Stroke

ORIGINAL CONTRIBUTION

Machine Learning Predictions of Recovery in Bilingual Poststroke Aphasia: Aligning Insights With Clinical Evidence

Manuel Jose Marte, MS; Erin Carpenter, MS; Michael Scimeca, MS; Marissa Russell-Meill, MS; Claudia Peñaloza, PhD; Uli Grasemann, PhD; Risto Miikkulainen, PhD; Swathi Kiran, PhD

BACKGROUND: Predicting treated language improvement (TLI) and transfer to the untreated language (cross-language generalization, CLG) after speech-language therapy in bilingual individuals with poststroke aphasia is crucial for personalized treatment planning. This study evaluated machine learning models to predict TLI and CLG and identified the key predictive features (eg, patient severity, demographics, and treatment variables) aligning with clinical evidence.

METHODS: Forty-eight Spanish-English bilingual individuals with poststroke aphasia received 20 sessions of semantic feature-based naming treatment in either their first or second language. Comprehensive language, cognitive, and background bilingual experience assessments were administered pre- and post-treatment. Sixteen curated features spanning demographics, language abilities, cognition, and bilingual experience were used as inputs to 6 machine learning algorithms to predict treatment responders versus nonresponders and CLG vs no CLG.

RESULTS: The top 2 machine learning models achieved F1 scores of 0.767±0.153 for TLI and 0.790±0.172 for CLG. Interpretability analyses revealed that aphasia severity in the trained language, education, and cognitive performance were key predictors of TLI. Aphasia severity in the untreated language and cognitive performance emerged as influential features of CLG. These aligned with expectations based on prior literature.

CONCLUSIONS: For the first time, machine learning models reveal that factors such as patient severity and demographics predict TLI and CLG after therapy in Spanish-English bilingual individuals with poststroke aphasia. Consideration of both treated and untreated language severity, as well as cognitive assessment performance, when forecasting treatment outcomes in an underserved population such Spanish-English stroke survivors, can meaningfully impact their short-term and long-term clinical care.

GRAPHIC ABSTRACT: A graphic abstract is available for this article.

Key Words: aphasia ■ cognition ■ language therapy ■ machine learning ■ multilingualism ■ stroke

Aphasia, a language disorder usually caused by brain injury, poses significant difficulties for bilingual people with aphasia (bPWA), particularly in the rapidly growing Hispanic/Latino population in the United States. Stroke, a primary cause of aphasia, affects this population disproportionately. Research has found that Hispanic/Latino individuals have a higher risk of stroke and experience lower levels of function after a stroke compared with other racial and ethnic groups.[1,2]

In bPWA, communication abilities can be differentially affected between their first- and second-acquired languages (L1 and L2) due to stroke-related and bilingual background factors.[3] Various approaches, including cognitive-based,[4] cognate-based,[4,5] and semantic feature–based treatments (SFT)[6–11] have shown efficacy in producing treated language (TL) improvement (TLI). Some studies have further shown improvements in naming in the untreated language (UL; hereafter

Correspondence to: Manuel Jose Marte, MS, Center for Brain Recovery, Boston University, 111 Cummington Mall, Boston, MA 02215. Email mjmarte@bu.edu
Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/STROKEAHA.124.047867.
For Sources of Funding and Disclosures, see page XXX.

Stroke is available at www.ahajournals.org/journal/str

ORIGINAL CONTRIBUTION

Nonstandard Abbreviations and Acronyms

AoA	age of acquisition
AQ	aphasia quotient
bPWA	bilingual people with aphasia
CLG	cross-language generalization
L1	first-acquired language
L2	second-acquired language
LDA	linear discriminant analysis
LUQ	Language Use Questionnaire
SFT	semantic feature–based treatment
SHAP	Shapley Additive Explanations
SVC	Support Vector Classifier
TL	treated language
UL	untreated language
TLI	treated language improvement
WAB-R	Western Aphasia Battery–Revised

cross-language generalization; CLG).[8–11] However, individual responses to such interventions exhibit notable variability in both the TL and UL.

It is well established that interindividual variability in TLI in poststroke aphasia is influenced by an array of factors, including, but not limited to, age,[12–15] education,[16–19] aphasia severity,[11,14,15,20,21] baseline cognitive abilities,[13,22,23] and, specific to bPWA, bilingual language experience and language of treatment (eg, whether therapy is delivered in L1 or L2).[24,25] Although several factors influencing language recovery in bPWA have been identified, their joint contribution to outcomes remains poorly understood due to small or heterogeneously sampled studies.

Management of bilingual aphasia is further complicated by the need to optimize treatment effects not only in the TL but also for CLG. Language needs may change due to socio-personal factors, such as the language spoken by caretakers, and clinical factors, such as the degree to which L1/L2 abilities are functionally preserved.[26] The Treatment Effects in Aphasia in Multilingual People model[27] proposes the consideration of bilingual history factors: age of acquisition (AoA), language use, exposure, and proficiency, and stroke-related factors: poststroke language abilities (ie, aphasia severity in each language) and months post-onset (MPO) as determinants of treatment response in multilingual aphasia. Crucially, a recent meta-analysis of 85 multilingual people with aphasia found that both TLI and CLG were most strongly influenced by the AoA of the TL, with the greatest effects observed when the TL was acquired in adulthood. Surprisingly, neither language proficiency nor aphasia severity affected TLI or CLG.[28] However, other studies focusing solely on bPWA have reported conflicting findings regarding the influence of aphasia severity in the treated and UL on TLI and CLG[7,10,29] underscoring the need for further research to clarify the contribution of these factors to treatment outcomes.

Given the sheer number of potential determinants of TLI and CLG in bilingual aphasia, machine learning (ML) emerges as a potential tool for analyzing and predicting the complex interaction of these factors beyond traditional analytical approaches. Recent explorations into ML in speech-language pathology have shown strong predictive performance of TLI using behavioral, demographic, and neuroimaging data.[30] This work demonstrated that top-performing ML models could accurately distinguish between treatment responders and nonresponders in monolingual people with aphasia, reflecting both high precision (ie, accurately identifying true responders among those predicted as responders), high sensitivity (ie, successfully capturing all true responders in its predictions) and not surprisingly, a combination of neuroimaging and behavioral factors predicted the outcome. Studies in aphasia have shown that certain ML algorithms outperform others on outcome prediction,[30,31] suggesting the need to explore a variety of algorithms. Finally, although outputs of such models are limited by a lack of, for example, beta coefficients, explainable ML techniques have recently emerged to mitigate these challenges.[32] However, no previous studies have used the ML approach to examine predictors of treatment outcomes in bPWA.

In this study, we sought to identify the independent and cumulative importance of demographic, baseline language and cognitive impairment, bilingual language experience, and treatment orientation data, to predict TLI and CLG in Spanish-English bPWA. Building on prior research demonstrating the success of SFT and the application of ML in predicting TLI in monolingual aphasia,[31] we investigated the efficacy of 6 different ML algorithms to predict language recovery after SFT. We hypothesized that model performance would be optimized by the combination of demographic, baseline language and cognitive abilities, bilingual language experience, and treatment orientation-based variables compared with single-feature set models. Furthermore, we expected that an explainable ML tool would reveal the relative importance of these variables in predicting TL response and CLG, broadly aligning with the clinical evidence described above for this population.

METHODS

This study was conducted in adherence to the Strengthening the Reporting of Observational Studies in Epidemiology and Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guidelines. This study was not registered. Data may be shared upon reasonable request based on a formal data-sharing agreement. Analytical code is available at https://osf.io/zkau8/

Study Design and Participants

This study involved 48 bPWA (mean age: 53.9 years, SD=15.9; mean MPO: 48.1, SD=82.4; Table S1). Following written informed consent under the Boston University Institutional Review Board approval, participants underwent a battery of standardized language, cognitive (cognitive assessments were administered in L1), and language history assessments. Participants then received 20 sessions of SFT in either their L1 or L2. Language for treatment was determined by randomized controlled trial protocol.[33] Extensive details regarding the randomized controlled trial have been described elsewhere and are summarized in the Supplemental Methods.[33]

Data Preparation and Feature Extraction

To compute treatment response measures, we first calculated the individual change in accuracy on the 3 naming probes administered pre- and post-treatment, in both the TL and UL. To operationalize TLI, we used a median split (55% change), yielding what we hereafter term robust responders (n=25, ≥55% change) and weaker responders (n=23, <55% change). For the UL, anticipating greater variability in CLG, we applied a 1 SD threshold (22% change) to delineate cross-language generalizers (n=13, ≥22% change) and non–cross-language generalizers (n=35, <22% change; see Supplemental Methods for further discussion).

Given the large number of assessments administered to bPWA, dimensionality reduction techniques were used to curate feature sets. One assessment included the Language Use Questionnaire (LUQ),[34,35] a tool designed to capture bilingual language experience patterns in healthy and aphasic bilinguals. First, 0.05% of missing LUQ data were imputed using the MICE[36] package in R (v4.3.2).[37] Then, using the "psych" package,[38] varimax-rotated principal component analyses were conducted on LUQ data for L1 and L2 separately across all bPWA in a dataset[35] including the 48 bPWA discussed here, yielding 2 rotated components per language (Table S2). The component loading scores corresponding to the 48 bPWA were extracted and sorted into TL and UL per randomized controlled trial assignment.

Next, all L1 and L2 assessment measures and subtests collected at pre-treatment[33] were examined via correlation analyses using the *cor()* function in R. Pairwise comparisons resulting in Pearson *r* values ≥0.7 were collapsed into composites or considered separately if below this value (Table S3). This resulted in the following remaining measures: Raven's Colored Progressive Matrices,[39] Cognitive Linguistic Quick Test[40] Symbol Trails, Cognitive Linguistic Quick Test Mazes, Cognitive Linguistic Quick Test Design Generation, Pyramids and Palm Trees,[41] a Naming Composite (Boston Naming Test[42] and an in-house 60-item naming screener), a Repetition Composite (Bilingual Aphasia Test[43] Word and Sentence Repetition subtests), a Morphology Composite (Bilingual Aphasia Test Morphological Opposites and Derivational Morphology subtests), and the Bilingual Aphasia Test Syntactic Comprehension, Auditory Verbal Discrimination, and Grammaticality subtests. These were input into 2 separate principal component analyses for L1 and L2 resulting in 2 components in L1 (1 linguistic and 1 nonlinguistic cognitive component with assessment directions provided in L1) and 1 linguistic component in L2 (Tables S4 and S5). Component loading scores were sorted into TL and UL per randomized controlled trial participant assignment.

Overall, 8 feature sets were used for predictive modeling as noted in Figure 1A: (1) demographics (years of education, age, and MPO), (2) LUQ (4 LUQ components), (3) severity-TL (TL Western Aphasia Battery–Revised [WAB-R][44] aphasia quotient [AQ]), (4) severity-UL (UL WAB-R AQ), (5) TL (linguistic component corresponding to the TL), (6) UL (linguistic component corresponding to the UL), (7) cognition (cognitive component from the assessment principal component analysis), and (8) patient language characteristics (binary variables indicating the language of treatment delivery [1=Spanish; 0=English], whether the TL was the patient's L1 [1=L1; 0=L2], and whether Spanish was the patient's L1 [1=Spanish; 0=English]), capturing the distinction between L1/L2 and the TL for each individual patient.

ML Algorithms

Using the data, scaled, described in Figure 1A as inputs, we used 6 supervised learning frameworks, to classify TLI and CLG outcomes (Figure 1B). These frameworks represent standard approaches in ML, ranging from simple (linear) to complex (nonlinear) models, and have been previously used in related literature (see Supplemental Methods for details). Logistic regression, linear discriminant analysis (LDA), random forest, extreme gradient boosting, support vector machine, and neural network algorithms were written using scikit-learn (v1.3.0)[45] in Python (v3.9.7).[46] Additional Python libraries used to support data analysis and visualization included pandas (2.0.3),[47] numpy (v1.24.4),[48] and matplotlib (v3.7.2).[49]

Cross-Validation and Hyperparameter Tuning

To ensure reliable model performance and generalizability, we used a nested cross-validation procedure[50] (Figure 1B) with 5×-repeated 5-fold cross-validation in the outer loop and 4×-repeated 4-fold cross-validation in the inner loop. This approach selects optimal hyperparameters in the inner loop, and then tests them on the unseen held-out fold in the outer loop. Given the imbalanced class distribution in the CLG analysis (13:35), we used stratified sampling to maintain consistent class distribution across folds.

Hyperparameter tuning was accomplished using Grid Search[51] implemented in scikit-learn. For each algorithm, we defined a hyperparameter search space based on its characteristics, selecting the permutation of hyperparameters that yielded the highest F1 score for each algorithm and feature set combination. With 8 feature sets (Figure 1A), we ran 256 optimized models per algorithm. Further details are in Table S6.

Evaluation Metrics

We evaluated the ML models using 6 metrics: (1) accuracy, which gauges the overall proportion of correct predictions, (2) F1 score, which balances precision and recall, providing a harmonic mean of the 2, (3) precision, which measures the accuracy of positive predictions and the model's ability to minimize false positives, (4) sensitivity (or recall), which assesses the model's capability to correctly identify all true positives, (5) specificity, which quantifies the model's success in recognizing true negatives, and (6) Matthews correlation coefficient, which captures the correlation between observed and predicted

ORIGINAL CONTRIBUTION

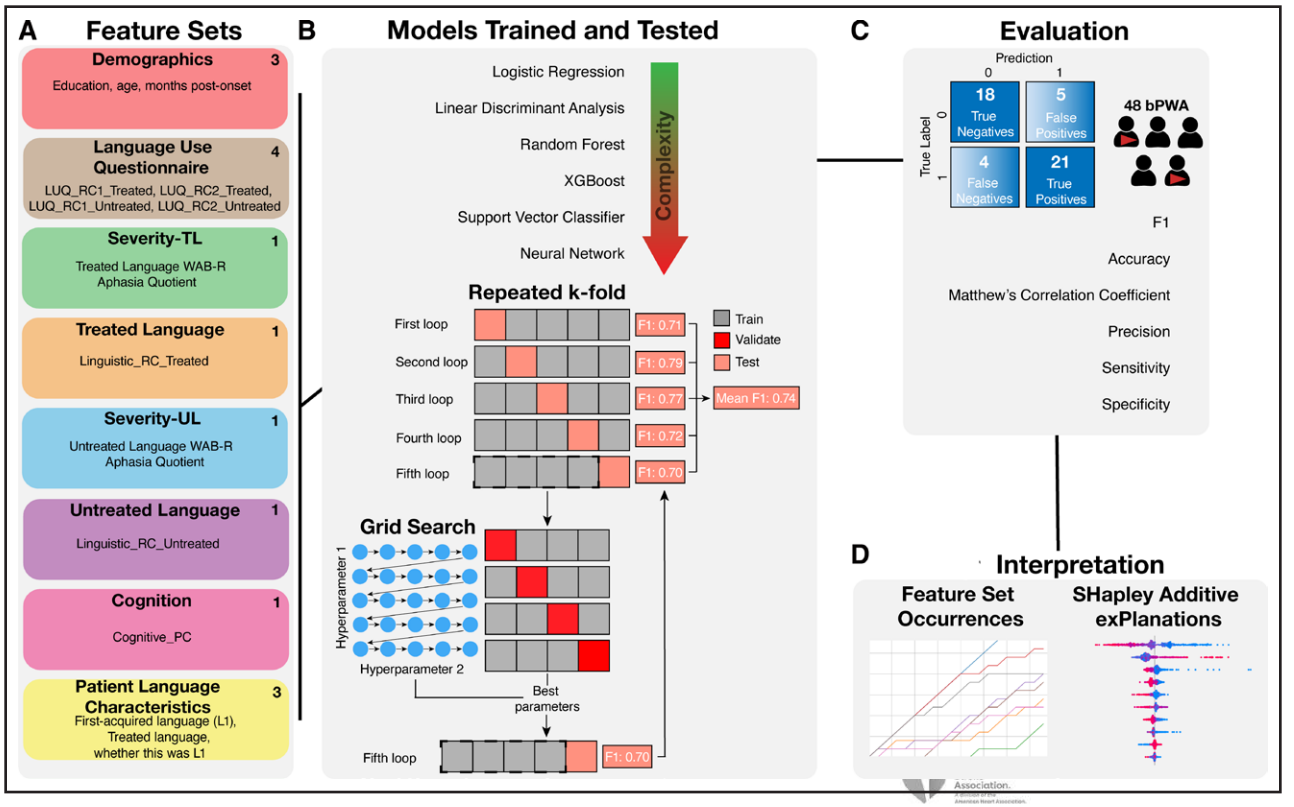

Figure 1. Workflow for evaluating treatment response in bilingual people with aphasia (bPWA).
A, Feature sets: 8 feature sets inform the training of 6 distinct machine learning algorithms. Individual features included in the feature set are noted below the name of the feature set, and the number of dimensions per feature set is noted in the **top-right** corner. **B**, Models trained and tested: the 6 models are trained and tested using a nested cross-validation procedure to validate the optimal hyperparameters and test performance for each individual model. **C**, Evaluation: performance evaluation of optimized models is conducted on 2 classification tasks, determining treated language (TL) improvement and cross-language generalization. **D**, Interpretation: models are interpreted by calculating feature set occurrences and by application of Shapley Additive Explanations values. LUQ indicates Language Use Questionnaire; MCC, Matthews correlation coefficient; PC, principal component; RC, rooted component; UL, untreated language; and WAB-R, Western Aphasia Battery–Revised.

classifications across all 4 categories of the confusion matrix. We present the top 5 models within the top-performing framework ranked by F1 score, as well as the metrics for models including each individual feature set and the model incorporating all feature sets.

Feature Interpretation

To interpret our top-performing ML models, we computed feature set occurrences to determine the regularity with which specific feature set combinations were selected by high-performing models. This involved counting the number of times each feature set was retained by a given algorithm throughout the Grid Search operation, that is, which feature sets played a crucial role in model predictions across the entire spectrum of optimized models, thereby spotlighting the most informative feature sets.

Next, we used Shapley Additive Explanations (SHAP)[32] to reveal the contribution of individual features within the retained feature sets to each prediction made by the top-performing model. SHAP values, which employ a game theoretic approach, assign each feature a value that represents its impact on the model's output, considering the interaction with other features. SHAP values are relative to each other, and the greater SHAP value, the greater the contribution to the overall prediction of TLI or CLG across all participants (Figure 1D).

RESULTS

Performance in TLI Analysis

The analysis of TL performance revealed the Support Vector Classifier (SVC) as the most proficient model in predicting TLI. The SVC model demonstrated a balanced predictive performance for TLI, evidenced by an F1 score of 0.767±0.153. The accuracy of 0.783±0.146 indicates that, on average, the model correctly classified a substantial proportion of instances overall, correctly identifying 38 out of 48 patients as treatment responders or nonresponders. The precision of 0.800±0.136 suggests that when the model predicted a positive treatment response, it was correct in ≈ 80% of cases. To illustrate, if the model made 25 predictions, about 20 would be correct. Similarly, the sensitivity of 0.811±0.136 indicates that the model successfully identified around 81% of all actual robust TLI cases (20 out of 25 responders).

The Matthews correlation coefficient value of 0.606±0.263 suggests a moderate to strong positive correlation between the predicted and actual classifications, confirming that, on balance, the model provides high-quality predictions. However, the specificity of 0.747±0.224 (correctly identifying 17 out of 23 nonresponders) was lower than the other metrics, suggesting that the model may have difficulty in distinguishing instances of weaker TLI, leading to a slightly higher rate of false positives compared with false negatives. Evaluation metrics for SVC models are reported in Table 1; across all models in this analysis, see Tables S7 through S11. Figure 2A visualizes performance metrics for the top-performing SVC model discussed, in addition to the optimal, all feature sets model, and the optimal single-feature set models for each feature set.

Feature Set Occurrences

Next, the evaluation of feature set occurrences within the top-performing SVC model revealed distinct patterns. Demographics and impairment-related feature sets pertaining to the TL, including severity-TL and TL, were all among the most present feature sets. Notably, cognition was also among the most present feature sets in high-performing models. Other feature sets showed relatively fewer occurrences, indicating secondary or specific influence within predictive models for TLI. See Figure S2A and S2B for visualization.

SHAP Value Analysis

The SHAP value analysis applied to the optimal SVC revealed several key findings. The top 5 most informative features, in rank order, were severity-TL, education, age, cognition, and MPO. A lower severity rating (ie, a high WAB-R AQ), greater performance on cognitive assessments, greater years of education, and younger age, were closely linked to predictions of TLI. The inverse held true for higher severity and fewer years of education, though older age was not nearly as impactful as younger age. See Figure 3A for visualization.

Performance in CLG Analysis

The analysis identified the LDA model, incorporating severity-UL, demographics, and cognition, as the optimal model for predicting CLG. The model demonstrated balanced predictive performance, evidenced by an F1 score of 0.790±0.172. An accuracy of 0.850±0.102 indicates that the model correctly classified 41 out of 48 patients as either cross-language generalizers or nongeneralizers. The precision of 0.808±0.170 suggests that, on average, when the model predicted a patient would exhibit CLG, it was correct ≈ 81% of the time (11 out of 13 cases). Similarly, the sensitivity of 0.819±0.178 indicates that the model successfully identified around 82% of actual CLG cases (11 out of 13 cross-language generalizers).

The Matthews correlation coefficient value of 0.624±0.326 suggests a moderate to strong positive correlation between the predicted and actual classifications. Notably, the specificity was higher than the other metrics at 0.884±0.126, indicating that the model was highly effective at correctly identifying true negative cases—on average, correctly classifying 31 out of 35 non–cross-language generalizers. This high specificity suggests that the model is particularly reliable in ruling out patients who are unlikely to exhibit CLG. Evaluation

Table 1. SVC Model Performance Metrics for Top-Performing, All Feature Sets, and Single-Feature Set Models, for the TL Improvement Analysis

Rank	Feature sets	Accuracy	F1	Precision	Recall	MCC	Specificity
1	Demographics, severity-TL, cognition	0.783±0.146	0.767±0.153	0.8±0.136	0.811±0.136	0.606±0.263	0.747±0.224
2	Severity-TL, cognition	0.774±0.165	0.761±0.171	0.79±0.147	0.804±0.144	0.589±0.286	0.792±0.222
3	Demographics, TL	0.778±0.127	0.759±0.139	0.782±0.128	0.802±0.124	0.578±0.243	0.786±0.209
4	Demographics, severity-TL, cognition, patient language characteristics	0.78±0.152	0.759±0.163	0.798±0.137	0.802±0.143	0.593±0.272	0.746±0.241
5	Demographics, TL, UL	0.763±0.123	0.746±0.127	0.766±0.121	0.786±0.121	0.548±0.232	0.803±0.187
45	All feature sets	0.721±0.161	0.703±0.173	0.738±0.154	0.751±0.154	0.486±0.303	0.747±0.205
56	Severity-TL	0.712±0.186	0.698±0.187	0.737±0.152	0.749±0.157	0.483±0.304	0.701±0.249
80	Demographics	0.713±0.141	0.688±0.161	0.71±0.149	0.716±0.169	0.425±0.31	0.674±0.239
146	TL	0.665±0.181	0.654±0.183	0.708±0.157	0.708±0.147	0.413±0.301	0.675±0.242
185	Severity-UL	0.646±0.13	0.632±0.136	0.673±0.118	0.682±0.114	0.352±0.227	0.654±0.202
207	UL	0.638±0.158	0.614±0.16	0.666±0.15	0.672±0.135	0.34±0.275	0.614±0.283
230	Cognition	0.603±0.138	0.581±0.142	0.643±0.133	0.638±0.135	0.276±0.257	0.557±0.22
252	LUQ	0.513±0.198	0.496±0.202	0.527±0.206	0.523±0.208	0.053±0.399	0.506±0.288
255	Patient language Characteristics	0.416±0.154	0.348±0.141	0.394±0.238	0.505±0.106	0.036±0.217	0.64±0.423

LUQ indicates Language Use Questionnaire; MCC, Matthews correlation coefficient; SVC, Support Vector Classifier; TL, treated language; and UL, untreated language.

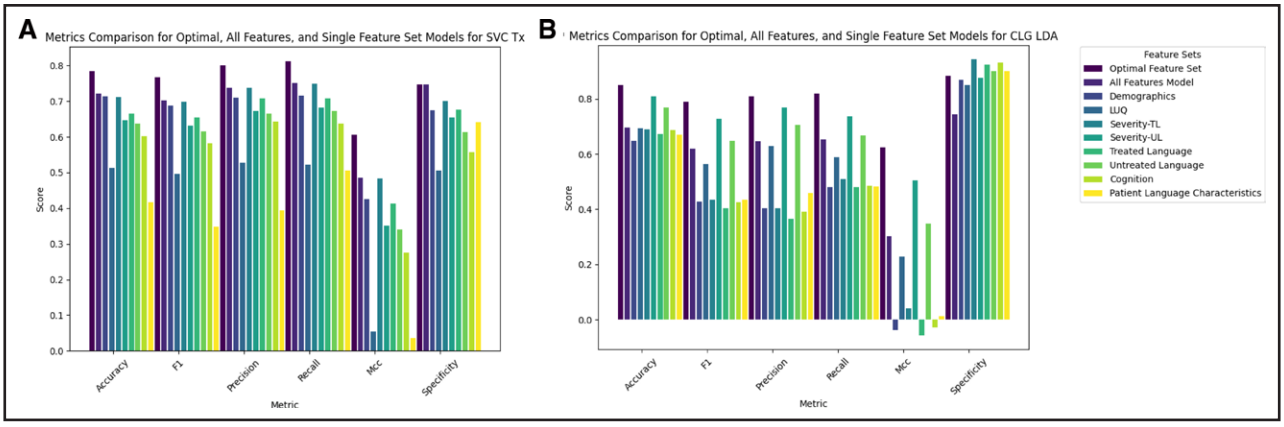

Figure 2. Comparison of metrics across models for treatment response prediction.
A illustrates metric scores using the Support Vector Classifier (SVC) algorithm, whereas **B** does the same for linear discriminant analysis (LDA). The *x* axis categorizes the performance metrics assessed. The *y* axis represents the metric scores. Each bar within a metric category corresponds to the score of an optimal, all feature set, or individual feature set model, with color coding distinguishing between different feature sets and feature set combinations. **A**, The optimal feature set model demonstrates superior performance across most metrics compared with individual feature sets in the SVC algorithm. **B**, A similar pattern is observed for the LDA models, albeit with varying degrees of metric scores. CLG indicates cross-language generalization; MCC, Matthews correlation coefficient; TL, treated language; and UL, untreated language.

metrics for LDA models are reported in Table 2; across all models in this analysis, see Tables S12 through S16. Figure 2B visualizes performance metrics for the top-performing LDA model discussed, in addition to the optimal all feature sets model, and the optimal single-feature set models for each feature set.

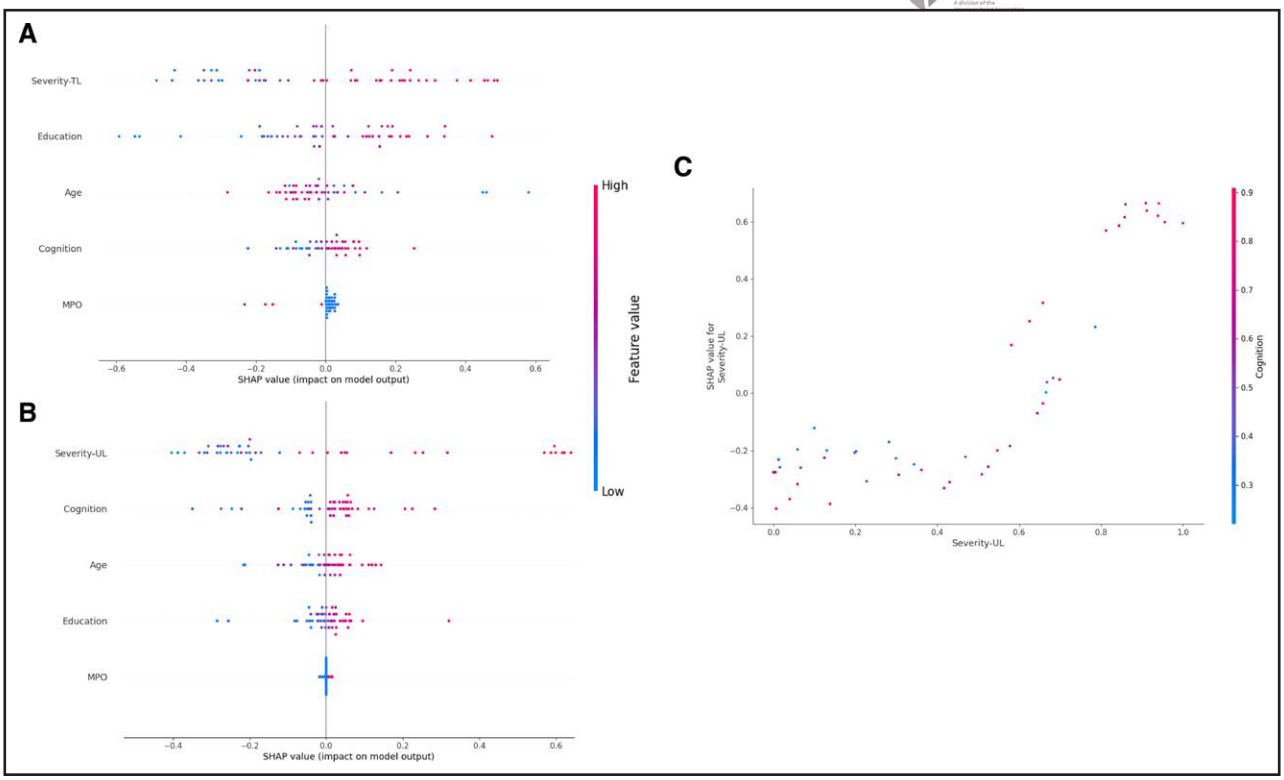

Figure 3. Shapley Additive Explanations (SHAP) value distributions for top-performing models and SHAP dependency plot.
A and **B**, SHAP values demonstrate the influence of each feature on the model's prediction, with values to the **right** indicating a positive impact on predicting robust treated language (TL) improvement (TLI; **A**) or cross-language generalization (CLG; **B**), and values to the **left** suggesting a negative influence. The *y* axis lists the features ranked by the absolute sum of SHAP values. The color gradient signifies the actual feature value, with blue indicating lower and pink indicating higher values. **A** illustrates the SHAP values for the TLI analysis, whereas **B** presents the SHAP values for the CLG analysis. **C**, SHAP dependency plot demonstrating the relationship between aphasia severity in the untreated language (severity-UL) and cognitive performance (cognition). The plot reveals a trend where higher values of severity-UL are associated with higher SHAP values, suggesting an increase in the feature's positive impact on the model's output as aphasia severity in the untreated language increases. MPO indicates months post-onset.

Table 2. LDA Model Performance Metrics for Top-Performing, All Feature Sets, and Single-Feature Set Models, for the Cross-Language Generalization Analysis

Rank	Feature sets	Accuracy	F1	Precision	Recall	MCC	Specificity
1	Demographics, severity-UL, cognition	0.850±0.102	0.790±0.172	0.808±0.17	0.819±0.178	0.624±0.326	0.884±0.126
2	Demographics, severity-UL, UL, cognition	0.817±0.131	0.744±0.206	0.772±0.191	0.755±0.224	0.528±0.397	0.86±0.137
3	Demographics, severity-UL, TL, cognition	0.802±0.117	0.744±0.172	0.77±0.158	0.785±0.188	0.554±0.309	0.826±0.146
4	Severity-UL, cognition	0.811±0.13	0.742±0.194	0.772±0.188	0.757±0.214	0.529±0.375	0.864±0.145
5	Demographics, severity-UL	0.812±0.127	0.738±0.186	0.752±0.174	0.771±0.202	0.519±0.361	0.832±0.126
6	Severity-UL	0.808±0.139	0.726±0.197	0.767±0.193	0.737±0.198	0.504±0.369	0.876±0.16
108	UL	0.767±0.154	0.649±0.223	0.706±0.262	0.667±0.202	0.349±0.393	0.901±0.209
147	All feature sets	0.695±0.179	0.620±0.214	0.645±0.205	0.652±0.239	0.303±0.416	0.744±0.188
207	LUQ	0.695±0.162	0.563±0.176	0.629±0.217	0.589±0.183	0.228±0.338	0.85±0.183
250	Severity-TL	0.688±0.144	0.436±0.121	0.403±0.154	0.508±0.123	0.042±0.19	0.942±0.141
251	Patient language characteristics	0.669±0.202	0.435±0.132	0.460±0.211	0.482±0.139	0.012±0.295	0.9±0.234

LDA indicates linear discriminant analysis; LUQ, Language Use Questionnaire; MCC, Matthews correlation coefficient; TL, treated language; and UL, untreated language.

Feature Set Occurrences

In evaluating feature set occurrences within the top-performing LDA model for CLG, Figure S2C and S2D shows the prominence of severity-UL and LUQ feature sets, highlighting their pivotal role in predicting CLG. Although the cognition, demographics, and severity-TL feature sets were also present among the top 20 SVC models, severity-UL dominated in many occurrences, indicating its significant predictive power.

SHAP Value Analysis for Feature Importance

The SHAP value analysis applied to the optimal LDA model for CLG revealed several key findings (Figure 3B and 3C). The top 5 most informative features, in rank order, were severity-UL, cognition, age, education, and MPO. Lower aphasia severity in the UL (ie, a higher WAB-R AQ score) and better performance on cognitive assessments were linked with predictions of CLG. Surprisingly, older age was also linked to predictions of CLG. As in the previous analysis, more years of education were linked to CLG, and by and large, the inverse held true for higher severity, worse performance on cognitive assessments, younger age, and fewer years of education.

The SHAP dependency plot (Figure 3C) further elucidates the relationship between severity-UL and cognition. The plot reveals that less severe aphasia in the UL is associated with higher SHAP values, indicating a greater positive impact on the model's CLG predictions. This relationship is modulated by cognitive performance, with higher cognitive scores (darker dots; cognition) corresponding to more positive SHAP values, particularly at higher scores. These findings suggest that the interaction between severity-UL and cognition plays a crucial role in determining CLG outcomes, with better cognitive performance potentially enhancing the beneficial effects of residual language function in the UL on CLG.

DISCUSSION

The present study provides the first-ever ML-based insights into the factors that explain treatment gains in bilinguals with poststroke aphasia, particularly in Spanish-English adult speakers. Our findings demonstrate the potential of ML models to accurately predict these outcomes, with the top-performing models attaining F1 scores of 0.766 for response in the TL and 0.790 for CLG. Notably, the optimal models incorporated a combination of primarily demographic and baseline language and cognitive features, rather than relying on a single feature or feature set. Further, using explainable ML, our study sheds light on the patient-specific factors that may determine TLI and CLG, offering new insights for optimizing rehabilitation outcomes in this population.

A key finding of our study is the prominent role of aphasia severity in predicting TLI and CLG outcomes. The interpretability analyses revealed that aphasia severity was closely linked to predictions of TLI and CLG. Although aligning with research in monolingual[11,14,15,20,21] and bilingual aphasia,[7,10,29] our findings diverge from the results of a recent meta-analysis.[28] Although we found that aphasia severity in the TL and UL predicted TLI and CLG, respectively, the aforementioned study[28] reported the AoA of the TL as the primary influence on CLG, with no influence of aphasia severity in either TLI or CLG.

Our study's focus on Spanish-English bilinguals provided a homogeneous sample, allowing for a more direct assessment of aphasia severity's role. In contrast, Goral et al[28] included multilinguals speaking more than 2 languages, potentially conflating aphasia severity effects with prestroke proficiency in additional languages (L3, L4). Moreover, their comparison of L1 to any other language (L*n*) may have diluted severity's impact.

Relatedly, we used single continuous measure of severity (WAB-R AQ), enabling analysis along a severity gradient, whereas Goral et al[28] used ordinal severity coding to collapse information across a variety of clinical instruments, potentially obscuring severity-outcome relationships. Notably, measures of language abilities may vary widely between such studies, as standardized assessments reflect both system damage and individual language mastery.[26,52]

Furthermore, our inclusion of AoA within a principal component that encompasses other aspects of bilingual experience contrasts with Goral et al[28] in the use of discrete life-stage ranges, potentially contributing to divergent findings. In addition, their inability to account for language use and exposure meant that in cases driving their strongest results, late-learned languages had often become primary due to prestroke immersion, suggesting that language use modulated AoA effects and may explain why AoA emerged as a significant factor in their analysis. Our results, on the other hand, align with a parsimonious interpretation of language processing in bilingual aphasia: the integrity of language abilities in the UL facilitates CLG via shared conceptual representations.[53,54]

Expanding on the role of aphasia severity and TLI, a meta-analysis of SFT in monolingual aphasia proposed that in individuals with milder aphasia, the language system is more preserved and responsive, providing the necessary resources to support the relearning of previously acquired but inconsistently accessible word forms and the reestablishment of connections between disrupted and available representations.[11] In SFT, greater language system integrity may facilitate the spread of activation to related semantic nodes, promoting generalization to untreated items and cross-language transfer.[7,11] In contrast, individuals with more severe aphasia may have a more degraded language system, limiting the available resources for treatment-induced recovery. The reduced integrity of the language network may hinder the effective spread of activation, resulting in limited generalization. Indeed, aphasia severity is often correlated with poststroke brain integrity, which ultimately determines the neural resources available to support recovery.[55]

Notably, our findings also emphasize the importance of baseline nonverbal cognitive function in predicting treatment outcomes, particularly for CLG. The SHAP dependency plot showed an interaction between severity-UL and cognition, suggesting that better-preserved cognitive abilities may enhance the beneficial effects of the residual integrity of the UL on CLG. This finding resonates with previous studies demonstrating the impact of cognitive factors, such as attention, memory, and executive function, on treatment-induced language recovery in monolingual aphasia.[13,22,23] The role of cognitive function in bilingual aphasia recovery may be even more pronounced, given the increased demands on cognitive control and language selection mechanisms in bilingual language processing.[56] In our study, the cognitive component included Raven Progressive Matrices, a nonverbal test closely related to executive function, and the Cognitive Linguistic Quick Test Symbol Trails, another measure of executive function, particularly of shifting and cognitive flexibility. These tasks tap into domain-general executive control abilities that are crucial for managing and coordinating multiple languages. Per a recent systematic review of executive control in bPWA, executive control interacts with language, and impairments in these functions may lead to difficulties in compensating for linguistic deficits in bPWA.[56] Thus, the increased demand for domain-general skills in bPWA could explain the observed interaction between severity-UL and cognitive performance in predicting CLG outcomes, in addition to TLI.

Another noteworthy finding is the influence of demographic factors, particularly education, on treatment response. Greater number of years of education were associated with predictions of TLI and CLG, consistent with prior research linking higher education levels to better aphasia recovery.[16–18] Indeed, a recent systematic review of social determinants of health found that while there is no evidence to support a role for education in language outcomes before 12 months poststroke, the limited research available with PWA at or beyond 12 months indicates education may play a role in language outcomes over time.[19] One potential hypothesis is that greater years of education support cognitive reserve,[57] though further research is needed given historically mixed findings.[16–19] See the Supplemental Discussion for further discussion of demographic factors.

Regarding the ML approach, the superior performance of the SVC and LDA models compared with other algorithms, including the neural network, may be due to its ability to find robust decision boundaries with limited data. The limited sample size might have been insufficient for neural networks to capture complex patterns without overfitting, leading to poorer generalization; future studies with larger samples could determine if an SVC or LDA advantage persists, or if neural networks improve. See the Supplemental Discussion for discussion of algorithmic performance.

Next, it is notable that when using a common predictor (WAB-R AQ), our models performed similarly to those in the most comparable study[30] (F1 scores of 0.842 for SVC and 0.771 for random forest models). Furthermore, our finding that ML models incorporating a combination

 Stroke. 2025;56:00–00. DOI: 10.1161/STROKEAHA.124.047867

of features outperform single-feature models aligns with previous research in monolingual aphasia.[30,31] Notably, our models' performance (F1 scores of 0.767 for TLI and 0.790 for CLG) approached that of Billot et al,[30] despite their inclusion of neuroimaging data. Their top 20 SVC and random forest models all incorporated neuroimaging-based features, highlighting the importance of such data in achieving higher performance (F1 scores of 0.941 for SVC and 0.873 for random forest).

In light of these findings, a promising future direction would be to incorporate neuroimaging data, which may provide critical information on the functional integrity of the language system and—unique to bilinguals—the interaction between L1 and L2 language networks and the concomitant integrity of control regions that can influence language processing and recovery in bPWA.[58–60]

Furthermore, using an explainable ML tool (ie, SHAP), we observed the critical influence of UL severity on CLG and its interaction with cognitive factors, aspects overlooked in recent meta-analyses.[28] Relatedly, our cognition feature was also a key predictor in both the TLI and CLG analyses independently. The influence of cognitive factors on treatment outcomes has been found in monolingual aphasia studies,[13,22,23] with stronger skills conferring better outcomes. These skills may be even more crucial in bilingual contexts due to the interaction between heightened executive control demands and the neural infrastructure supporting bilingual lexical access.[56,58–60] These ML-derived insights, consistent with theoretical and clinical evidence, emphasize the need to consider both language-specific and domain-general cognitive abilities in bPWA assessment and intervention.

Although our study provides novel insights into the prediction of treatment response in bPWA, it is essential to acknowledge several limitations. This study represents the largest homogeneous sample in a single treatment study of bPWA to date, and future research with diverse samples can further explore generalizability to the broader bPWA population. Next, given the limited precedence of comparable thresholds in monolingual or bilingual aphasia treatment literature, our study provides a pioneering quantification of response to bilingual treatment and CLG. The novelty of applying ML to bilingual aphasia treatment response means standardized benchmarks are currently non-existent. This scarcity of comparable studies and metrics underscores the need for future research to establish benchmarks, advancing treatment outcome predictions for bPWA.

Finally, it is important to note that our study focused only on trained items and their direct translations, not addressing within-language generalization. Although within-language generalization is a well-studied aspect of aphasia rehabilitation in monolingual populations,[11] we chose to focus on CLG due to the lack of sufficient statistical power in a previous bilingual meta-analytic study to identify its predictors,[61] and because, specific to bilingual aphasia, understanding factors that promote benefits across both languages is crucial when treatment is often delivered in only one language. This focus on CLG and baseline behavioral predictors may limit the generalizability of our findings to other aspects of language recovery in bPWA. Future research examining within-language generalization is needed, which may reveal shared and distinct mechanisms across different types of generalization.

In conclusion, our study demonstrates the promising application of ML models in predicting TLI and CLG in Spanish-English bPWA, with top-performing models achieving strong predictive performance. The alignment between the most influential predictive features and established clinical evidence underscores the validity and interpretability of our findings. Importantly, our study identified several novel insights, including the prominent role of aphasia severity in predicting treatment outcomes, challenging recent meta-analytic findings, and the influence of cognitive ability on both TLI and CLG outcomes. Despite limitations, our results provide a foundation for future research and contribute to the development of more effective and personalized rehabilitation strategies for this underserved population.

ARTICLE INFORMATION

Received May 22, 2024; final revision received October 17, 2024; accepted November 26, 2024.

Affiliations

Center for Brain Recovery, Boston University, MA (M.J.M., E.C., M.S., M.R.-M., S.K.). Department of Cognition, Development and Educational Psychology, Faculty of Psychology (C.P.) and Institute of Neurosciences (C.P.), University of Barcelona, Spain (C.P.). Cognition and Brain Plasticity Unit, Bellvitge Biomedical Research Institute-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain (C.P.). Department of Computer Sciences, University of Texas at Austin (U.G., R.M.).

Acknowledgments

The authors wish to thank the participants and their families for participating in the study. The authors also thank Xinyi Hu for her helpful comments on the manuscript.

Sources of Funding

This research was supported by U01 DC014922. M.J. Marte was partially supported by T32 DC013017. M. Scimeca is supported by 1F31DC021628-01A1. E. Carpenter is supported by the Dudley Allen Sargent Research Award, 1F31DC021385-01A1, and by a gift from the American Speech-Language-Hearing Foundation (New Century Scholars Doctoral Scholarship). Dr Peñaloza is supported by grant RYC2021-034561-I funded by MCIN/AEI/10.13039/501100011033 and the European Union NextGeneration/Recovery, Transformation, and Resilience Plan (PRTR).

Disclosures

Dr. Kiran serves as a cofounder and advisor to Constant Therapy Health. There is no scientific overlap between this work and the entity. The other authors report no conflicts.

Supplemental Material

Supplemental Methods
Supplemental Discussion
Figures S1–S2
Tables S1–S16
References 62–77
STROBE Checklist
TRIPOD+AI Checklist

REFERENCES

1. Rodriguez CJ, Allison M, Daviglus ML, Isasi CR, Keller C, Leira EC, Palaniappan L, Piña IL, Ramirez SM, Rodriguez B, et al; American Heart Association Council on Epidemiology and Prevention. Status of cardiovascular disease and stroke in Hispanics/Latinos in the United States: a science advisory from the American Heart Association. *Circulation*. 2014;130:593–625. doi: 10.1161/CIR.0000000000000071
2. Simmonds KP, Luo Z, Reeves M. Race/ethnic and stroke subtype differences in poststroke functional recovery after acute rehabilitation. *Arch Phys Med Rehabil*. 2021;102:1473–1481. doi: 10.1016/j.apmr.2021.01.090
3. Peñaloza C, Kiran S. Recovery and rehabilitation patterns in bilingual and multilingual aphasia. In: Schwieter JW, Paradis M, eds. *The Handbook of the Neuroscience of Multilingualism*. John Wiley & Sons, Ltd; 2019: 553–571
4. Kohnert K. Cognitive and cognate-based treatments for bilingual aphasia: a case study. *Brain Lang*. 2004;91:294–302. doi: 10.1016/j.bandl.2004.04.001
5. Kurland J, Falcon M. Effects of cognate status and language of therapy during intensive semantic naming treatment in a case of severe nonfluent bilingual aphasia. *Clin Linguist Phon*. 2011;25:584–600. doi: 10.3109/02699206.2011.565398
6. Peñaloza C, Scimeca M, Gaona A, Carpenter E, Mukadam N, Gray T, Shamapant S, Kiran S. Telerehabilitation for word retrieval deficits in bilinguals with aphasia: effectiveness and reliability as compared to in-person language therapy. *Front Neurol*. 2021;12:589330. doi: 10.3389/fneur.2021.589330
7. Scimeca M, Peñaloza C, Kiran S. Multilevel factors predict treatment response following semantic feature-based intervention in bilingual aphasia. *Biling (Camb Engl)*. 2023;27:246–262. doi: 10.1017/s1366728923000391
8. Croft S, Marshall J, Pring T, Hardwick M. Therapy for naming difficulties in bilingual aphasia: which language benefits? *Int J Lang Commun Disord*. 2011;46:48–62. doi: 10.3109/13682822.2010.484845
9. Kiran S, Roberts PM. Semantic feature analysis treatment in Spanish–English and French–English bilingual aphasia. *Aphasiology*. 2010;24:231–261. doi: 10.1080/02687030902958365
10. Kiran S, Sandberg C, Gray T, Ascenso E, Kester E. Rehabilitation in bilingual aphasia: evidence for within- and between-language generalization. *Am J Speech Lang Pathol*. 2013;22:S298–S309. doi: 10.1044/1058-0360(2013/12-0085)
11. Quique YM, Evans WS, Dickey MW. Acquisition and generalization responses in aphasia naming treatment: a meta-analysis of semantic feature analysis outcomes. *Am J Speech Lang Pathol*. 2019;28:230–246. doi: 10.1044/2018_AJSLP-17-0155
12. Nardo D, Holland R, Leff AP, Price CJ, Crinion JT. Less is more: neural mechanisms underlying anomia treatment in chronic aphasic patients. *Brain*. 2017;140:3039–3054. doi: 10.1093/brain/awx234
13. Seniów J, Litwin M, Leśniak M. The relationship between non-linguistic cognitive deficits and language recovery in patients with aphasia. *J Neurol Sci*. 2009;283:91–94. doi: 10.1016/j.jns.2009.02.315
14. Persad C, Wozniak L, Kostopoulos E. Retrospective analysis of outcomes from two intensive comprehensive aphasia programs. *Top Stroke Rehabil*. 2013;20:388–397. doi: 10.1310/tsr2005-388
15. Nakagawa Y, Sano Y, Funayama M, Kato M. Prognostic factors for long-term improvement from stroke-related aphasia with adequate linguistic rehabilitation. *Neurol Sci*. 2019;40:2141–2146. doi: 10.1007/s10072-019-03956-7
16. Lazar RM, Antoniello D. Variability in recovery from aphasia. *Curr Neurol Neurosci Rep*. 2008;8:497–502. doi: 10.1007/s11910-008-0079-x
17. Hillis AE, Tippett DC. Stroke recovery: surprising influences and residual consequences. *Adv Med*. 2014;2014:1–10. doi: 10.1155/2014/378263
18. Ramsey LE, Siegel JS, Lang CE, Strube M, Shulman GL, Corbetta M. Behavioural clusters and predictors of performance during recovery from stroke. *Nat Hum Behav*. 2017;1:1–10. doi: 10.1038/s41562-016-0038
19. O'Halloran R, Renton J, Harvey S, McSween MP, Wallace SJ. Do social determinants influence post-stroke aphasia outcomes? A scoping review. *Disabil Rehabil*. 2023;46:1274. doi: 10.1080/09638288.2023.2193760
20. Doogan C, Dignam J, Copland D, Leff A. Aphasia recovery: when, how and who to treat? *Curr Neurol Neurosci Rep*. 2018;18:90. doi: 10.1007/s11910-018-0891-x
21. Efstratiadou EA, Papathanasiou I, Holland R, Archonti A, Hilari K. A systematic review of semantic feature analysis therapy studies for aphasia. *J Speech Lang Hear Res*. 2018;61:1261–1278. doi: 10.1044/2018_JSLHR-L-16-0330
22. Lambon Ralph MA, Snell C, Fillingham JK, Conroy P, Sage K. Predicting the outcome of anomia therapy for people with aphasia post CVA: both language and cognitive status are key predictors. *Neuropsychol Rehabil*. 2010;20:289–305. doi: 10.1080/09602010903237875
23. Dignam J, Copland D, Brien Kate O, Burfein P, Khan A, Rodriguez AD. Influence of cognitive ability on therapy outcomes for anomia in adults with chronic poststroke aphasia. *J Speech Lang Hear Res*. 2017;60:406–421. doi: 10.1044/2016_JSLHR-L-15-0384
24. Kohnert K. Cross-language generalization following treatment in bilingual speakers with aphasia: a review. *Semin Speech Lang*. 2009;30:174–186. doi: 10.1055/s-0029-1225954
25. Faroqi-Shah Y, Frymark T, Mullen R, Wang B. Effect of treatment for bilingual individuals with aphasia: a systematic review of the evidence. *J Neurolinguistics*. 2010;23:319–341. doi: 10.1016/j.jneuroling.2010.01.002
26. Kiran S, Gray T. Diagnosis, assessment and rehabilitation: chapter 17. Understanding the nature of bilingual aphasia. In: Miller D, Bayram F, Rothman J, Serratrice L, eds. *Bilingual Cognition and Language: The State of the Science Across its Subfields*. John Benjamins Publishing Company; 2018:371–400
27. Goral M, Lerman A. Variables and mechanisms affecting response to language treatment in multilingual people with aphasia. *Behav Sci (Basel)*. 2020;10:144. doi: 10.3390/bs10090144
28. Goral M, Norvik MI, Antfolk J, Agrotou I, Lehtonen M. Cross-language generalization of language treatment in multilingual people with post-stroke aphasia: a meta-analysis. *Brain Lang*. 2023;246:105326. doi: 10.1016/j.bandl.2023.105326
29. Grasemann U, Peñaloza C, Dekhtyar M, Miikkulainen R, Kiran S. Predicting language treatment response in bilingual aphasia using neural network-based patient models. *Sci Rep*. 2021;11:10497. doi: 10.1038/s41598-021-89443-6
30. Billot A, Lai S, Varkanitsa M, Braun EJ, Rapp B, Parrish TB, Higgins J, Kurani AS, Caplan D, Thompson CK, et al. Multimodal neural and behavioral data predict response to rehabilitation in chronic poststroke aphasia. *Stroke*. 2022;53:1606–1614. doi: 10.1161/STROKEAHA.121.036749
31. Kristinsson S, Zhang W, Rorden C, Newman-Norlund R, Basilakos A, Bonilha L, Yourganov G, Xiao F, Hillis A, Fridriksson J. Machine learning-based multimodal prediction of language outcomes in chronic aphasia. *Hum Brain Mapp*. 2021;42:1682–1698. doi: 10.1002/hbm.25321
32. Lundberg S, Lee SI. A unified approach to interpreting model predictions. *arXiv*. Preprint posted online May 22, 2017. doi: 10.48550/arXiv.1705.07874
33. Peñaloza C, Dekhtyar M, Scimeca M, Carpenter E, Mukadam N, Kiran S. Predicting treatment outcomes for bilinguals with aphasia using computational modeling: study protocol for the PROCoM randomised controlled trial. *BMJ Open*. 2020;10:e040495. doi: 10.1136/bmjopen-2020-040495
34. Kastenbaum JG, Bedore LM, Peña ED, Sheng L, Mavis I, Sebastian-Vaytadden R, Rangamani G, Vallila-Rohter S, Kiran S. The influence of proficiency and language combination on bilingual lexical access. *Biling (Camb Engl)*. 2019;22:300–330. doi: 10.1017/S1366728918000366
35. Marte MJ, Carpenter E, Falconer IB, Scimeca M, Abdollahi F, Peñaloza C, Kiran S. LEX-BADAT: language experience in bilinguals with and without aphasia dataset. *Front Psychol*. 2022;13:875928. doi: 10.3389/fpsyg.2022.875928
36. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Soft*. 2011;45:1–67. doi: 10.18637/jss.v045.i03
37. R Core Team. R: the R project for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Accessed March 13, 2024. https://www.r-project.org/
38. Makowski D. The psycho package: an efficient and publishing-oriented workflow for psychological science. *J Open Source Softw*. 2018;3:470. doi: 10.21105/joss.00470
39. Helm-Estabrooks N. Cognitive Linguistic Quick Test. In:Kreutzer J, DeLuca J, Caplan B, eds. *Encyclopedia of Clinical Neuropsychology*. Cham: Springer International Publishing; 2018:1–4.
40. Raven J. Raven Progressive Matrices. In: McCallum RS, ed. *Handbook of Nonverbal Assessment*. Boston, MA: Springer US. 2003:223–237.
41. Howard D. The Pyramids and Palm Trees Test: A Test of Semantic Access from Wordsand Pictures. Thames Valley Test Company; 1992. Accessed July 10, 2023. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Cognition-%26-Neuro/The-Pyramids-and-Palm-Trees-Test/p/100000185.html
42. Kaplan E, Goodglass H, Weintraub S. *Boston Naming Test*. Pro-Ed; 2001.
43. Paradis M, Libben G. *The Assessment of Bilingual Aphasia*. Psychology Press; 2014:1–258.
44. Kertesz A. *Western Aphasia Battery—Revised*. Pearson Assessment; 2012.
45. 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. *J Mach Learn Res*. 2011;12:2825–2830. doi: 10.48550/arXiv.1201.0490

46. Van Rossum G, Drake FL. *Python 3 Reference Manual*. CreateSpace. 2009:1–242.
47. McKinney W. *Data Structures for Statistical Computing in Python*. In: van der Walt S, Millman J, eds.Proceedings of the 9th Python in Science Conference. 2010:56–61.
48. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. Array programming with NumPy. *Nature*. 2020;585:357–362. doi: 10.1038/s41586-020-2649-2
49. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9:90–95. doi: 10.1109/mcse.2007.55
50. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079–2107. doi: 10.1093/humrep/deu295
51. LaValle SM, Branicky MS, Lindemann SR. On the relationship between classical grid search and probabilistic roadmaps. *Int J Robot Res*. 2004;23:673–692. doi: 10.1177/0278364904045481
52. Lerman A, Goral M, Obler LK. The complex relationship between pre-stroke and post-stroke language abilities in multilingual individuals with aphasia. *Aphasiology*. 2020;34:1319–1340. doi: 10.1080/02687038.2019.1673303
53. Kroll JF, Hell JGV, Tokowicz N, Green DW. The revised hierarchical model: a critical review and assessment. *Biling (Camb Engl)*. 2010;13:373–381. doi: 10.1017/S136672891000009X
54. Dijkstra T, Wahl A, Buytenhuijs F, Halem NV, Al-Jibouri Z, Korte MD, Rekké S. Multilink: a computational model for bilingual word recognition and word translation. *Biling (Camb Engl)*. 2019;22:657–679. doi: 10.1017/S1366728918000287
55. Marebwa BK, Fridriksson J, Yourganov G, Feenaughty L, Rorden C, Bonilha L. Chronic post-stroke aphasia severity is determined by fragmentation of residual white matter networks. *Sci Rep*. 2017;7:8188. doi: 10.1038/s41598-017-07607-9
56. Mooijman S, Schoonen R, Roelofs A, Ruiter MB. Executive control in bilingual aphasia: a systematic review. *Biling (Camb Engl)*. 2022;25:13–28. doi: 10.1017/s136672892100047x
57. Rosenich E, Hordacre B, Paquet C, Koblar SA, Hillier SL. Cognitive reserve as an emerging concept in stroke recovery. *Neurorehabil Neural Repair*. 2020;34:187–199. doi: 10.1177/1545968320907071
58. Calabria M, Costa A, Green DW, Abutalebi J. Neural basis of bilingual language control: bilingual language control. *Ann NY Acad Sci*. 2018;1426:221–235. doi: 10.1111/nyas.13879
59. Abutalebi J, Green D. Bilingual language production: the neurocognition of language representation and control. *J Neurolinguistics*. 2007;20:242–275. doi: 10.1016/j.jneuroling.2006.10.003
60. Radman N, Mouthon M, Di Pietro M, Gaytanidis C, Leemann B, Abutalebi J, Annoni JM. The role of the cognitive control system in recovery from bilingual aphasia: a multiple single-case fMRI study. *Neural Plast*. 2016;2016:1–22. doi: 10.1155/2016/8797086
61. Lee S, Faroqi-Shah Y. A Meta-analysis of anomia treatment in bilingual aphasia: within- and cross-language generalization and predictors of the treatment outcomes. *J Speech Lang Hear Res*. 2024;67:1558–1600. doi: 10.1044/2024_JSLHR-23-00026
62. Iosa M, Morone G, Antonucci G, Paolucci S. Prognostic factors in neurorehabilitation of stroke: a comparison among regression, neural network, and cluster analyses. *Brain Sciences*. 2021;11:1147. doi: 10.3390/brainsci11091147
63. Day M, Dey RK, Baucum M, Paek EJ, Park H, Khojandi A. Predicting severity in people with aphasia: a natural language processing and machine learning approach. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021: 2299–2302
64. Basilakos A, Yourganov G, den ODB, Fogerty D, Rorden C, Feenaughty L, Fridriksson J. A multivariate analytic approach to the differential diagnosis of apraxia of speech. *J Speech Lang Hear Res*. 2017;60:3378–3392. doi: 10.1044/2017_JSLHR-S-16-0443
65. Tafuri B, De Blasi R, Nigro S, Logroscino G. Explainable machine learning radiomics model for primary progressive aphasia classification. *Front Syst Neurosci*. 2024;18:1324437. doi: 10.3389/fnsys.2024.1324437
66. Breier JI, Juranek J, Maher LM, Schmadeke S, Men D, Papanicolaou AC. Behavioral and neurophysiologic response to therapy for chronic aphasia. *Arch Phys Med Rehabil*. 2009;90:2026–2033. doi: 10.1016/j.apmr.2009.08.144
67. Kristinsson S, Basilakos A, Elm J, Spell LA, Bonilha L, Rorden C, den Ouden DB, Cassarly C, Sen S, Hillis A, et al. Individualized response to semantic versus phonological aphasia therapies in stroke. *Brain Commun*. 2021;3:fcab174. doi: 10.1093/braincomms/fcab174
68. Latimer NR, Dixon S, Palmer R. Cost-utility of self-managed computer therapy for people with aphasia. *Int J Technol Assess Health Care*. 2013;29:402–409. doi: 10.1017/S0266462313000421
69. Kuhn M, Johnson K. *Applied Predictive Modeling*. Springer; 2013.
70. Aban I, Cutter G. Defining responders and non-responders. In: Filippi M, Rovaris M, Comi G, eds. *Neurodegeneration in Multiple Sclerosis*. Springer Milan; 2007:113–125.
71. Laska E, Siegel C, Lin Z. A likely responder approach for the analysis of randomized controlled trials. *Contemp Clin Trials*. 2022;114:106688. doi: 10.1016/j.cct.2022.106688
72. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychol Methods*. 2002;7:19–40. doi: 10.1037/1082-989x.7.1.19
73. Cordella C, Marte MJ, Liu H, Kiran S. An introduction to machine learning for speech-language pathologists: concepts, terminology, and emerging applications. *Perspect ASHA Spec Interest Groups*. doi: 10.1044/2024_PERSP-24-00037
74. Erickson BJ, Kitamura F. Magician's corner: 9. performance metrics for machine learning models. *Radiol Artif Intell*. 2021;3:e200126. doi: 10.1148/ryai.2021200126
75. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21:6. doi: 10.1186/s12864-019-6413-7
76. Ben-Hur A, Weston J. A user's guide to support vector machines. *Methods Mol Biol*. 2010;609:223–239. doi: 10.1007/978-1-60327-241-4_13
77. Stasenko A, Kleinman D, Gollan TH. Older bilinguals reverse language dominance less than younger bilinguals: evidence for the inhibitory deficit hypothesis. *Psychol Aging*. 2021;36:806–821. doi: 10.1037/pag0000618