Interpretable Machine Learning approach to predict discharge Functional Independence Measure (FIM) scores for Stroke Rehabilitation.

Khush Patel, MD[1]

[1]School of Biomedical Informatics UTHealth, Houston, TX 77030, USA.

Abstract

Background and Purpose

Stroke is the leading cause of disability in the United States. Rehabilitation is vital in stroke recovery. Functional Independence Measure (FIM) is a validated survey instrument comprising of an eighteen-item, seven-level ordinal scale measured at the time of admission and discharge from the rehabilitation center. Predicting all individual 18 items at the time of admission to the rehabilitation center, although difficult, can help plan a better-personalized rehabilitation program and answer the expectations of patients and their families. Explaining the individual item predictions at the patient level can help identify the primary outcome predictors and further individualize the rehabilitation plan.

Methods

The study population consisted of retrospectively collected data from 803 patients (52% male, 45% Caucasian, 18% African American, 79% ischemic stroke) admitted to Memorial Hermann Comprehensive Stroke Center, Houston, Texas, USA. Popular machine learning and deep learning models like Bayesian Ridge Regression, XGBoost, Lightgbm, Random Forest, TabNet were developed. SHAP (Shapley Additive explanations) values were obtained to explain the predictions.

Results

Predictions for all 18 individual items in FIM were obtained. The best-performing model was a chained regression model using Bayesian ridge regression. The uniform mean absolute error for all 18 items was 0.80. Patient-level and population-level interpretability was obtained with the help of SHAP values.

Conclusion

Our findings strongly suggest that although predicting individual items in the FIM instrument is challenging, it can be done using state-of-the-art machine learning models. The predictions, along with the explanations, can help develop a personalized rehabilitation plan.

Introduction

Stroke is the leading cause of disability in the United States (US), with more than 795,000 people experiencing a stroke every year[1,2]. The total cost of stroke in the US was $46 billion between 2014-2015[1,2]. The major component of this cost was rehabilitation services[3] and indirect costs from missed workdays and salary difference[4]. Rehabilitation is a systematic approach to limit and reverse the impact of stroke-related brain damage on daily life[5]. It helps improve patient outcomes, including long-term survival, and as a result, reduces healthcare costs. It has been validated by numerous clinical studies[6]. Inpatient rehabilitation facility (IRFs)[7] is recommended for patients who need ongoing supervision by healthcare staff and have adequate fitness to participate in therapy.

Prediction of functional outcomes at admission to an inpatient rehabilitation facility (IRF) can benefit both the patients and the center. It can help better plan a personalized therapy and better manage the center's resources. It can help answer common questions from the patients and their family members in challenging times. The Functional Independence Measure (FIM)[8,9] is a

validated instrument used to measure disability. FIM score is assessed within 72 hours of the start of a rehabilitation program and 72 hours before the end of the rehabilitation program. It is a common practice to predict discharge time FIM scores at admission[10] to predict functional outcomes.

FIM comprises eighteen items divided into two subscales – motor and cognitive. Motor subscale comprises 13 items like eating, grooming, bathing, and others, and cognition subscale comprises five items like comprehension, expression, and others. Each FIM category has seven levels. 1 indicates total assistance, and 7 indicates complete independence. The individual items within each subcategory, although related, are different, like eating vs. bladder management. Each category provides valuable information. Prior approaches have predicted changes in total FIM score[11–13], total motor or cognitive FIM score[14], or a classification problem predicting whether total discharge FIM score is less than a specific number[15]. Ignoring individual categories and predicting overall cumulative scores loses essential information, although it shows good overall evaluation metrics. Predicting all eighteen individual FIM categories can help better plan the rehabilitation program and improve patient outcomes by providing detailed information. However, predicting eighteen FIM categories with each category having seven levels is quite a challenging problem[16].

With the recent advances in artificial intelligence, advanced predictive modeling techniques have been developed and validated in many data challenges. The previous approaches mainly use limited traditional machine learning approaches like linear and logistic regression[15], support vector machine[14], regression with L1 penalty[12]. We developed and compared current state-of-the-art machine learning and deep learning approaches. Another limitation of prior

approaches was the lack of a reasonable sample size[12]. We have a sufficient size feature-rich dataset.

Recently, there has been much research in explainable AI. We use SHAP[17,18] which uses the game theory approach to explain the model's output at the patient and population levels. The population-level explanations can help understand the contributions of each feature responsible for the predicted score. The individual-level explanations of the predictions can help identify the main contributing features for that individual and develop a better personalized rehabilitation plan.

The primary purpose of this study is to develop a model predicting all 18 items in FIM with an explanation for predicted scores at individual and population levels to aid the personalization of rehabilitation plans.

Methods

The study population consisted of 803 total patients with 637 ischemic stroke patients and 166 hemorrhagic stroke patients. The data was retrospectively collected from Memorial Hermann Health System, Houston, Texas. The baseline characteristics of patients at rehabilitation admission are shown in Table 1. FIM scores measured within 72 hours of the start of a rehabilitation program and the end of the rehabilitation program are shown in figure 1.

Table I: Characteristics of patients admitted to the rehabilitation center.

	Overall n = 803	Ischemic n = 637	Hemorrhagic n = 166
Age at onset			
Mean (SD)	68.02 (13.86)	69.58 (13.30)	62.01 (14.35)
Missing (%)	0 (0.00)		
Sex			
Female (%)	382 (0.48)	307 (0.48)	75 (0.45)
Missing (%)	0 (0.00)		

Marital Status			
Married (%)	391 (0.49)	300 (0.47)	91 (0.55)
Not Married: widow, single, divorced (%)	412 (0.51)	337 (0.53)	75 (0.45)
Missing (%)	0 (0.00)		
Race			
White (%)	361 (0.45)	291 (0.46)	70 (0.42)
Black (%)	143 (0.18)	36 (0.06)	107 (0.64)
Hispanic (%)	17 (0.02)	14 (0.02)	3 (0.02)
Asian (%)	18 (0.02)	11 (0.02)	7 (0.04)
Others (%)	214 (0.27)	168 (0.26)	46 (0.28)
Missing (%)	0 (0.00)		
Laterality			
Left (%)	363 (0.45)	290 (0.46)	73 (0.44)
Right (%)	359 (0.44)	285 (0.78)	74 (0.45)
Bilateral	51 (0.06)	32 (0.05)	19 (0.11)
Missing (%)	0 (0.00)		
Past medical history of Hypertension			
Yes (%)	640 (0.80)	512 (0.80)	128 (0.77)
Missing (%)	59 (0.07)		
Past medical history of Diabetes			
Yes (%)	287 (0.36)	249 (0.39)	38 (0.23)
Missing (%)	67 (0.08)		
Past medical history of Hyperlipidemia			
Yes (%)	359 (0.45)	311 (0.49)	48 (0.29)
Missing (%)	69 (0.09)		
Current smoker			
Yes (%)	175 (0.22)	144 (0.23)	31 (0.19)
Missing (%)	78 (0.10)		
Previous Stroke			
Yes (%)	220 (0.27)	186 (0.29)	34 (0.20)
Missing (%)	80 (0.10)		
TPA* used			
Yes (%)	130 (0.16)	128 (0.20)	2 (0.01)
Missing (%)	85 (0.11)		
IAT* used			
Yes (%)	79 (0.09)	79 (0.12)	0 (0.00)
Missing (%)	88 (0.11)		
CMI*			
Mean (SD)	1.47 (0.55)	1.43 (0.53)	1.61 (0.59)
Missing	0.00		
NIHSS* at admission			
Mean (SD)	8.12 (7.14)	7.53 (6.76)	10.36 (8.03)
Missing	47 (0.06)		
Days from onset to inpatient rehabilitation admit			
Mean (SD)	10.16 (39.92)	9.80 (44.61)	11.56 (8.89)
Missing	40 (0.05)		
Length of stay in inpatient rehabilitation^{+}			
Mean (SD)	16.70 (8.46)	15.87 (6.93)	19.86 (12.25)
Missing	0 (0.00)		
Total therapy days in two weeks duration^{+}			
Mean (SD)	12.38 (2.82)	12.31 (2.87)	12.67 (2.61)

Missing	20 (0.02)

*NIHSS: NIH Stroke Scale/Score (NIHSS) for quantifying stroke severity., TPA: Tissue Plasminogen Activator, IAT: Intra-arterial therapy (IAT), CMI: Cognitive motor interference

[+] length of stay in inpatient rehabilitation and Total therapy days in two weeks duration were not used in model building, as explained later

Figure I: Mean FIM score during the admit and discharge from the rehabilitation center.

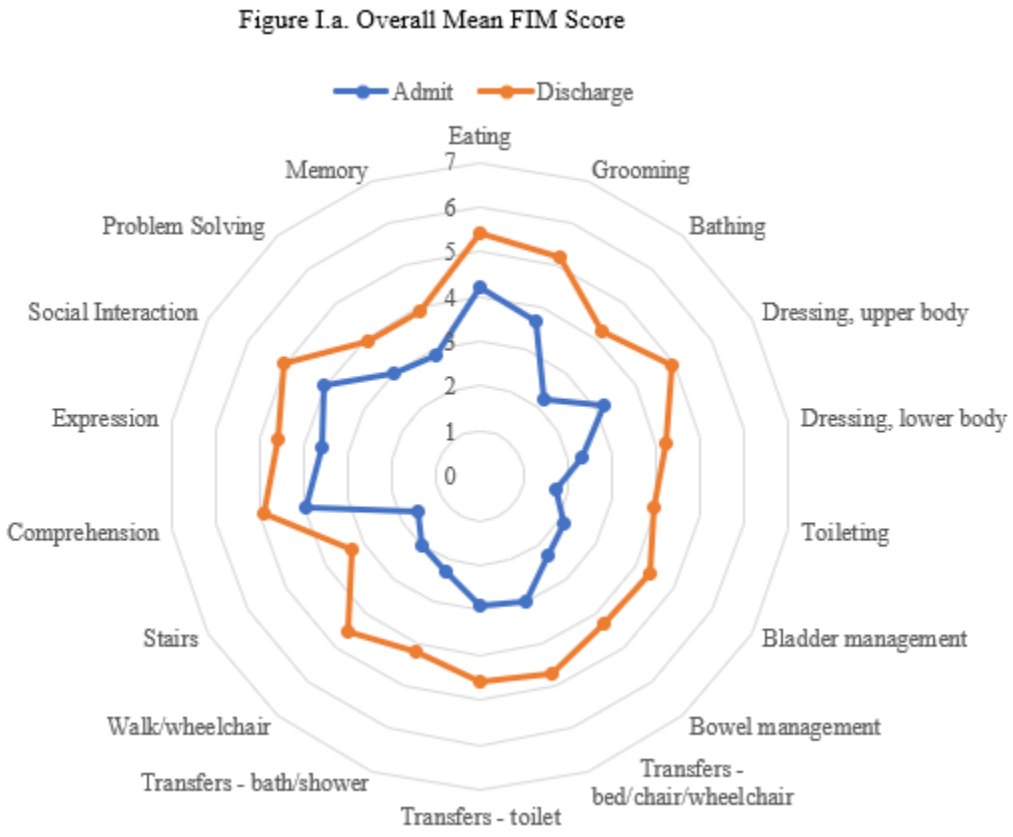


Figure 1.b. Ischemic Stroke: Mean FIM Scores

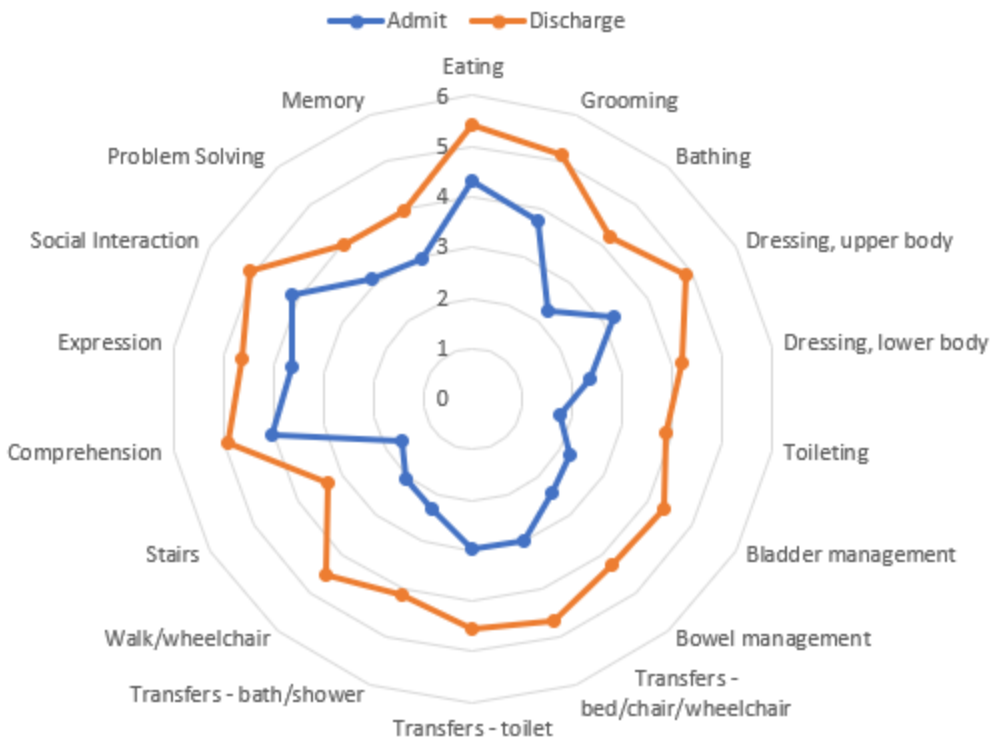
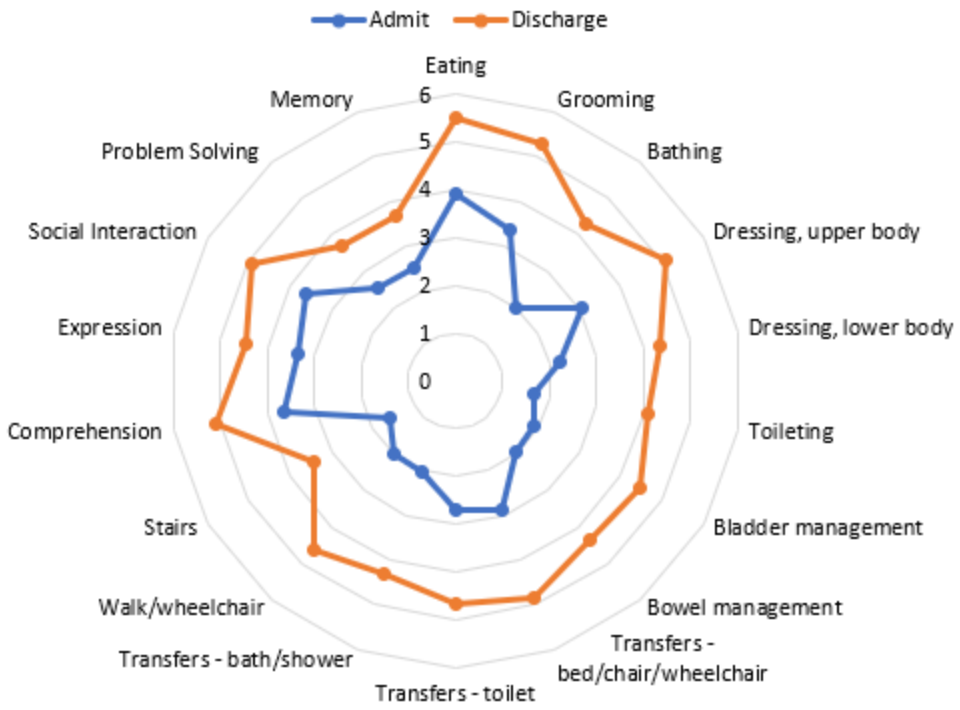


Figure 1.c. Hemorrhagic Stroke: Mean FIM Scores



Data preparation

Data leakage[19] is a challenging problem in developing machine learning models in healthcare. It happens when information outside of training data is used to create a model which may show better results but performs poorly in a real-world scenario. Features that are not present at the time of admission were removed from the data. These include the length of stay in inpatient rehabilitation and total therapy days in two weeks duration. Also, we divided the dataset into train, validation, and test data set after shuffling the data randomly. 20 % of the data was divided into a test data set (161 patients), and 80 % of the data set was divided into the train (577 patients) and validation set (65 patients) in a 9:1 ratio. The validation set was used to test hyperparameters and target sequences for the chained regression model explained later. All the reported results are on the test data set.

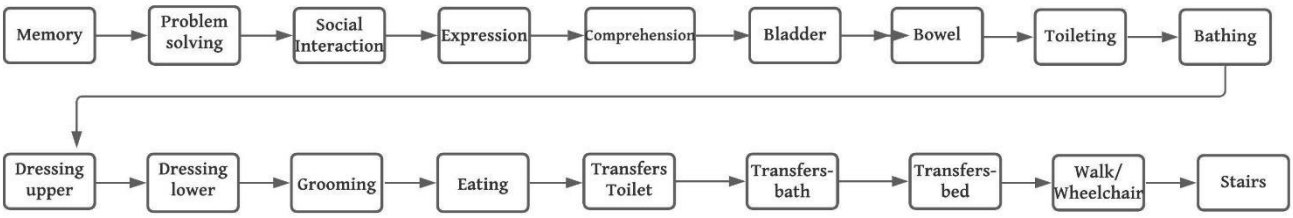
For various reasons, clinical datasets may contain missing values. We included the missing value numbers in table I. We inferred the missing values from known parts of the data using multivariate feature imputation[20]. It models missing values as a function of other features and uses them to impute the missing values. We used Bayesian Ridge Regression[21,22] as the model for multivariate feature imputation. We used Scikit-learn implementation of Bayesian Ridge Regression[23].

Model development

We developed and tested a vast number of current state-of-the-art approaches for developing predictive models. We developed a deep learning model using sequential attention for feature selection at each step called TabNet[24]. We also used gradient boosted, tree-based ensemble approaches like XGBoost[25], LightGBM[26], Random Forest[27], Bayesian approach called Bayesian Ridge Regression[21,22], and traditional approaches like linear regression with and without regularization[28]. Since the individual FIM items are ordinal, we found regression models to perform better than classification models.

Simultaneously predicted 18 FIM items make it a multi-task problem. Some techniques like Random Forest have inherent multioutput prediction capabilities. For others, we can develop a separate model for each item. However, individual FIM items strongly correlate statistically, as shown in the supplementary figure and intuitively. Due to correlation among target features, we use predictions from the previous step as input for the next step along with the baseline features in a chained approach. The sequence in which we make predictions is as mentioned in Figure II. For example, to predict expression in the FIM item, our training dataset will have all the baseline features and predictions of Memory, Problem Solving and Social Interactions.

Figure II. The sequence in which predictions are made for discharge FIM scores using a chained regression model.



Explaining predictions

The most common method used before to explain FIM score predictions was from coefficients learned from each feature[13]. They tell us how much the predicted value will change when we change the given feature; however, it is not the best way to measure the overall importance of the feature. We used SHAP values[17,18] based on Shapley values, a concept based on game theory to explain an individual feature's contribution to the predicted value at the patient and global data set levels.

Computational resources

We used python 3.7 along with packages scikit-learn 0.24[23], Pytorch 1.7.0 [29], Pytorch implementation of TabNet. We used Optuna 2.5[30] for hyperparameter tuning using Tree-structured Parzen Estimator (TPE) algorithm, a Bayesian approach to hyperparameter tuning. We used NVIDIA 2080 11 GB VRAM to run Pytorch deep learning framework.

Results

Predictive performance of the model

The predictions for individual FIM items were obtained. We calculated and reported the mean absolute error for individual FIM items. The best performing model was the chained regression model using Bayesian Ridge Regression with a mean absolute error of 0.80. Table II reports the

mean predicted scores and mean absolute error for each item. Figure III plots the predicted mean FIM scores for each item against the actual values and the admission FIM scores. Table III in the supplement reports mean absolute error using different machine learning and deep learning approaches.

Table II: Test data set: mean admission FIM scores, mean discharge FIM scores (actual values), mean predicted FIM scores, mean absolute error using chained Bayesian Ridge Regression.

Items	**mean admission FIM scores**	**mean discharge FIM scores changes (actual values)**	**mean predicted discharge FIM scores**	**Mean absolute error**
Eating	4.02	5.3	5.29	0.78
Grooming	3.66	5.13	5.08	0.66
Bathing	2.22	4.12	4.13	0.62
Dressing, upper body	3.17	4.69	4.84	0.72
Dressing, lower body	2.25	4.03	4.14	0.88
Toileting	1.55	3.67	3.89	1.17
Bladder management	2.02	4.02	4.37	1.30
Bowel management	2.20	4.06	4.28	1.18
Transfers - bed/chair/wheelchair	2.94	4.54	4.66	0.63
Transfers - toilet	2.88	4.44	4.53	0.68
Transfers - bath/shower	2.28	4	4.16	0.63
Walk/wheelchair	1.91	4.49	4.52	0.95
Stairs	1.55	3.09	3.25	1.04
Motor FIM score	**32.65**	**55.58**	**57.14**	
Comprehension	3.97	4.86	4.9	0.57
Expression	3.55	4.5	4.57	0.61
Social Interaction	4.07	4.99	5.11	0.63
Problem Solving	2.96	3.87	3.92	0.67
Memory	2.91	3.89	3.9	0.70
Cognitive FIM score	**17.46**	**22.11**	**22.4**	
Total	**50.11**	77.69	79.54	**0.80***

Figure III: Test data set admission FIM scores, discharge FIM scores, and predicted discharge FIM scores using chained Bayesian Ridge Regression model. Test data set, n=161 (0.20% of the randomly shuffled entire dataset)

Interpretability:

A: Individual-level explanation of model:

A random patient from the test set had been selected to understand the contribution of features for the predicted score for comprehension. The predicted value for the change in FIM item Comprehension at discharge from the model was calculated as 0.70 with a base value of around 0.90 (model with no features). Comprehension is at the fifth position in the regression chain, and the prior predicted values in the chain used in training the model were predicted memory, problem-solving, social interaction, expression, and baseline features. Features in decreasing order of importance were Comprehension score at admission, problem-solving score at admission, expression score at admission, memory score at admission, predicted memory score,

and predicted expression score at discharge from prior models in the chain, Initial NIHSS score. Figure IV A. explains the scores. Supplementary Figures contain individual-level explanations for other FIM items.

B: Population-level explanation of model:

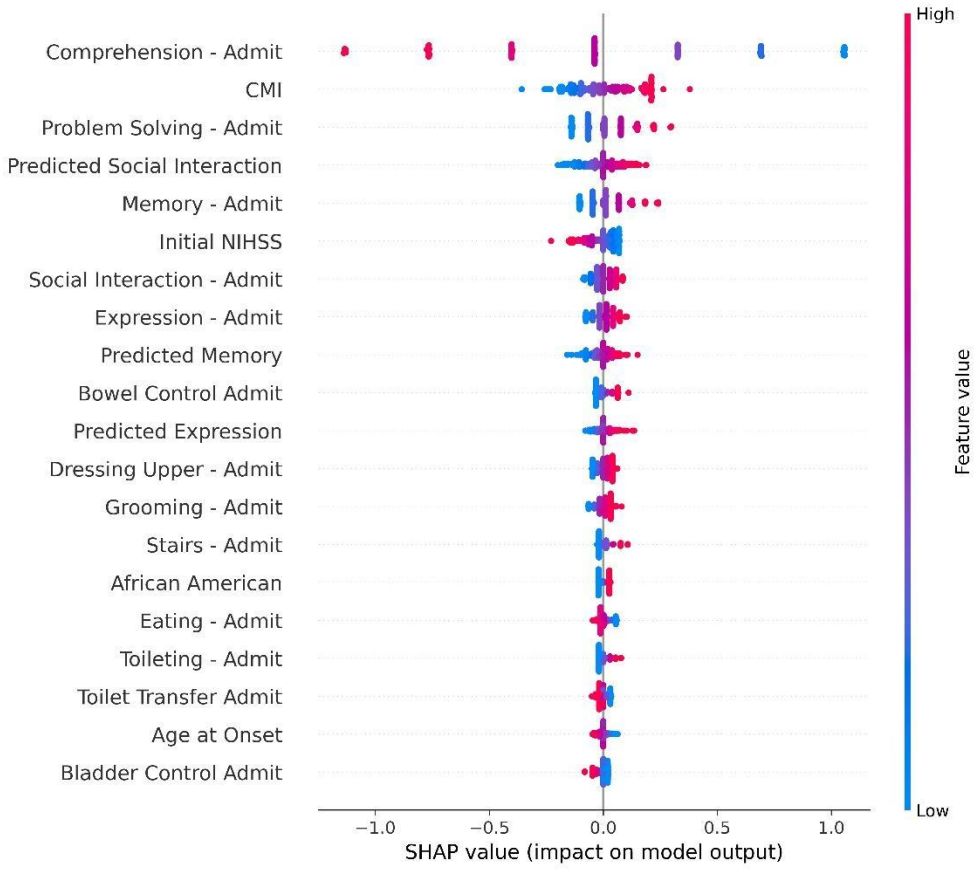
The complete test dataset was used to generate SHAP values to understand the contribution of each feature in predicting change in comprehension FIM score. Population-level explanation helps to understand the model better. Features in decreasing order of importance were Comprehension score at admission, CMI, problem-solving score at admission, predicted social interaction, memory score at admission, Initial NIHSS score, social interaction score at admission, expression score at admission, predicted memory score. Figure IV B. explains the scores. Supplementary Figures contain global level explanations for other FIM items.

Figure IV: Explaining the model:

A: Individual-level explanation of model:

Explaining predicted changes in Comprehension score for a random patient from test data set. Red indicates negatively contributing, and blue indicates positively contributing to the final predicted value.

B: Global level explanation of model: Explanation of prediction of change in Comprehension score at discharge for patients in test data set. Red shows negatively contributing, and blue shows positively contributing to the final predicted value. Each dot represents a patient.



Discussion:

We developed a machine learning approach to predict all 18 items in the FIM instrument. The model takes baseline features like demographics, past medical history, stroke episode data like NIHSS, along with predicted FIM scores in a chained fashion. We provided explanations for predicted changes in FIM score for each item at the patient and population levels.

Low mean absolute errors (MAE) obtained for individual items shows that our modeling approach reasonably predicted individual items in the FIM. Figure III shows that our mean predictions for each item are close to actual values. We found that cognitive FIM items have lower MAE than motor FIM items. A high statistical correlation between cognitive FIM items helps take advantage of the chained structure of the model.

Comparing predictive performance on a private dataset is not possible. We highlighted the importance of predicting individual items to improve predictive performance and provide meaningful information.

We explained predictions using SHAP values at both patient-level and population levels. When we obtain the contribution of factors for predicting change in comprehension FIM score as shown in Figure IV, we see that the admission level FIM score for comprehension was the most important contributor. Items like problem-solving, social interaction, memory, and stroke characteristics like CMI and NIHSS scores were essential. These explained results are in concurrence with clinical interpretation. The rehabilitation center can assess each item's score and the way it is contributing to the prediction of individual patient outcomes. This can help in the creation of a personalized rehabilitation plan.

Limitations:

However, our study does have some limitations. The data is collected from a single health system located in one city, which helps uniformity in data collection and treatment. However, for machine learning models to have generalizability, it is essential to have training data from different health systems and remove any unrecognized biases arising from the data. It will also help to validate the models. Our data is collected from the US-based hospital where the median length of stay is 16 days[7]. Our mean length of stay was 16.70 days. However, the median length of stay for other countries is different. For European countries, it could be 44.5 days in the UK to around 66 days in Belgian[31]. This makes it essential to collect data from IRFs throughout the world to generalize the model better.

The study design is a retrospective cohort study that relies on accurate record entries in the past, which may introduce information bias. Also, the data has been extracted by trained research professionals; however, there could be an error in the data entered.

Conclusion

18 items 7 level FIM score prediction, although a challenging problem, can be achieved using our machine learning-based model. The predictions and their explanation at the patient and population levels can help develop a better personalized rehabilitation plan and efficiently manage rehabilitation facility resources.

Disclosures

None.

References

1. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, et al. Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association. *Circulation*. 2020;141:e139–e596.

2. Stroke Facts | cdc.gov [Internet]. 2021 [cited 2021 Jun 27];Available from: https://www.cdc.gov/stroke/facts.htm

3. Rajsic S, Gothe H, Borba HH, Sroczynski G, Vujicic J, Toell T, Siebert U. Economic burden of stroke: a systematic review on post-stroke care. *Eur J Health Econ*. 2019;20:107–134.

4. Girotra T, Lekoubou A, Bishu KG, Ovbiagele B. A contemporary and comprehensive analysis of the costs of stroke in the United States. *Journal of the Neurological Sciences* [Internet]. 2020 [cited 2021 Jun 27];410. Available from: https://www.jns-journal.com/article/S0022-510X(19)32408-6/abstract

5. Young J, Forster A. Review of stroke rehabilitation. *BMJ*. 2007;334:86–90.

6. Organised inpatient (stroke unit) care for stroke. *Cochrane Database Syst Rev*. 2013;2013:CD000197.

7. Dobkin BH. Rehabilitation after Stroke. *N Engl J Med*. 2005;352:1677–1684.

8. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. *Arch Phys Med Rehabil*. 1993;74:531–536.

9. Ring H, Feder M, Schwartz J, Samuels G. Functional measures of first-stroke rehabilitation inpatients: usefulness of the Functional Independence Measure total score with a clinical rationale. *Arch Phys Med Rehabil*. 1997;78:630–635.

10. Chumney D, Nollinger K, Shesko K, Skop K, Spencer M, Newton RA. Ability of Functional Independence Measure to accurately predict functional outcome of stroke-specific population: Systematic review. *JRRD*. 2010;47:17.

11. Brown AW, Therneau TM, Schultz BA, Niewczyk PM, Granger CV. Measure of Functional Independence Dominates Discharge Outcome Prediction After Inpatient Rehabilitation for Stroke. *Stroke*. 2015;46:1038–1044.

12. Harari Y, O'Brien MK, Lieber RL, Jayaraman A. Inpatient stroke rehabilitation: prediction of clinical outcomes using a machine-learning approach. *Journal of NeuroEngineering and Rehabilitation*. 2020;17:71.

13. Meyer MJ, Pereira S, McClure A, Teasell R, Thind A, Koval J, Richardson M, Speechley M. A systematic review of studies reporting multivariable models to predict functional outcomes after post-stroke inpatient rehabilitation. *Disabil Rehabil*. 2015;37:1316–1323.

14. Sprint G, Cook DJ, Weeks DL, Borisov V. Predicting Functional Independence Measure Scores During Rehabilitation With Wearable Inertial Sensors. *IEEE Access*. 2015;3:1350–1366.

15. Scrutinio D, Lanzillo B, Guida P, Mastropasqua F, Monitillo V, Pusineri M, Formica R, Russo G, Guarnaschelli C, Ferretti C, et al. Development and Validation of a Predictive Model for Functional Outcome After Stroke Rehabilitation: The Maugeri Model. *Stroke*. 2017;48:3308–3315.

16. Xu D, Shi Y, Tsang IW, Ong Y-S, Gong C, Shen X. A Survey on Multi-output Learning. *arXiv:1901.00248 [cs, stat]* [Internet]. 2019 [cited 2021 Jun 30];Available from: http://arxiv.org/abs/1901.00248

17. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]* [Internet]. 2017 [cited 2021 Jul 1];Available from: http://arxiv.org/abs/1705.07874

18. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2:749–760.

19. Chiavegatto Filho A, Batista AFDM, dos Santos HG. Data Leakage in Health Outcomes Prediction With Machine Learning. Comment on “Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning.” *J Med Internet Res*. 2021;23:e10969.

20. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*. 2015;15:30.

21. Tipping ME. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* 2001;1:211–244.

22. MacKay DJC. Bayesian Interpolation. *Neural Computation*. 1992;4:415–447.

23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.

24. Arik SO, Pfister T. TabNet: Attentive Interpretable Tabular Learning. *arXiv:1908.07442 [cs, stat]* [Internet]. 2020 [cited 2021 Jul 2];Available from: http://arxiv.org/abs/1908.07442

25. XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. [cited 2021 Jul 2];Available from: https://dl.acm.org/doi/10.1145/2939672.2939785

26. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 3149–3157.

27. Ho TK. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1998;20:832–844.

28. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996;58:267–288.

29. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* [Internet]. 2019 [cited 2021 Jul 2];32. Available from: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

30. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv:1907.10902 [cs, stat]* [Internet]. 2019 [cited 2021 Jul 2];Available from: http://arxiv.org/abs/1907.10902

31. De Wit L, Putman K, Schuback B, Komárek A, Angst F, Baert I, Berman P, Bogaerts K, Brinkmann N, Connell L, et al. Motor and functional recovery after stroke: a comparison of 4 European rehabilitation centers. *Stroke*. 2007;38:2101–2107.

SUPPLEMENT

Table III. Comparison of machine learning models in terms of mean absolute error (MAE)

FIM items	Chained Bayesian Ridge Regression	Bayesian Ridge Regression one model per target	Chained Linear Regression	Chained Lasso Regression (Linear model with L1 penalty)	Chained XGBoost Regression	Random Forest Regressor with inherent support for multioutput predictions	Chained Lightgbm Regressor	TabNet with inherent support for multi-output predictions
Eating	0.78	0.78	0.77	0.99	0.79	1.04	0.78	0.97
Grooming	0.66	0.67	0.66	0.87	0.80	0.87	0.79	0.83
Bathing	0.62	0.61	0.66	1.09	0.75	0.91	0.63	0.92
Dressing, upper body	0.72	0.72	0.75	1.02	0.75	0.87	0.80	0.93
Dressing, lower body	0.88	0.87	0.89	1.15	0.93	0.92	0.98	1.10
Toileting	1.17	1.19	1.18	1.63	1.20	1.19	1.20	1.42
Bladder management	1.30	1.29	1.33	1.56	1.39	1.27	1.61	1.51
Bowel management	1.18	1.19	1.19	1.58	1.29	1.22	1.42	1.24
Transfers - bed/chair/ wheelchair	0.63	0.65	0.65	0.84	0.81	0.84	0.78	0.86
Transfers - toilet	0.68	0.66	0.73	0.9	0.81	0.86	0.72	0.84
Transfers - bath/shower	0.63	0.66	0.70	1.11	0.80	0.88	0.74	1.09
Walk/wheelchair	0.95	0.99	1.04	1.13	0.93	1.09	0.99	1.10
Stairs	1.04	1.06	1.02	1.48	0.94	0.97	0.99	1.24
Comprehension	0.57	0.57	0.62	0.7	0.67	0.68	0.62	0.78
Expression	0.61	0.62	0.7	0.72	0.78	0.79	0.77	0.80
Social Interaction	0.63	0.65	0.65	0.83	0.86	0.76	0.80	0.91
Problem Solving	0.67	0.73	0.71	0.73	0.88	0.76	0.88	0.82
Memory	0.70	0.70	0.76	0.76	0.92	0.76	0.83	0.86

Uniform average of MAE	0.80	0.81	0.83	1.06	0.90	0.93	0.91	1.01

Figure V: Correlation heatmap for all target features