMU = inertial measurement unit; ML = machine learning; PI = patient information; TUG = Timed "Up & Go" test; *X*1, *X*2, *X*3, and *XN* = example features extracted from admission data.

Feature Extraction

Features are measurable, independent variables used as an input to a machine learning algorithm to make predictions. Three feature categories were defined in this study: PI, FA scores, and wearable sensor (IMU) data. To reduce the dimensionality of the feature space and increase robustness to the sensor placement, IMU features were computed from the Euclidean norm of the triaxial accelerometer and gyroscope signals. IMU features for the BBS were supplemented with measures of postural sway, which were computed from the mediolateral and anteroposterior axes of the lumbar sensor (Suppl. Tab. 1). We applied 1-hot encoding to categorical variables to prevent ordinality issues. Supplementary Table 2 summarizes characteristics of the PI and FA features for our different training and testing datasets.

Combinations of these feature categories were used to train prediction models, creating 3 different types of models for comparison: a benchmark model (PI + FA, no sensor data), including both PI and FAs, a streamlined sensor model (PI + IMU), including easily obtained PI, and a comprehensive model (PI + FA + IMU), including all feature types. The PI + FA benchmark served as a comparative point of reference to determine the impact of sensor data on predicting each discharge outcome.

Model Architecture and Training

We trained separate supervised learning classifiers to predict 3 different discharge outcomes: ambulation, independence, and risk of falling (Fig. 2). For each outcome, we defined 2 classes of patient function at discharge; namely, household versus community ambulators (based on 10MWT score[17,18]), low versus high independence (based on FIM motor subscore[19,20]), and high versus normal risk of falling (based on BBS score[21]).

Classifiers were developed using the Scikit-Learn (0.23.2) library in Python 3.8.8. We selected L1-penalized logistic regression, given its ability to handle the high dimensionality, relatively small sample size, and the varying degrees of class imbalance. L1-penalized logistic regression also requires few hyperparameters and calculates feature importance scores, simplifying the training and interpretation processes for more direct comparison between the models. Models were trained and tested to predict the 3 discharge outcomes for the ambulatory and nonambulatory populations by using nested leave-1-subject-out crossvalidation (Suppl. Fig. 2).

Models predicting ambulatory outcomes at discharge were exclusively trained and tested using the 32 patients who were ambulatory and had IMU data available for both the 10WMT and BBS (Fig. 1). To determine the most predictive sensor tasks for patients who were ambulatory, we compared model performance when training with IMU features from BBS only (IMU_{BBS}), 10MWT only (IMU_{10MWT}), and BBS and 10MWT combined ($IMU_{10MWT+BBS}$).

Models predicting nonambulatory outcomes at discharge were trained using data from the combined ambulatory and nonambulatory populations to maximize the availability of the BBS IMU data. We refer to these models as nonambulatory

models because they were tested and intended exclusively for the 8 patients who were nonambulatory. This combined training was adopted to increase the sample size and heterogeneity of discharge outcomes for model learning compared to the nonambulatory cohort alone.

Model Interpretation

The primary performance metric was the weighted F1 score (WF1), defined as the harmonic mean of the precision and recall, computed separately for each class j, and weighted by the number of samples n_j within each class, with the highest possible value of 1.0 indicating perfect precision and recall.[22]

$$\mathrm{WF1} = \frac{\sum_{j=1}^{L} 2 \cdot \frac{\mathrm{precision}_j \cdot \mathrm{recall}_j}{\mathrm{precision}_j + \mathrm{recall}_j} \cdot n_j}{\sum_{j=1}^{L} n_j}.$$

Secondary performance metrics were accuracy and log-loss scores. Accuracy is the ratio of correct predictions to the total number of samples, with the highest value of 1.0:

$$\mathrm{Accuracy} = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn}},$$

where tp, tn, fp, and fn are the numbers of true positives, true negatives, false positives, and false negatives, respectively. Positive classes were household ambulation ability, low independence, and high risk of falling.

Log-loss measures the variation between prediction probabilities and true classes, wherein lower values indicate greater certainty about the predictions.[23] Given a true label y_i and the prediction probability $p_i = \Pr(y_i = 1)$, log-loss is computed as:

$$\text{Log-loss} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \cdot \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i) \right).$$

Confusion matrices were generated for the best-performing models to examine misclassifications for each outcome and patient group. These were also compared to the benchmark PI + FA model. Parameter importance was determined from the coefficients fit in the best model, taking the median coefficient value and the 25th and 75th percentiles across all participants. Parameters with median, 25th, and 75th percentile values equal to 0 were discarded.

Role of the Funding Source

The funders played no role in the design, conduct, or reporting of this study.

Results

Classification performance for each model and feature set is presented in the Table and is summarized below.

Ambulation

For patients who were ambulatory, the benchmark PI + FA ambulation model had a WF1 of 0.709. Gait-based IMU features, either alone or combined with balance features, improved performance in both the streamlined and comprehensive sensor model configurations by 19.6%. Balance-based IMU features alone did not affect the ambulation predictions.

Table. Performance Metrics for Ambulatory and Nonambulatory Patients for Each Prediction Model and Feature Set[a]

Patient Group	Prediction Model	Benchmark			Sensor Models							
		PI + FA			IMU Task	Streamlined – PI + IMU			Comprehensive – PI + FA + IMU			
		WF1	Accuracy	Log-Loss		WF1	Accuracy	Log-Loss	WF1	Accuracy	Log-Loss
Ambulatory ($N = 32$)	Ambulation	0.709	0.719	0.904	BBS	0.688	0.688	0.594	0.688	0.688	0.484
					10MWT[b]	**0.848**[b]	0.844	0.546	**0.848**[b]	0.844	0.545
					BBS + 10MWT	0.838	0.844	0.292	0.838	0.844	0.294
	Independence	0.685	0.688	1.236	BBS	0.588	0.594	1.016	0.622	0.625	1.044
					10MWT[b]	0.657	0.656	1.211	**0.688**[b]	0.688	1.346
					BBS + 10MWT	0.563	0.563	1.422	0.622	0.625	1.081
	Risk of falling	0.534	0.531	1.380	BBS	0.347	0.344	1.116	0.566	0.563	0.727
					10MWT[b]	**0.659**[b]	0.656	0.974	0.628	0.625	0.877
					BBS + 10MWT	0.597	0.594	1.788	0.658	0.656	0.785
Nonambulatory ($N = 8$)	Ambulation	0.643	0.750	0.787	**BBS**[b]	0.300	0.250	0.916	**0.859**[b]	0.875	0.392
	Independence	**0.933**[b]	0.875	0.782	**BBS**[b]	0.933	0.875	0.284	**0.933**[b]	0.875	0.407
	Risk of falling	**1.000**[b]	1.000	0.246	**BBS**[b]	0.933	0.875	0.294	**1.000**[b]	1.000	0.184

[a]10MWT = 10-Meter Walk Test; BBS = Berg Balance Scale; FA = functional assessments; IMU = inertial measurement unit; PI = patient information; WF1 = weighted F1 score. [b]Highest WF1 for each model and patient group shown in **bold**, indicating the best-performing parameter sets and IMU tasks to predict discharge outcomes.

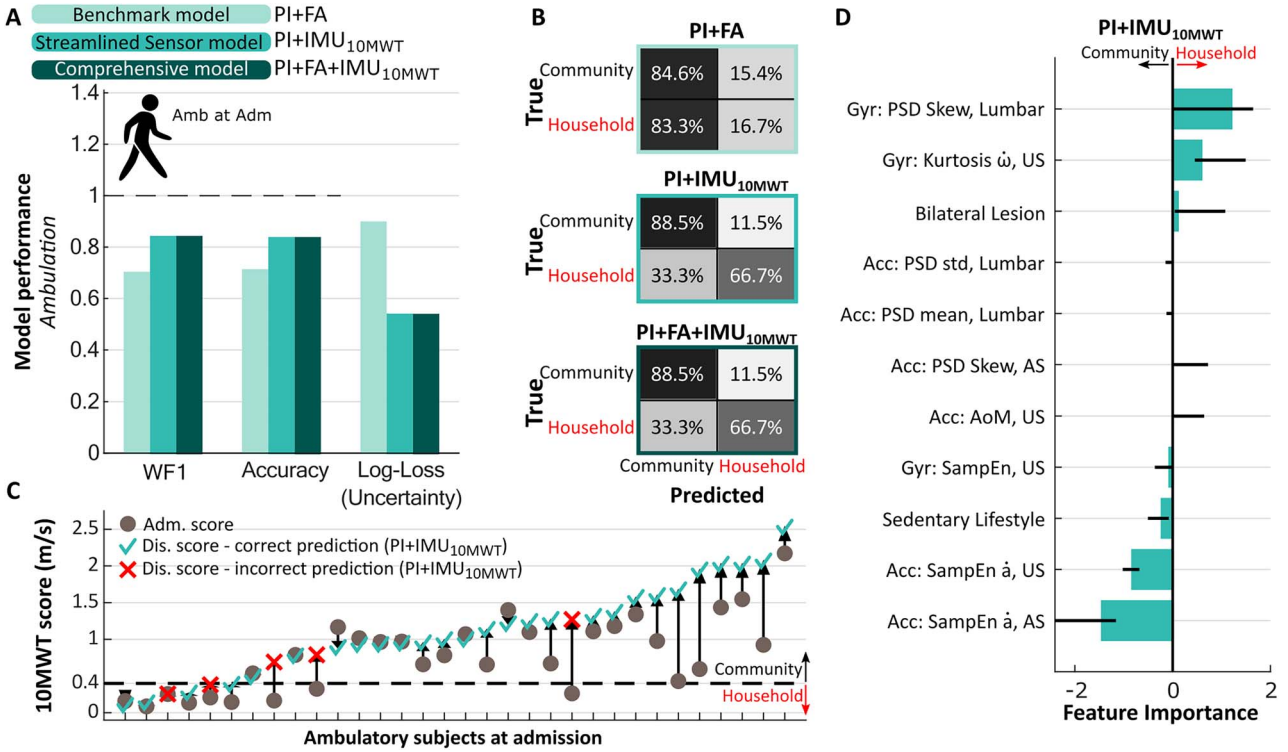


Figure 3. Prediction models for ambulation at discharge (ambulatory at admission). (A) WF1, accuracy, and log-loss for the benchmark model (PI + FA), streamlined sensor model (PI + IMU_{10MWT}), and comprehensive model (PI + FA + IMU_{10MWT}). (B) Confusion matrices. (C) 10MWT score at admission (circles) and discharge timepoints. Values at discharge are marked with ticks if correctly predicted by the best-performing model (simplest model with the highest WF1), or with crosses if incorrectly predicted. (D) Median and interquartile ranges of the coefficients fit to the most important features for the best-performing model. 10MWT = 10-Meter Walk Test; ȧ = derivative of acceleration; $\dot{\omega}$ = derivative of gyroscope; Acc = accelerometer; Adm = admission; Amb = ambulatory; AoM = amount of motion; AS = affected side; Dis = discharge; FA = functional assessments; Gyr = gyroscope; IMU = inertial measurement unit; PI = patient information; PSD = power spectral density; SampEn = sample entropy; US = unaffected side; WF1 = weighted F1 score.

The gait-based streamlined sensor model, PI + IMU_{10MWT}, was selected as the best model for patients who were ambulatory, given its simple configuration and highest WF1 (Fig. 3A). The streamlined sensor model outperformed the benchmark, correctly identifying more patients who were household (4 vs 1 patient[s]) and community (23 vs 21 patients) ambulators at discharge (Fig. 3B). The PI + IMU_{10MWT} model also correctly identified 27 of 29 patients who did not change the ambulation category from IRF admission to discharge, though it misclassified 3 patients who improved from household to community ambulators (Fig. 3C). Eleven features were selected for the PI + IMU_{10MWT} model, including lesion location, activity lifestyle, and IMU features from all sensor locations (Fig. 3D).

For patients who were nonambulatory, the comprehensive model trained on the combined dataset was the best ambulation model, achieving a WF1 of 0.859 (Suppl. Fig. 3A). The PI + FA + IMU_{BBS} model correctly classified 1 of 2 individuals who were nonambulatory and progressed to community ambulators as well as all 6 individuals who were nonambulatory and were discharged as household ambulators (Suppl. Fig. 3B and C). Notably, 2 individuals remained nonambulatory at discharge, with 1 completing the 6MWT but was unable to complete the 10MWT. Among the 28 features selected for the comprehensive model, the admission 10MWT score and IMU balance features were the most important predictors of community and household ambulation, respectively (Suppl. Fig. 3D).

Independence

For patients who were ambulatory, the benchmark PI + FA independence model had a WF1 of 0.685. Gait-based IMU features yielded a similar WF1, while balance features performed slightly worse in both the streamlined (−14.2%) and comprehensive (−9.2%) sensor models. Combining gait and balance IMU features further decreased WF1 (up to −17.8%) for both sensor models.

The gait-based comprehensive model, PI + FA + IMU_{10MWT}, was the best-performing model according to WF1 (Fig. 4A). Compared to benchmark, the comprehensive model correctly classified more individuals who were ambulatory and were discharged with low independence (11 vs 9 patients), though with fewer correct predictions for individuals with high independence (11 vs 13 patients) (Fig. 4B). Misclassifications were higher among participants with discharge FIM motor scores close to the class threshold. The PI + FA + IMU_{10MWT} model correctly identified 10 out of the 16 patients who transitioned from low independence to high independence (Fig. 4C). Fourteen features were selected for this model, including gyroscope features from the lumbar and unaffected-side ankle. Participant age was the most discriminative feature for low independence at discharge, while the 10MWT and BBS admission scores indicated high independence (Fig. 4D).

For patients who were nonambulatory, independence predictions achieved the same WF1 of 0.933 across models, with the least uncertainty in the comprehensive model (Suppl. Fig. 4A–C). We selected the benchmark as the best model, which

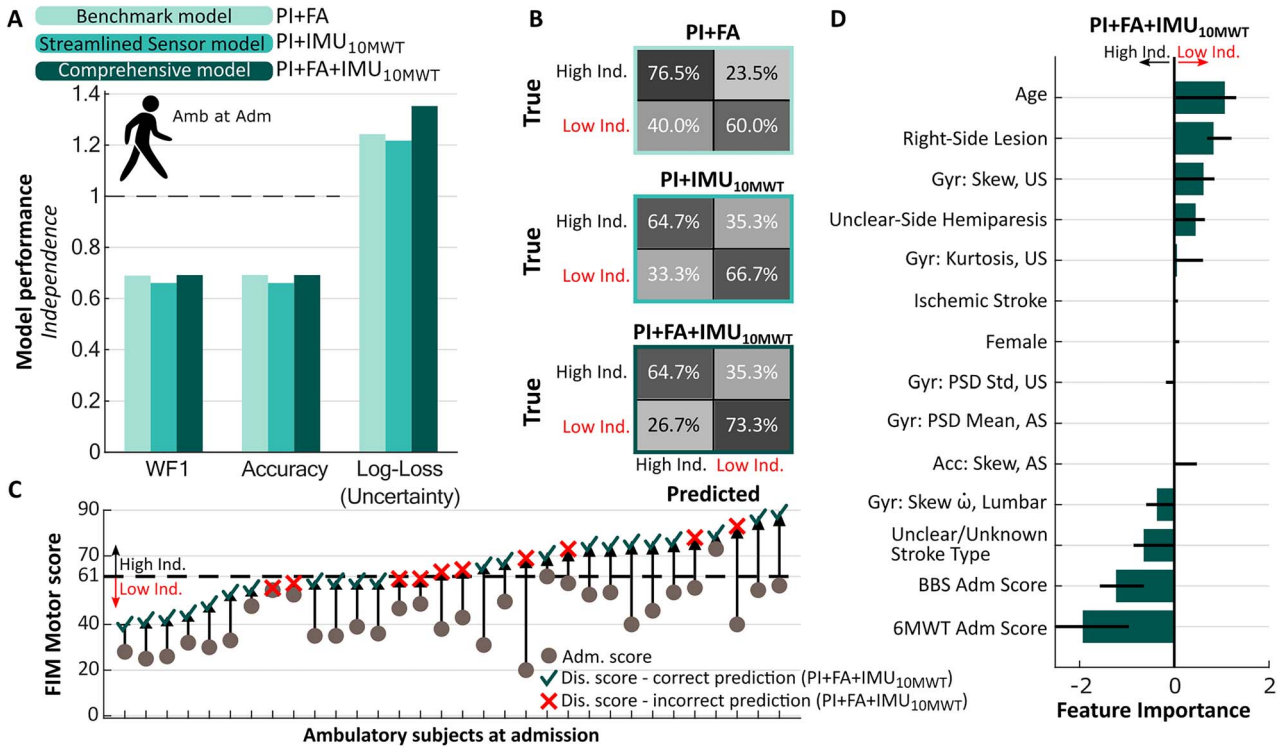


Figure 4. Prediction models for independence at discharge (ambulatory at admission). (A) WF1, accuracy, and log-loss for the benchmark model (PI + FA), streamlined sensor model (PI + IMU_{10MWT}), and comprehensive model (PI + FA + IMU_{10MWT}). (B) Confusion matrices. (C) 10MWT score at admission (circles) and discharge timepoints. Values at discharge are marked with ticks if correctly predicted by the best-performing model (simplest model with the highest WF1), or with crosses if incorrectly predicted. (D) Median and interquartile ranges of the coefficients fit to the most important features for the best-performing model. 6MWT = 6-Minute Walk Test; 10MWT = 10-Meter Walk Test; $\dot{\omega}$ = derivative of rotational velocity (from gyroscope); Acc = accelerometer; Adm = admission; Amb = ambulatory; AS = affected side; BBS = Berg Balance Scale; Dis = discharge; FA = functional assessments; FIM = Functional Independence Measure; Gyr = gyroscope; IMU = inertial measurement unit; Ind = independence; PI = patient information; PSD = power spectral density; US = unaffected side; WF1 = weighted F1 score.

used simple features, such as age, admission 6MWT, and admission BBS scores, to differentiate between the 2 levels of discharge independence (Suppl. Fig. 4D).

Risk of Falling

For patients who were ambulatory, the benchmark risk-of-falling model had a WF1 of 0.534 (Fig. 5A). Balance-based IMU features decreased performance in the streamlined sensor model to 0.347, but they slightly increased performance in the comprehensive model to 0.566. Gait-based IMU features improved performance relative to the benchmark model in both the streamlined (23.4%) and comprehensive (17.6%) sensor models. Combined gait and balance IMU features did not increase performance further in either model configuration.

The gait-based streamlined sensor model, PI + IMU_{10MWT}, was selected as the best risk-of-falling model. Compared to the benchmark, the streamlined sensor model correctly classified more individuals who were ambulatory and were discharged with both high risk (12 vs 9 patients) and normal risk (9 vs 8 patients) (Fig. 5B). Incorrect predictions were more likely when the BBS discharge score was near the cut-off value. The PI + IMU_{10MWT} model correctly predicted 5 patients who transitioned from high to normal risk (out of 8 total) (Fig. 5C). Of the 23 features selected for this model, various IMU and demographic features had similar average importance to distinguish individuals with high and normal risk of falling (Fig. 5D).

For patients who were nonambulatory, risks of falling predictions were perfectly accurate for the benchmark and comprehensive models, whereas the streamlined sensor model exhibited marginally lower performance (Suppl. Fig. 5A). Both the benchmark and the comprehensive models identified all individuals who were nonambulatory with high risk of falling (Suppl. Fig. 5B and C). The benchmark was selected as the best model by utilizing the simplest set of 4 features with relatively low uncertainty. Lifestyle and left-side hemiparesis were markers for high fall risk, whereas the BBS admission score had the highest importance to identify individuals with normal risk (Suppl. Fig. 5D).

Discussion

For patients who were ambulatory at admission, we found that the IMU sensor data recorded from the lumbar and ankles during walking tasks improved early predictions of poststroke inpatient rehabilitation outcomes (ambulation, independence, and risk of falling) compared to benchmark predictions derived from EMR-based PI and standardized FA scores. For ambulation and risk of falling, IMU features extracted during a 10-m walking bout increased the WF1 in a streamlined sensor model (PI+ IMU_{10MWT}), while FA

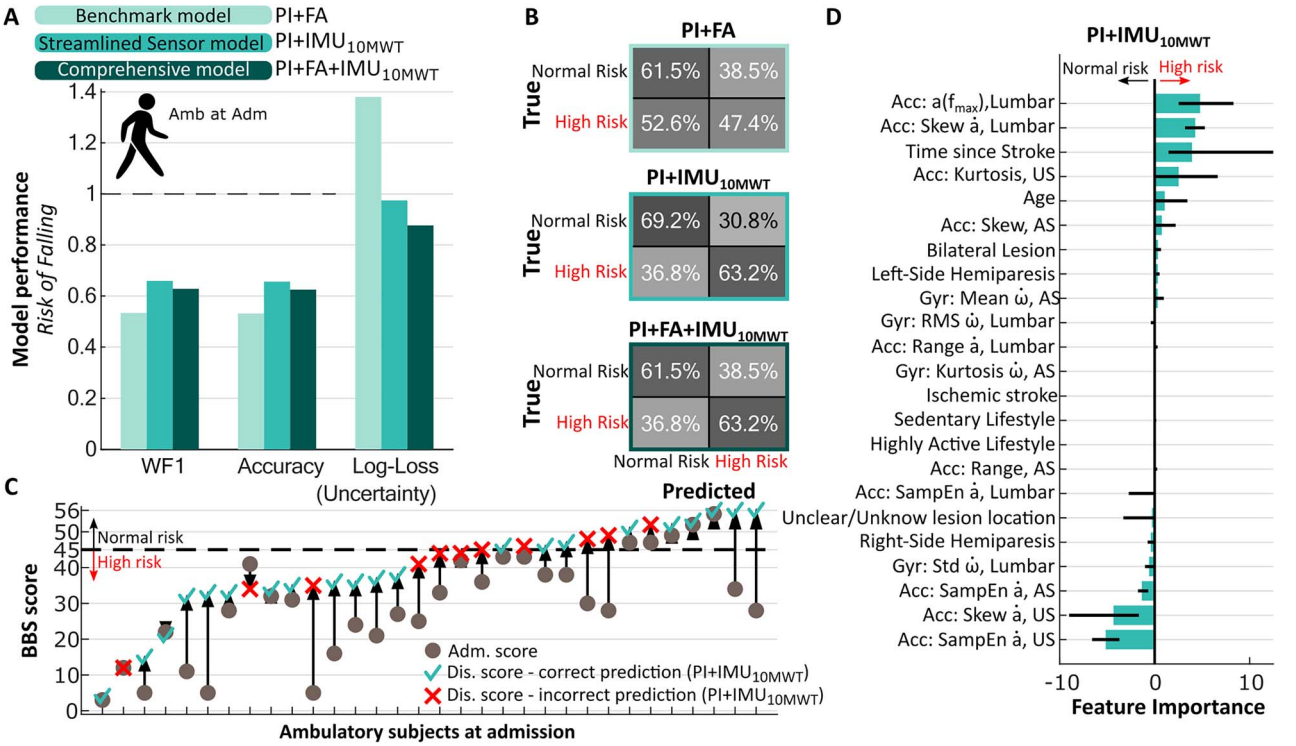


Figure 5. Prediction models for risk of falling at discharge (ambulatory at admission). (A) WF1, accuracy, and log-loss for the benchmark model (PI + FA), streamlined sensor model (PI + IMU_{10MWT}), and comprehensive model (PI + FA + IMU_{10MWT}). (B) Confusion matrices. (C) 10MWT score at admission (circles) and discharge timepoints. Values at discharge are marked with ticks if correctly predicted by the best-performing model (simplest model with the highest WF1), or with crosses if incorrectly predicted. (D) Median and interquartile ranges of the coefficient fit to the most important features for the best-performing model. 10MWT = 10-Meter Walk Test; ȧ = derivative of acceleration; $\dot{\omega}$ = derivative of rotational velocity (from gyroscope); $a(f_{max})$ = amplitude at maximum frequency; Acc = accelerometer; Adm = admission; Amb = ambulatory; AS = affected side; BBS = Berg Balance Scale; Dis = discharge; FA = functional assessments; Gyr = gyroscope; IMU = inertial measurement unit; PI = patient information; RMS = root mean square; SampEn = sample entropy; US = unaffected side; WF1 = weighted F1 score.

features from admission further improved the predictions for independence. Similar to our previous work, including sensor data improved the predictions of discharge ambulation compared to a benchmark model.[16] This finding was repeatable across the 2 studies despite using different modeling approaches and algorithms (random forest vs L1-penalized logistic regression). A streamlined PI + IMU_{10MWT} model improved the ambulation predictions by 19.6% over the benchmark performance, achieving an 84.4% accuracy to predict the community/household ambulators based on the 10MWT score. A comprehensive PI + FA + IMU_{10MWT} model performed similarly to the benchmark independence performance, scoring an accuracy of 68.8% to classify high/low independence based on the FIM motor subscore. A streamlined PI + IMU_{10MWT} model improved the benchmark risk-of-falling performance by 23.4%, achieving a 65.9% accuracy to classify the normal/high risk based on the BBS score. Most misclassifications occurred when patients had admission or discharge scores near the class boundary (Figs. 3 and 5C).

For patients who were nonambulatory at admission, incorporating the IMU data from simple balance tasks added less value to predicting discharge ambulation function. The comprehensive models were as accurate as the benchmark models for independence and risk of falling outcomes, with lower log-loss values indicating less uncertainty due to better convergence between prediction probabilities and actual classes. Interpretation of the nonambulatory models is challenging, given the small, imbalanced sample size and similar discharge outcomes for these patients, which likely limited the model's ability to learn from the available nonambulatory patient data.

A growing body of research focuses on development and testing early prediction tools after stroke. Stinear et al[8] provide a detailed review of models predicting the functional and motor-related outcomes, enumerating the strengths and limitations of methods published up to 2019. Several previous studies have developed predictive models for IRF discharge, with most incorporating FAs and therapist evaluations obtained at admission.[24–26] For example, Bland et al[24] use the BBS and FIM walk scores at admission to predict ambulation at IRF discharge according to the 10MWT, with greater sensitivity (91%–94%, household ambulation), but lower specificity (60%–65%, community ambulation), compared to our findings. We have previously developed sensor-free regression models to predict discharge scores using similar PI + FA features and 50 participants from this study with mean average errors of 0.3 m/s, 9.5 points, and 7.4 points for the 10MWT, FIM, and BBS, respectively.[9] The TWIST model[27] is another promising approach for predictions outside of the IRF setting, utilizing age, BBS, and knee extension grade at 1-week poststroke to predict independent walking according to Functional Ambulation Categories at 4, 6, 9, 16, or 26 weeks after stroke, with 83% accuracy across all timepoints. Only, recently, has the research community begun investigating the predictive value of wearable sensor data for similar

prognostic applications.[14,15] However, the utility of sensor data in regression models or long-term outcomes remains uncertain.

Accurately predicting expected posttreatment outcomes early in rehabilitation would improve the discharge planning for clinicians, patients, families, and insurance companies by providing a roadmap of the patient's care needs after leaving the hospital. In this study, sensor features were important predictors for individuals discharged with limited ambulation ability and high risk of falling, providing quantitative measures of movement symmetry (eg, the skewness) and repeatability (eg, sample entropy) for treatment monitoring.[9,28,29] Sensor models could replace or reduce the reliance on FA scores, as less time is needed to collect the data. Consumer-grade devices and an about 5-minute sequence of simple physical activities (brief walking bout, standing, stand-to-sit, sitting, and sit-to-stand) would enable quicker and more frequent evaluations than the longer and more complex standardized FAs. The assessments considered in this study (10MWT, 6MWT, BBS, TUG, and FIM) are typically collected during IRF admission in the USA for clinical evaluation and insurance reporting. However, completing these assessments upon admission can be challenging due to time limitations during intake/treatment and varied patient impairments, including fatigability and physical or cognitive deficits.

Our results should be considered in context of previous findings for clinical machine learning models—namely that appropriate choices of target population,[15] activities,[14] sensor modalities,[14] and prediction outcomes[5] are paramount to design a successful model.[6] For instance, in the case of patients were ambulatory, IMU data from the 10MWT and BBS were less impactful for predicting discharge independence, as defined by the FIM motor subscore. This is unsurprising, considering that FIM motor assessment evaluates a breadth of functional activities—including walking, stair climbing, transfers, dressing, bathing, grooming, toileting, and bowel or bladder management—and some of these activities may not be well characterized by gait or balance movements at IRF admission. Sensor features from other physical activities may better capture biomarkers of motor independence according to the FIM. Similarly, predictions for patients who were nonambulatory did not significantly benefit from sensor data, revealing the need for alternative modeling approaches for patients with severe gait impairment.

Limitations

The number of incorrect predictions is a primary limitation of the models presented in this study. Indeed, a naive model predicting no change in the outcome classification from admission to discharge would generally perform well for this study sample since only a fraction of the patients changed classes in our study (ie, 9%–50% of patients who were ambulatory, or 0%–25% of patients who were nonambulatory patients, depending on the outcome). However, such a model will always fail to identify individuals who improved functional classes, who are arguably the most difficult and clinically meaningful cases to predict. By contrast, our models could identify some individuals who improved in the independence (10 out of 16) and risk of falling (5 out of 8) functional classes. The small and unbalanced populations in our single-site study may limit the sensitivity, generalizability, and utility of the proposed models, with a potential risk of overfitting in these high-dimensional feature sets. Larger sample sizes, particularly for patients who are nonambulatory at admission and achieve heterogenous discharge outcomes, will be crucial to further train and validate sensor-based prediction models.

Future work should also investigate sensor regression models that predict continuous outcome scores at discharge rather than classification models that predict categories based on a cut-off score. Regression models may offer greater clinical utility by removing reliance on predefined classification boundaries and providing higher-resolution discharge predictions, though possibly with greater sensitivity to error.[6,14] Alternative clinical outcomes (eg, Fugl-Meyer Assessment), sensor placements (eg, upper limbs), and functional abilities (eg, endurance) should also be considered in these models for various clinical applications.

We did not evaluate other machine learning algorithms, which may outperform L1-penalized logistic regression. Rather, we sought to understand the relative value of sensor data using a single, well-performing and interpretable algorithm for each of these outcomes and patient groups. Alternative algorithms and extended hyperparameter tuning could improve the prediction performance shown here.

A potential disadvantage of models trained to predict outcomes at hospital discharge is the use of hospital- and care-specific data. Because treatment strategies and patient characteristics can vary nationally and internationally, a model trained using data from one location may not generalize to others. For example, the PREP2 model[13,30]—which demonstrated 75% accuracy in New Zealand for categorizing 3-month upper limb function after 1-week poststroke—had drastically lower accuracy for patients in the USA and Europe.[28,29] This highlights the necessity for additional testing and external validation to determine whether site-specific training data are essential for prediction models, or whether combined training data from multiple sites would broaden the generalization across IRFs.

Conclusions

This study affirms that motion-based measures from wearable sensors can be beneficial for predicting certain patient outcomes following acute poststroke rehabilitation. We have highlighted the potential and open challenges of moving these machine learning algorithms into clinical practice to inform tailored and effective rehabilitation therapies. While sensor-based models may increase predictive performance, additional research is needed to refine and validate these models for new patients and IRF settings.

Author Contributions

Conception and design: M. K. O'Brien, R. L. Lieber, A. Jayaraman
Administrative support: R. L. Lieber, A. Jayaraman
Provision of study materials or patients: R. L. Lieber, R. Ghaffari, J. A. Rogers, A. Jayaraman
Collection and assembly of data: M. K. O'Brien
Data analysis and interpretation: M. K. O'Brien., F. Lanotte, R. Khazanchi, S. Y. Shin
Manuscript writing: M. K. O'Brien, F. Lanotte, R. Khazanchi, S. Y. Shin, R. L. Lieber, R. Ghaffari, J. A. Rogers, A. Jayaraman
Final approval of manuscript: M. K. O'Brien, F. Lanotte, R. Khazanchi, S. Y. Shin, R. L. Lieber, R. Ghaffari, J. A. Rogers, A. Jayaraman

Acknowledgments

The authors thank Nsude Okeke Ewo, Alexander J. Boe, Marco Hidalgo-Araya, Sara Prokup, Matthew Giffhorn, Kelly McKenzie, Kristen Hohl, and Matthew McGuire for their help in patient recruitment and data collection.

Ethics Approval

All individuals (or a proxy) provided written informed consent, and the study was approved by the Institutional Review Board of Northwestern University (Chicago, IL; STU00205532).

Funding

This work was supported by the Shirley Ryan AbilityLab, with partial support from the National Institutes of Health under institutional training grants at Northwestern University (T32HD007418 to M.K.O.), center grant to establish the Center for Smart Use of Technology to Assess Real-world Outcomes (C-STAR, P2CHD101899 to R.L.L.), and the National Institute on Aging of the NIH (R43AG067835 to R.L.L.). This work was also supported in part by Research Career Scientist Award from the US Department of Veterans Affairs Rehabilitation R&D Service (IK6 RX003351 to R.L.L.).

Disclosures

The authors completed the ICMJE Form for Disclosure of Potential Conflicts of Interest and reported no conflicts of interest.

References

1. Centers for Disease Control and Prevention (CDC). Prevalence and most common causes of disability among adults—United States, 2005. *MMWR Morb Mortal Wkly Rep*. 2009;58:421–426.
2. Le Danseur M. Stroke rehabilitation. *Crit Care Nurs Clin North Am*. 2020;32:97–108. https://doi.org/10.1016/j.cnc.2019.11.004.
3. Brandstater ME, Shutter LA. Rehabilitation interventions during acute care of stroke patients. *Top Stroke Rehabil*. 2002;9:48–56. https://doi.org/10.1310/YGAX-X5VK-NHVD-HGPA.
4. *Report to the Congress: Medicare Payment Policy – MedPAC*. Medicare Payment Advisory Commission. Washington, DC; March 2022. Accessed January 11, 2024. https://www.medpac.gov/document/march-2022-report-to-the-congress-medicare-payment-policy/.
5. Kwah LK, Herbert RD. Prediction of walking and arm recovery after stroke: a critical review. *Brain Sci*. 2016;6:53. https://doi.org/10.3390/brainsci6040053.
6. Campagnini S, Arienti C, Patrini M, Liuzzi P, Mannini A, Carrozza MC. Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: a systematic review. *J Neuroeng Rehabil*. 2022;19:1–22. https://doi.org/10.1186/s12984-022-01032-4.
7. Harvey RL. Predictors of functional outcome following stroke. *Phys Med Rehabil Clin N Am*. 2015;26:583–598. https://doi.org/10.1016/j.pmr.2015.07.002.
8. Stinear CM, Smith MC, Byblow WD. Prediction tools for stroke rehabilitation. *Stroke*. 2019;50:3314–3322. https://doi.org/10.1161/STROKEAHA.119.025696.
9. Harari Y, O'Brien MK, Lieber RL, Jayaraman A. Inpatient stroke rehabilitation: prediction of clinical outcomes using a machine-learning approach. *J Neuroeng Rehabil*. 2020;17:1–10. https://doi.org/10.1186/s12984-020-00704-3.
10. Piron L, Piccione F, Tonin P, Dam M. Clinical correlation between motor evoked potentials and gait recovery in poststroke patients. *Arch Phys Med Rehabil*. 2005;86:1874–1878. https://doi.org/10.1016/j.apmr.2005.03.007.
11. Stinear CM, Barber PA, Petoe M, Anwar S, Byblow WD. The PREP algorithm predicts potential for upper limb recovery after stroke. *Brain*. 2012;135:2527–2535. https://doi.org/10.1093/brain/aws146.
12. Rondina JM, Filippone M, Girolami M, Ward NS. Decoding post-stroke motor function from structural brain imaging. *Neuroimage Clin*. 2016;12:372–380. https://doi.org/10.1016/j.nicl.2016.07.014.
13. Stinear CM, Byblow WD, Ackerley SJ, Smith MC, Borges VM, Barber PA. PREP2: a biomarker-based algorithm for predicting upper limb function after stroke. *Ann Clin Transl Neurol*. 2017;4:811. https://doi.org/10.1002/acn3.488.
14. Adans-Dester C, Hankov N, O'Brien A et al. Enabling precision rehabilitation interventions using wearable sensors and machine learning to track motor recovery. *NPJ Digital Medicine*. 2020;3:1–10. https://doi.org/10.1038/s41746-020-00328-w.
15. Lee SI, Adans-Dester CP, Obrien AT et al. Predicting and monitoring upper-limb rehabilitation outcomes using clinical and wearable sensor data in brain injury survivors. *IEEE Trans Biomed Eng*. 2021;68:1871–1881. https://doi.org/10.1109/TBME.2020.3027853.
16. O'Brien MK, Shin SY, Khazanchi R et al. Wearable sensors improve prediction of post-stroke walking function following inpatient rehabilitation. *IEEE J Transl Eng Health Med*. 2022. 10:1–11. https://doi.org/10.1109/JTEHM.2022.3208585.
17. Perry J, Garrett M, Gronley JK, Mulroy SJ. Classification of walking handicap in the stroke population. *Stroke*. 1995;26:982–989. https://doi.org/10.1161/01.STR.26.6.982.
18. Bowden MG, Balasubramanian CK, Behrman AL, Kautz SA. Validation of a speed-based classification system using quantitative measures of walking performance poststroke. *Neurorehabil Neural Repair*. 2008;22:672–675. https://doi.org/10.1177/1545968308318837.
19. Alexander MP. Stroke rehabilitation outcome: a potential use of predictive variables to establish levels of care. *Stroke*. 1994;25:128–134. https://doi.org/10.1161/01.STR.25.1.128.
20. Teasell R, Norhayati H, Foley N. Managing the stroke rehabilitation triage process. *Evidence Based Review of Stroke Rehabilitation*. 2018. Accessed January 11, 2024. http://www.ebrsr.com/evidence-review/4-managing-stroke-rehabilitation-triage-process.
21. Berg K, Wood-Dauphinee S, Williams JI. The balance scale: reliability assessment with elderly residents and patients with an acute stroke. *Scand J Rehabil Med*. 1995;27:27–36.
22. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45:427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
23. Murphy KP. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press; 2012.
24. Bland MD, Sturmoski A, Whitson M et al. Prediction of discharge walking ability from initial assessment in a stroke inpatient rehabilitation facility population. *Arch Phys Med Rehabil*. 2012;93:1441–1447. https://doi.org/10.1016/j.apmr.2012.02.029.
25. Scrutinio D, Lanzillo B, Guida P et al. Development and validation of a predictive model for functional outcome after stroke rehabilitation. *Stroke*. 2017;48:3308–3315. https://doi.org/10.1161/STROKEAHA.117.018058.
26. Henderson CE, Fahey M, Brazg G, Moore JL, Hornby TG. Predicting discharge walking function with high-intensity stepping training during inpatient rehabilitation in nonambulatory patients poststroke. *Arch Phys Med Rehabil*. 2022;103:S189–S196. https://doi.org/10.1016/j.apmr.2020.10.127.
27. Smith M-C, Barber AP, Scrivener BJ et al. The TWIST tool predicts when patients will recover independent walking after stroke: an observational study. *Neurorehabil Neural Repair*. 2019;2022:461–471.

28. Barth J, Waddell KJ, Bland MD, Lang CE. Accuracy of an algorithm in predicting upper limb functional capacity in a United States population. *Arch Phys Med Rehabil*. 2022;103:44–51. https://doi.org/10.1016/j.apmr.2021.07.808.
29. Lundquist CB, Nielsen JF, Arguissain FG, Brunner IC. Accuracy of the upper limb prediction algorithm PREP2 applied 2 weeks poststroke: a prospective longitudinal study. *Neurorehabil Neural Repair*. 2021;35:68–78. https://doi.org/10.1177/1545968320971763.
30. Smith MC, Ackerley SJ, Barber PA, Byblow WD, Stinear CM. PREP2 algorithm predictions are correct at 2 years poststroke for most patients. *Neurorehabil Neural Repair*. 2019;33:635–642. https://doi.org/10.1177/1545968319860481.