JoS JOURNAL OF STROKE

Journal of Stroke 2022;24(1):108-117
https://doi.org/10.5853/jos.2021.02061

Original Article

Deep Learning Approach Using Diffusion-Weighted Imaging to Estimate the Severity of Aphasia in Stroke Patients

Soo Jeong,[a,*] Eun-Jae Lee,[a,*] Yong-Hwan Kim,[b] Jin Cheol Woo,[c] On-Wha Ryu,[a] Miseon Kwon,[a] Sun U Kwon,[a] Jong S. Kim,[a] Dong-Wha Kang[a]

[a]Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[b]Nunaps Inc., Seoul, Korea
[c]Asan Institute for Life Sciences, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Background and Purpose This study aimed to investigate the applicability of deep learning (DL) model using diffusion-weighted imaging (DWI) data to predict the severity of aphasia at an early stage in acute stroke patients.

Methods We retrospectively analyzed consecutive patients with aphasia caused by acute ischemic stroke in the left middle cerebral artery territory, who visited Asan Medical Center between 2011 and 2013. To implement the DL model to predict the severity of post-stroke aphasia, we designed a deep feed-forward network and utilized the lesion occupying ratio from DWI data and established clinical variables to estimate the aphasia quotient (AQ) score (range, 0 to 100) of the Korean version of the Western Aphasia Battery. To evaluate the performance of the DL model, we analyzed Cohen's weighted kappa with linear weights for the categorized AQ score (0–25, very severe; 26–50, severe; 51–75, moderate; ≥76, mild) and Pearson's correlation coefficient for continuous values.

Results We identified 225 post-stroke aphasia patients, of whom 176 were included and analyzed. For the categorized AQ score, Cohen's weighted kappa coefficient was 0.59 (95% confidence interval [CI], 0.42 to 0.76; $P<0.001$). For continuous AQ score, the correlation coefficient between true AQ scores and model-estimated values was 0.72 (95% CI, 0.55 to 0.83; $P<0.001$).

Conclusions DL approaches using DWI data may be feasible and useful for estimating the severity of aphasia in the early stage of stroke.

Keywords Stroke; Aphasia; Magnetic resonance imaging; Deep learning

Correspondence: Dong-Wha Kang
Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: +82-2-3010-3440
Fax: +82-2-474-4691
E-mail: dwkang@amc.seoul.kr
https://orcid.org/0000-0002-2999-485X

Received: June 8, 2021
Revised: September 16, 2021
Accepted: October 5, 2021

*These authors contributed equally to the manuscript as first author.

Introduction

Post-stroke aphasia is a major stroke sequela and burden which significantly and negatively affects the quality of life of patients.[1-4] Identifying the severity of aphasia in the early stage of stroke is crucial in predicting the prognosis of patients.[5,6] However, detailed examination of post-stroke aphasia during the acute phase is often difficult because of various stroke-related conditions (e.g., altered mental status, poor cooperation, confusion, and unstable medical conditions). Stroke imaging may help evaluate the severity and prognosis of aphasia; it has been suggested to be useful in estimating roughly categorized aphasia outcomes (e.g., good vs. bad outcome) after stroke;[6-9] however, detailed aphasia characteristics (e.g., continuous

Copyright © 2022 Korean Stroke Society
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

score value) have yet to be estimated from stroke imaging analysis.

Deep learning (DL) techniques, which are applications of artificial intelligence, have recently emerged and are now rigorously applied in the medical field, especially in outcome prediction studies using imaging data.[10,11] Because DL techniques may use multiple features that are invisible to humans,[12,13] we hypothesized that DL methods may be helpful in estimating the severity of aphasia at an acute stage after stroke using magnetic resonance imaging (MRI) data. Consequently, we developed a DL model using diffusion-weighted imaging (DWI) to estimate the aphasia quotient (AQ) score of the Korean version of the Western Aphasia Battery (K-WAB) (range, 0 to 100),[14] which reflects the severity of aphasia in stroke patients with acute stroke, and evaluated its performance by comparing it with the true values of each patient. In the meantime, we also developed logistic regression models that estimate the AQ score to evaluate the performance of the DL method, as compared to that of the conventional machine-learning approach.

Methods

Data availability

Anonymized data are available on reasonable request from any qualified investigator.

Study design and patient selection

We retrospectively analyzed consecutive acute stroke patients with aphasia who visited the Asan Medical Center (Seoul, South Korea) within 7 days of symptom onset between January 2011 and December 2013. Aphasia was defined as the presence of a score greater than 1 in the best language category of the National Institutes of Health Stroke Scale.[15] In Asan Medical Center, during the study period, we usually performed MRI, including DWI and fluid-attenuated inversion recovery sequences, in stroke patients. If patients were considered candidates for acute revascularization, brain computed tomography was the *a priori* imaging modality in Asan Medical Center to exclude the presence of hemorrhage; MRI was subsequently performed in the emergency room immediately after the initiation of alteplase injection. The choice to perform intra-arterial thrombolysis/thrombectomy was determined according to a comprehensive analysis considering MRI lesion characteristics, vessel status, and patients' neurological status. Therefore, DWI was performed after the initiation of alteplase before mechanical thrombectomy. After neurological stabilization at the discretion of the attending physicians, the K-WAB was routinely performed in patients with aphasia, unless patients were vitally unstable or uncooperative.

We included patients who had ischemic lesions in the left middle cerebral artery (MCA) territory and those who underwent K-WAB within 14 days of symptom onset. Patients with multiple lesions in multiple vascular territories were also included if they had aphasia and stroke lesions in the left MCA territory. Only patients who used Korean as their first language were included. We excluded patients who did not undergo DWI or who demonstrated old stroke lesions within the left MCA territory on fluid-attenuated inversion recovery or gradient echo sequences of MRI.[16] Patients who had previously been diagnosed with dementia that caused communication difficulties before the index stroke were also not included in the study. In addition, left-handed patients were excluded to ensure study consistency.

DWI data processing

The patients underwent 1.5-T MRI (Magnetom Avanto, Siemens Healthineers, Erlangen, Germany). The ischemic lesion mask was extracted in the native DWI space using the FSLView toolbox in the functional MRI of the brain software library (FSL, developed by Oxford Center for Functional MRI of the Brain, Oxford, UK). Initial DWI was used for the analysis. A stroke neurologist, who was blinded to all clinical information, manually segmented the DWI high-signal intensity area to measure the ischemic lesion. In patients with lesions in multiple vascular territories, lesions outside the left MCA territory were also included in the lesion volume analysis. When we obtained image features, only lesions in the left hemisphere were targeted and analyzed. Affine transformation and nonlinear warping coefficients were estimated between the DWI using a b-value of zero in the native space and standard Montreal Neurological Institute 152 T2 template images via the Functional Magnetic Resonance Imaging of the Brain's linear or nonlinear image registration tool. The estimated parameters were applied to the delineated lesion masks.

To cover the whole brain area and to consider white matter regions, we used five brain ATLAS templates (Broadmann's Area [BA], Automated Anatomical Labeling [AAL], Harvard-Oxford [HO], JHU white matter label [WM-Label], and JHU white matter track [WM-Track]) for analyzing the brain imaging information. The lesion occupying ratio in each individual cortex was calculated for five human brain ATLASs (BA, 41 left hemisphere regions among 82 region labels; AAL, 54 among 116; HO, 55 among 110; WM-Label, 19 among 44; and WM-Track, nine out of 20) in the standard Montreal Neurological Institute space to quantitatively calculate the regional damage as follows: (volume of lesion inside the individual gyrus/total volume of indi-

vidual gyrus)×100 (%).[17]

Data collection

To develop the DL model, we obtained 184 features, including 178 lesion occupying ratio features derived from manually segmented infarct lesions on DWI, and six clinical features (age, sex, interval between the K-WAB test and onset, interval between DWI and onset, education, and lesion volume), which are regarded as important prognostic factors for post-stroke aphasia.[18] Clinical features were obtained from electronic medical records. As an outcome parameter, the AQ score, a composite score of the K-WAB, was adopted and collected. In Asan Medical Center, the K-WAB was performed by two dedicated speech therapists with experience in speech evaluation and therapy for more than 5 years. The AQ score ranges from 0 to 100, with a higher score representing better language performance.

Deep learning model

We utilized a deep feed-forward network (DFFN) for the DL model and did not use convolutional neural network approaches. We chose 178 lesion occupying ratio features associated with the left hemisphere regions among various atlases on DWI and clinical features as input parameters. DFFN was implemented using PyTorch (developed by Facebook's AI Research lab). DFFN consists of four fully connected layers: an input layer (184 nodes), 1st hidden layer (90 nodes), 2nd hidden layer (30 nodes), and an output layer. The hard hyper-tangent activation function was adopted for the 1st and 2nd hidden layers with minimum/maximum values of −1 to 1. Finally, the output layer is fully connected and concatenates the outputs of the 1st and 2nd layers, resulting in a final score with a hyper-tangent activation function having a range of 0 to 1. During training of the model, mean-squared error as a loss function and Adam optimizer with a learning rate of 0.001 with default parameters (beta=0.9 and 0.999; eps=1e-08; weight decay=0; AMSgrad=false) were adopted with 50% of dropout nodes on the 1st and 2nd hidden layers. A mini-batch of 50 samples for an epoch was used for the training, and the training was stopped when the loss from the validation data was no longer minimized compared to the best loss over 50 epochs. The maximum number of training epochs was limited to 500.

For feature selection, we adopted a dropout technique, a regularization method to overcome redundancies in image features, and overfitting.[19] A schematic diagram of the DFFN structure is shown in Supplementary Figure 1. To develop the DL model, we allocated patients admitted before January 2013 to the training set, and the remaining patients were assigned to the test set. The ratio of patients in the training and test sets was 3:1. The DL model was established with randomly selected 70% of the training set at 1,000 times (i.e., 1,000 DL models were generated by the bagging strategy). For the test set, the predicted scores from 1,000 models were averaged.

Logistic regression model

We developed a conventional machine-learning model using logistic regression with Lasso (least absolute shrinkage and selection operator) regularization.[20] Lasso regression is a model that shrinks regression coefficients towards zero, thereby effectively reducing feature redundancies in the imaging features, driven by using multiple brain ATLAS templates. We used a 5-fold cross-validation to yield the optimal regularization parameter. To examine whether the logistic regression model can process a large number of image features, as well as a small number of clinical features, to improve its performance with increasing datasets, we developed multiple regression models by inputting different sets of features: (1) clinical features only, (2) imaging features only, and (3) clinical+imaging features.

Statistical analysis

Baseline characteristics between the training and test sets were compared using the chi-square test or Fisher's exact test for categorical variables and the t-test or Mann-Whitney U test for continuous variables, as appropriate. To compare the voxel-wise frequency difference of lesions between the training and test groups, the Bernoulli model-based two-sample t-test was performed for each voxel.[21] We calculated Cohen's weighted kappa with linear weights[22] between the categorized model-estimated scores and true AQ scores (0–25, very severe; 26–50, severe; 51–75, moderate; ≥76, mild).[23] The degree of agreement using Cohen's weighted kappa were as follows: poor (<0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), and very good (0.81–1.00). In addition, we estimated the Pearson's correlation coefficients between the model-estimated and true values of the continuous AQ scores. Interpretations of the correlation coefficient (r) were as follows: very weak ($0.00<|r|<0.20$), weak ($0.20\leq|r|<0.40$), moderate ($0.40\leq|r|<0.60$), strong ($0.60\leq|r|<0.80$), and very strong ($0.80\leq|r|<1.00$). We further evaluated the correlation coefficients between model-estimated and true values of sub-domain scores of the AQ score (spontaneous speech, comprehension, repetition, and naming).[14] Patients showing notable discrepancies between model-estimated and true AQ scores were defined as those showing studentized residuals larger than 2 (in absolute value).[24] A two-sided $P<0.05$ indicated statistical significance. All statistical analyses were performed using the

https://doi.org/10.5853/jos.2021.02061

SPSS statistical software version 25.0 (IBM Co., Armonk, NY, USA).

Standard protocol approvals, registrations, and patient consents

The study was performed in accordance with the Good Clinical Practice guidelines and the Declaration of Helsinki, and was approved by the Institutional Review Board of Asan Medical Center (IRB No. 2020-1794). Written informed consent was waived owing to the retrospective nature of the study.

Results

Participants

During the study period, a total of 225 acute stroke patients with aphasia visited Asan Medical Center and were included in the study. The K-WAB test was performed for all patients. Of these, 49 were excluded (no available DWI data [n=16], crossed aphasia with right MCA lesions [n=7], old stroke lesion in the left MCA territory [n=13], left-handedness [n=10], and delayed K-WAB test [>14 days after symptom onset, n=3]) (Figure 1). The included patients had a mean age of 66.5±11.8 years, mean time interval from onset to MRI of 2.4±2.7 days, and mean time interval from onset to K-WAB of 3.8±2.7 days. A total of 36 patients received alteplase or intra-arterial thrombolysis/thrombectomy treatment. Of the 176 patients who were finally included, 127 were assigned to the training set and 49 were assigned to the test set. There were no significant differences in the baseline characteristics between the training and test sets (Table 1). The spatial patterns of lesions were analyzed by comparing the lesion proportion in every single voxel between the training and test groups using the Bernoulli model-based two-sample t-test. There were no significant differences between the groups (Figure 2).

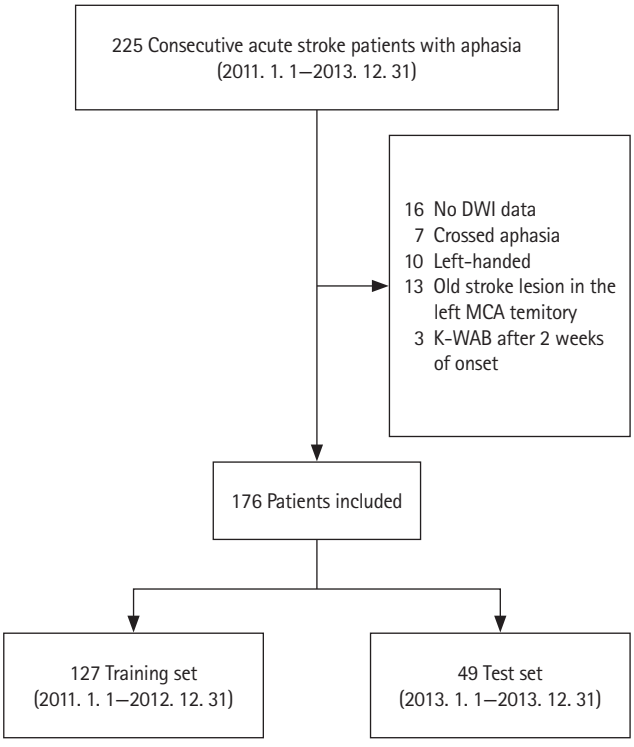


Figure 1. Flowchart showing patient selection. DWI, diffusion-weighted image; MCA, middle cerebral artery; K-WAB, Korean version of the Western Aphasia Battery.

Performance of the machine-learning model

For categorized AQ score, the DL model showed an accuracy of 61% in total, with Cohen's weighted kappa of 0.59 (95% confidence interval [CI], 0.42 to 0.76; $P<0.001$) (Table 2). In addition, the DL model showed an accuracy of 50% or more in all severity categories. The largest figure (73%) was observed in the very severe aphasia group.

For continuous AQ score, the correlation coefficient between the true AQ score and model-estimated AQ score was 0.72 (95% CI, 0.55 to 0.83; $P<0.001$) (Figure 3). For sub-domain scores of the AQ score, the DL model showed competent performance with strong correlation coefficient values in all categories as follows: spontaneous speech (r=0.75; 95% CI, 0.59 to 0.85; $P<0.001$); comprehension (r=0.71; 95% CI, 0.54 to 0.83; $P<0.001$); repetition (r=0.65; 95% CI, 0.44 to 0.78; $P<0.001$); and naming (r=0.71; 95% CI, 0.54 to 0.83; $P<0.001$) (Supplementary Figure 2).

We performed an additional analysis using logistic regression to evaluate the performance of the DL method, as compared to that of the conventional machine-learning approach (Supplementary Tables 1 and 2). In this process, we developed multiple regression models to determine whether a conventional model could handle large data such as image features by including different sets of input features. Consequently, logistic regression with clinical variables (only six features) showed a greater performance (correlation coefficient for AQ scores, 0.70), as compared to the large feature situation (178 image features only, 0.54; 178 image features+6 clinical features together, 0.63) in the test set (Supplementary Table 2).

Cases with discrepancy and those with acute intervention

Three patients demonstrated notable discrepancies between the model-estimated and true AQ scores (Table 3 and Figure 3). Cases 1 and 2 showed more severe aphasia in their K-WAB tests compared to the results predicted by the DL model, while Case 3 demonstrated opposite results with milder aphasia in

Table 1. Baseline characteristics of patients in the training and test set

Characteristic	127 Training set	49 Test set	*P*
Age (yr)	65.9±11.8	68.0±12.0	NS
Female sex	46 (36.2)	14 (28.6)	NS
Hypertension	60 (47.2)	31 (63.3)	NS
Diabetes	40 (31.5)	14 (28.6)	NS
NIHSS on admission	8.3±5.5	8.4±5.5	NS
Years of education	10.1±5.1	10.1±4.7	NS
Lesion volume (cm^3)	58.4±75.7	38.9±49.9	NS
MRI from onset (hr)	56.1±66.7	60.4±59.2	NS
K-WAB from onset (day)	3.9±3.0	3.4±2.1	NS
AQ score	43.6±30.3	37.7±31.1	NS
Mild (≥76)	29 (22.8)	13 (26.5)	NS
Moderate (51–75)	28 (22.0)	6 (12.2)	NS
Severe (26–50)	27 (21.3)	8 (16.3)	NS
Very severe (0–25)	43 (33.9)	22 (44.9)	NS

Values are presented as mean±standard deviation or number (%).
NS, non-significant; NIHSS, National Institutes of Health Stroke Scale; MRI, magnetic resonance imaging; K-WAB, Korean version of the Western Aphasia Battery; AQ, aphasia quotient.

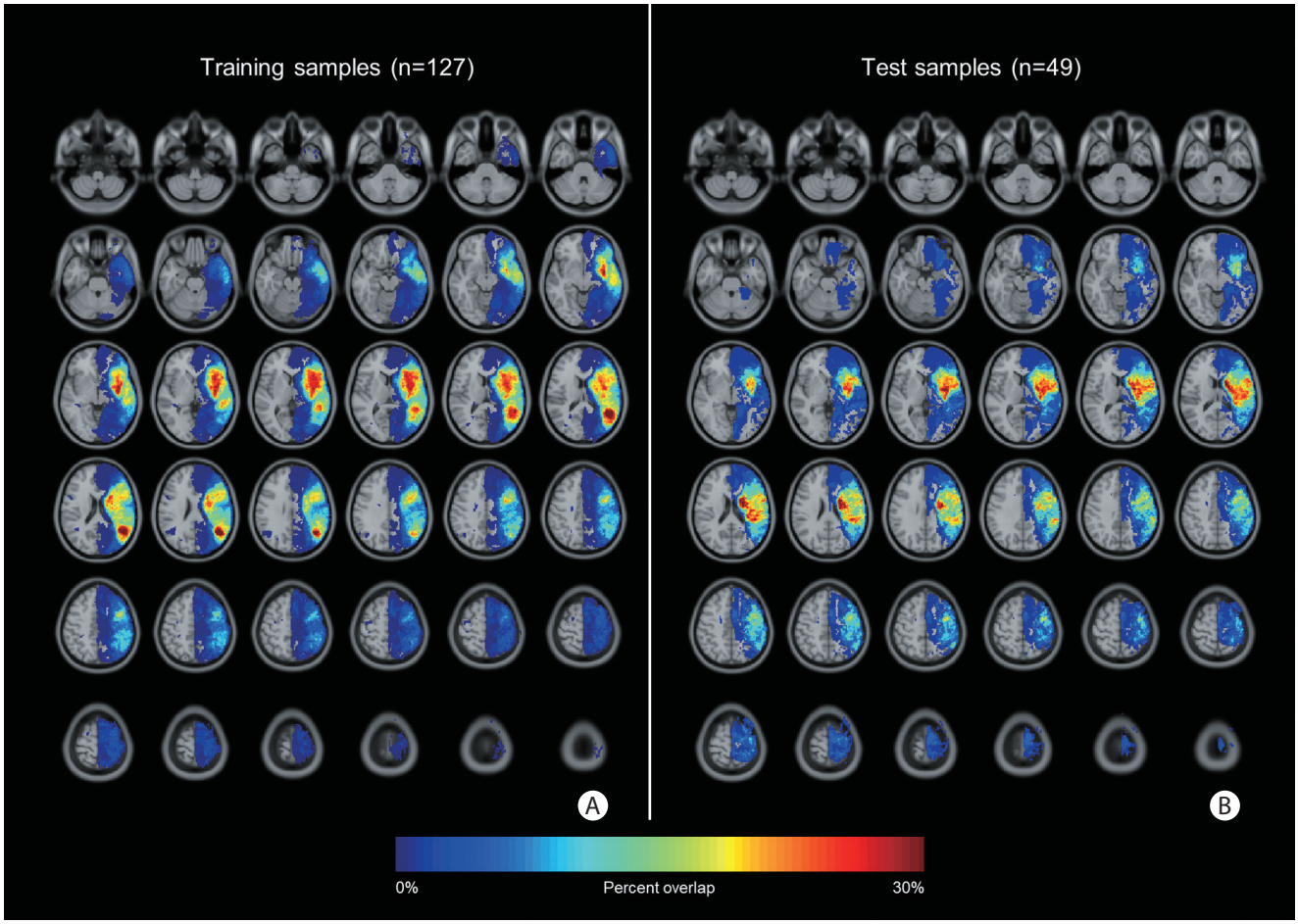


Figure 2. Lesion pattern heat maps of (A) training and (B) test groups. A heat map was used to visualize the proportion of lesions in each voxel. We compared the lesion proportion in every voxel between the training and test groups using the Bernoulli model-based two-sample t-test, but found no difference between the training and test groups (P>0.05).

https://doi.org/10.5853/jos.2021.02061

Table 2. Contingency table between the true score and model-estimated score

	True score				Total
	Very severe	Severe	Moderate	Mild	
Model-estimated score					
Very severe	16	3	2	0	21
Severe	4	4	1	2	11
Moderate	1	1	3	4	9
Mild	1	0	0	7	8
Total	22	8	6	13	49
Accuracy	16/22 (73)	4/8 (50)	3/6 (50)	7/13 (54)	30/49 (61)

Values are presented as number (%). Cohen's weighted kappa, κ=0.59 (95% confidence interval, 0.42 to 0.76; *P*<0.001).

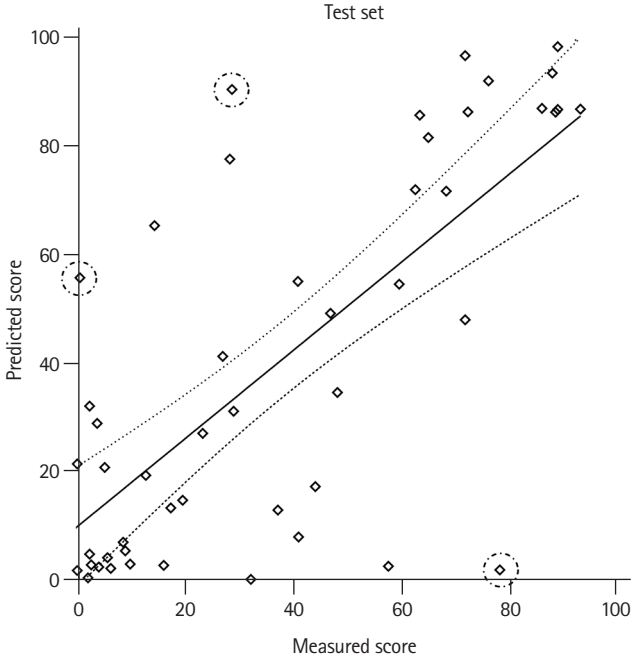


Figure 3. Correlation analysis between the true aphasia quotient (AQ) score and predicted AQ score in the test set. The correlation coefficient was 0.72 (95% confidence interval, 0.55 to 0.83; *P*<0.001); solid black line, regression line; dotted black line, 95% confidence limit; dotted circle, cases with notable discrepancy between the true AQ score and predicted AQ score with studentized residual larger than 2 (in absolute value).

the K-WAB test than in the DL model. These patients had distinctive clinical features. Case 1 had a small embolic ischemic lesion on acute DWI (3.6 hours after symptom onset) with underlying left proximal internal carotid artery stenosis, which decreased perfusion in the left MCA territory. Case 2 had multiple scattered lesions that involved the bilateral frontal lobes, including the anterior cingulate cortex, on DWI (83 hours after symptom onset). Finally, Case 3 involved a patient with in-hospital stroke after endovascular coiling of a left posterior communicating artery aneurysm. This patient showed large lesions involving most of the left MCA territory on hyperacute DWI (within 1 hour after symptom onset) and subsequently underwent successful acute intra-arterial thrombectomy (Figure 4).

We subsequently evaluated Cohen's weighted kappa value and correlation coefficients only in patients who had undergone acute thrombolysis/thrombectomy. Because the number of these patients was small (n=24 for the training group; n=12 for the test group), we combined two groups in this analysis to increase the number of patients (n=36). As a result, the values of the variables (κ=0.60; 95% CI, 0.41 to 0.78; *P*<0.001) (r=0.73; 95% CI, 0.53 to 0.85; *P*<0.001) were comparable to those in the total group of patients (Figure 3 and Supplementary Table 3). Notably, among the patients who underwent acute intervention, the performance of the DL model was good in patients with very severe (accuracy, 92%) or mild (accuracy, 70%) aphasia, while it was less remarkable in patients with moderate and severe aphasia.

Discussion

In this study, we demonstrated that DL techniques using DWI and clinical data could be used to estimate the severity of aphasia in ischemic stroke patients at an early stage. The DL model showed good performance in estimating the severity of post-stroke aphasia, as compared to the actual AQ score values, within 14 days after symptom onset.

Aphasia is one of the most devastating cognitive sequelae of stroke, which results in difficulties in activities of daily living in stroke patients.[1-4] There are many factors that are associated with post-stroke aphasia recovery, including age, sex, handedness, lesion location and size, stroke and aphasia severity, and even non-linguistic cognitive abilities such as emotional and social characteristics.[25] Among these factors, the severity of initial aphasia is one of the best predictors of aphasia outcome at a later stage.[26-28] However, it is often difficult to thoroughly

Table 3. Clinical features of cases with significant discrepancies between model-estimated score and true AQ score

	Case 1	Case 2	Case 3
Age (yr)	81	75	54
Sex	Male	Male	Female
Years of education	9	6	12
Model-estimated AQ score	90.5	55.7	1.8
True AQ score	28.6	0.5	78.1
NIHSS on admission	10	9	9
MRI from onset (hr)	3.6	83	0.8
K-WAB from onset (day)	3	9	5
Prime suspect for discrepancy	Low perfusion due to left proximal ICA stenosis	Abulia due to bilateral frontal lesions	Early revascularization

AQ, aphasia quotient; NIHSS, National Institutes of Health Stroke Scale; MRI, magnetic resonance imaging; K-WAB, Korean version of the Western Aphasia Battery; ICA, internal carotid artery.

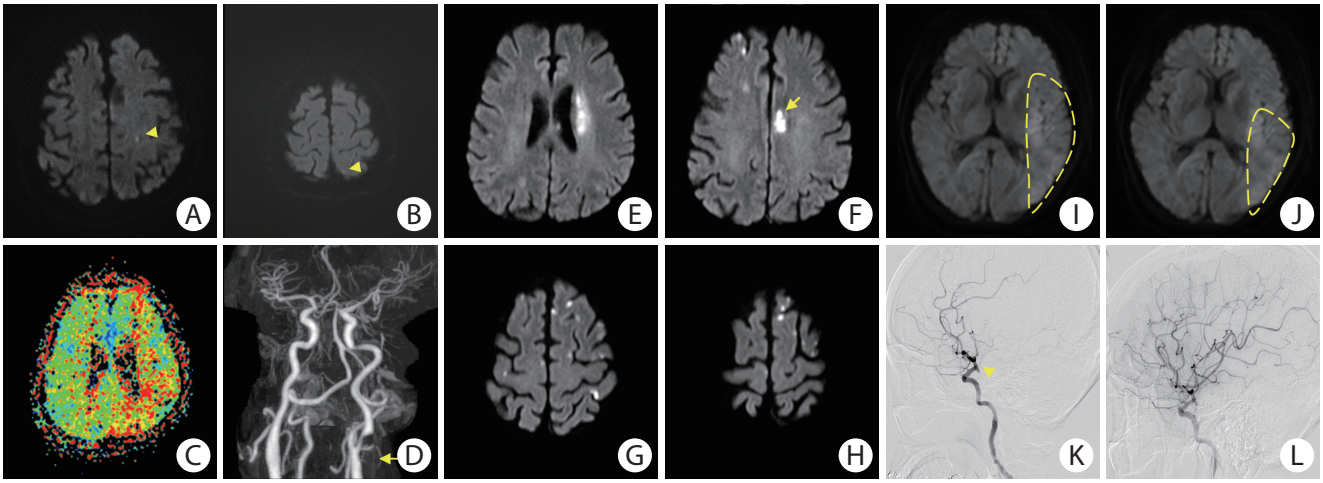


Figure 4. Imaging characteristics of cases with notable discrepancies. Panels represent (A-D) Case 1, (E-H) Case 2, and (I-L) Case 3, respectively. (A, B) Small acute lesions in the left internal carotid artery (ICA) border zone (yellow arrowheads) on diffusion-weighted imaging (DWI); (C) increased time-to-peak value in the left middle cerebral artery (MCA) territory; (D) severe stenosis in the left proximal ICA (yellow arrow); (E-H) acute infarction in the left corona radiata and bilateral anterior cerebral artery territory on DWI, especially in the left anterior cingulate cortex (yellow arrow); (I, J) large but subtle DWI high-signal intensity in the left MCA territory (yellow dashed lines); (K) occlusion of the left MCA inferior division (yellow arrow) on conventional angiography; (L) subsequent recanalization after immediate mechanical thrombectomy.

examine language function in stroke patients during the acute stage. Therefore, a DL model using acute MRI data may be useful in clinical practice because it can evaluate language function regardless of patient cooperation and/or concomitant medical conditions.

The major strength of this study is that we developed a DL model to estimate aphasia severity in detail. Although several models have been used to predict the outcome of aphasia after stroke,[7,9,29,30] these models use conventional logistic regression to construct a model that uses only categorized image information (e.g., small or large lesions) or outcomes (e.g., good or bad outcomes). This process could have caused a marked loss of information. However, DL techniques can handle original data without significant modification, resulting in only a modest loss of information. Another strength of the DL model is its ability to handle large datasets. Although logistic regression approaches can also give rise to a model dealing with image features, the overfitting issue is always a concern in conventional models. In this study, we developed a logistic regression model, showing a comparable performance with our DL model, with only six clinical features (correlation coefficients for AQ scores, 0.70 vs. 0.72). However, logistic regression approaches failed to enhance performance when adding image features to clinical features. This unstable performance of logistic regression models raises the importance of DL algorithms, which can successfully process large-sized image features.[31] Additionally, in this study, we used imaging (DWI lesions) data as an input without categorization, and we developed a DL model that

https://doi.org/10.5853/jos.2021.02061

could estimate not only the total AQ score (continuous variable), but also the sub-domain scores of language function. The resulting performance is noteworthy. The DL model could estimate the categorized AQ score with moderate agreement (κ=0.59) and continuous AQ score with strong correlation (r=0.72, P<0.001). More detailed input data can improve model performance, while more detailed outputs can provide more information to clinicians and patients.

We noticed three patients with notable discrepancies between the model-estimated and true AQ scores (Table 3 and Figure 3). Case 1 presented with a few dot-like DWI lesions in the left internal carotid artery border zone territory on acute DWI with perfusion delay in the MCA area due to proximal internal carotid artery stenosis. Decreased perfusion and subsequent lesions in the language cortex may have resulted in severe aphasia in the true AQ score of the K-WAB test. However, we could not check the final lesions because follow-up MRI was not available for this patient. Case 2 exhibited scattered lesions in both anterior cerebral arterial territories involving the left anterior cingulate cortex, and the lesions and network disruption have been shown to be associated with apathy or an abulic state.[32,33] A decreased responsiveness to language tasks may have resulted in a poor K-WAB test. Alternatively, we should also consider that other unknown combined conditions not captured in the medical records (e.g., fluctuating delirium) may have affected the results. Case 3 underwent DWI immediately after symptom onset and early revascularization (onset to needle time, 45 minutes). Although we did not perform follow-up MRI, we speculated that acute lesions on initial DWI may have been reversed with neurological recovery. Reversible acute DWI lesions in patients thrombolyzed within 4.5 hours have been reported, and they are associated with early neurological improvement.[34] Taken together, these cases suggest that our DL model results should be interpreted individually according to each patient's clinical situation.

The DL technique also estimated the AQ score of patients who received acute thrombolysis/thrombectomy in our cohort. This is remarkable in that MRI findings at a time point early enough to perform acute interventions may also be useful in estimating the degree of aphasia at an early stage. However, the performance of the DL model in an acute intervention setting is likely to be reliable in patients with mild or very severe aphasia, but not in those with a medium degree of aphasia. Because intervention procedures may result in early neurological improvement or alteration, the model-predicted aphasia outcomes should be interpreted with caution in patients undergoing acute interventions (as in Case 3 in Table 3). Notwithstanding these aspects, the good performance of the DL model to estimate aphasia severity in many patients undergoing acute intervention may further raise the applicability of our DL technique in real-world practice.

In the present study, the DL model used a skip connection instead of a plain network (i.e., concatenated outputs of the 1st and 2nd hidden layers for input in the output layer; this is called the residual neural network). Previous studies have shown that the residual neural network architecture is useful to avoid the problem of vanishing gradients in multi-layer neural networks in case of an image recognition field.[35,36] In our DL model, it was also useful to reduce the time in the convergence speeds for training. Our DL model involved 90 and 30 nodes in the 1st and 2nd hidden layers, respectively, and showed the best performance in the measurement of correlation coefficients between measured and predicted AQ scores in our additional evaluations among the tested combinations of 10, 30, 50, 70, and 90 nodes for the 1st and 2nd hidden layers. The optimal number of nodes for the hidden layers may be limited in this study; however, our approach for DL modeling is promising for predicting AQ scores.

Our study has some limitations. First, it was a retrospective observational study with a single-center cohort, and therefore has a certain level of inherent bias. Furthermore, all patients used only Korean as their first language, which may have decreased the generalizability of the study results. In addition, we did not correlate early aphasia results with long-term aphasia. Although the severity of early aphasia is an important predictor of chronic aphasia,[5,26-28] most previous studies dealing with post-stroke aphasia have been dedicated to predicting the prognosis at 1 year after stroke, which is regarded as a "plateau" status. Therefore, long-term and prospective studies are warranted to broaden our results.[7,9,29,30] Third, we arbitrarily allocated patients into training and test sets according to the admission time. Although the baseline characteristics were not statistically different between the two groups, a multicenter cohort study with external validation could improve the generalizability of our model. Fourth, because we used multiple brain ATLAS templates with overlaps, image features may have been calculated and derived from overlapping templates. Thus, redundancies in image-feature information may be possible. However, using multiple templates is unavoidable to cover the entire brain area related to language function. Moreover, it should also be noted that we adopted regularization methods, such as Lasso, for logistic regression and dropout approach for DFFN, to minimize these biases in feature selection. Finally, we used only DWI data for the analysis and did not consider various MRI modalities. Multiple sequences of MRI may be helpful in estimating neurological phenotypes. For example, fluid-at-

tenuated inversion recovery and susceptibility-weighted imaging sequences show underlying old ischemic or hemorrhagic lesions; perfusion MRI demonstrates the penumbra that may predict the final infarction territory; and diffusion tensor tractography could predict the integrity of fasciculi of language domains. However, including only a simple MRI sequence may have enhanced the clinical utility of our DL model. Nevertheless, future studies to evaluate the usefulness of multiple MRI sequences in estimating the severity of aphasia are warranted.

Conclusions

Our study suggests that the DL model using DWI data may be feasible and useful in estimating the severity of aphasia in patients with acute stroke at an early stage. These findings warrant further research to evaluate the applicability of DL model in different study populations.

Supplementary materials

Supplementary materials related to this article can be found online at https://doi.org/10.5853/jos.2021.02061.

Disclosure

Yong-Hwan Kim was employed by company Nunaps Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (HI18C2383) and the Ministry of Science and ICT (NRF-2018M3A9E8066249), Republic of Korea.

References

1. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics-2019 update: a report from the American Heart Association. *Circulation* 2019;139:e56-e528.
2. Ellis C, Dismuke C, Edwards KK. Longitudinal trends in aphasia in the United States. *NeuroRehabilitation* 2010;27:327-333.
3. Ellis C, Simpson AN, Bonilha H, Mauldin PD, Simpson KN. The one-year attributable cost of poststroke aphasia. *Stroke* 2012;43:1429-1431.
4. Charidimou A, Kasselimis D, Varkanitsa M, Selai C, Potagas C, Evdokimidis I. Why is it difficult to predict language impairment and outcome in patients with aphasia after stroke? *J Clin Neurol* 2014;10:75-83.
5. Pedersen PM, Jørgensen HS, Nakayama H, Raaschou HO, Olsen TS. Aphasia in acute stroke: incidence, determinants, and recovery. *Ann Neurol* 1995;38:659-666.
6. Nouwens F, de Jong-Hagelstein M, De Lau LM, Dippel DW, Koudstaal PJ, van de Sandt-Koenderman WM, et al. Severity of aphasia and recovery after treatment in patients with stroke. *Aphasiology* 2014;28:1168-1177.
7. El Hachioui H, Lingsma HF, van de Sandt-Koenderman MW, Dippel DW, Koudstaal PJ, Visch-Brink EG. Long-term prognosis of aphasia after stroke. *J Neurol Neurosurg Psychiatry* 2013;84:310-315.
8. Nouwens F, Visch-Brink EG, El Hachioui H, Lingsma HF, van de Sandt-Koenderman MW, Dippel DWJ, et al. Validation of a prediction model for long-term outcome of aphasia after stroke. *BMC Neurol* 2018;18:170.
9. Payabvash S, Kamalian S, Fung S, Wang Y, Passanese J, Kamalian S, et al. Predicting language improvement in acute stroke patients presenting with aphasia: a multivariate logistic model using location-weighted atlas-based analysis of admission CT perfusion scans. *AJNR Am J Neuroradiol* 2010;31:1661-1668.
10. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the brain: artificial intelligence in stroke imaging. *J Stroke* 2017;19:277-285.
11. Lee H, Lee EJ, Ham S, Lee HB, Lee JS, Kwon SU, et al. Machine learning approach to identify stroke within 4.5 hours. *Stroke* 2020;51:860-866.
12. Kassner A, Thornhill RE. Texture analysis: a review of neurologic MR imaging applications. *AJNR Am J Neuroradiol* 2010;31:809-816.
13. Wernick MN, Yang Y, Brankov JG, Yourganov G, Strother SC. Machine learning in medical imaging. *IEEE Signal Process Mag* 2010;27:25-38.
14. Kim H, Na DL. Normative data on the Korean version of the Western Aphasia Battery. *J Clin Exp Neuropsychol* 2004;26:1011-1020.
15. Brott T, Adams HP Jr, Olinger CP, Marler JR, Barsan WG, Biller J, et al. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke* 1989;20:864-870.
16. Forsting M, Janen O. MR Neuroimaging: Brain, Spine, Peripheral Nerves. New York, NY: Thieme, 2017.
17. Kim BJ, Kim YH, Kim N, Kwon SU, Kim SJ, Kim JS, et al. Lesion location-based prediction of visual field improvement after cerebral infarction. *PLoS One* 2015;10:e0143882.

18. Plowman E, Hentz B, Ellis C Jr. Post-stroke aphasia prognosis: a review of patient-related and stroke-related factors. *J Eval Clin Pract* 2012;18:689-694.
19. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;15:1929-1958.
20. Raita Y, Goto T, Faridi MK, Brown DF, Camargo CA Jr, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019;23:64.
21. Chung MK. Statistical and Computational Methods in Brain Image Analysis. Boca Raton, FL: CRC Press, 2013.
22. Martinez-Martin P, Radicati FG, Rodriguez Blazquez C, Wetmore J, Kovacs N, Ray Chaudhuri K, et al. Extensive validation study of the Parkinson's disease composite scale. *Eur J Neurol* 2019;26:1281-1288.
23. Kertesz A. Western Aphasia Battery-Revised (WAB-R): Examiner's Manual. San Antonio, TX: PsychCorp, 2006.
24. Andersen R. Modern Methods for Robust Regression. Los Angeles, CA: Sage Publications, 2007.
25. Watila MM, Balarabe SA. Factors predicting post-stroke aphasia recovery. *J Neurol Sci* 2015;352:12-18.
26. Laska AC, Hellblom A, Murray V, Kahan T, Von Arbin M. Aphasia in acute stroke and relation to outcome. *J Intern Med* 2001;249:413-422.
27. Lazar RM, Minzer B, Antoniello D, Festa JR, Krakauer JW, Marshall RS. Improvement in aphasia scores after stroke is well predicted by initial severity. *Stroke* 2010;41:1485-1488.
28. Glize B, Villain M, Richert L, Vellay M, de Gabory I, Mazaux JM, et al. Language features in the acute phase of poststroke severe aphasia could predict the outcome. *Eur J Phys Rehabil Med* 2017;53:249-255.
29. Blom-Smink MR, van de Sandt-Koenderman MW, Lingsma HF, Heijenbrok-Kal MH, Ribbers GM. Predicting everyday verbal communicative ability after inpatient stroke rehabilitation in patients with moderate and severe aphasia at admission: validation of a prognostic model. *Eur J Phys Rehabil Med* 2019;55:532-534.
30. Godecke E, Rai T, Ciccone N, Armstrong E, Granger A, Hankey GJ. Amount of therapy matters in very early aphasia rehabilitation after stroke: a clinical prognostic model. *Semin Speech Lang* 2013;34:129-141.
31. Liu X, Gao K, Liu B, Pan C, Liang K, Yan L, et al. Advances in deep learning-based medical image analysis. *Health Data Sci* 2021;2021:8786793.
32. Das JM, Saadabadi A. Abulia. In: StatPearls. Treasure Island, FL: StatPearls Publishing, 2019. https://www.ncbi.nlm.nih.gov/books/NBK537093. Assessed December 1, 2021.
33. Siegel JS, Snyder AZ, Metcalf NV, Fucetola RP, Hacker CD, Shimony JS, et al. The circuitry of abulia: insights from functional connectivity MRI. *Neuroimage Clin* 2014;6:320-326.
34. Labeyrie MA, Turc G, Hess A, Hervo P, Mas JL, Meder JF, et al. Diffusion lesion reversal after thrombolysis: a MR correlate of early neurological improvement. *Stroke* 2012;43:2986-2991.
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, NV. IEEE; 2016;770-778. https://www.computer.org/csdl/proceedings-article/cvpr/2016/8851a770/12OmNxvwoXv. Assessed December 1, 2021.
36. Srivastava RK, Greff K, Schmidhuber J. Highway networks. arXiv 2015 Nov 3. https://arxiv.org/abs/1505.00387.

Supplementary Table 1. Performance of logistic regression and deep learning methods depending on input features in the training set

Performance in training set* (n=127)	AQ score	Spontaneous speech	Comprehension	Repetition	Naming
Logistic regression with LASSO					
Clinical feature only (6 features)	0.74	0.70	0.72	0.71	0.74
Imaging feature only (178 features)	0.79	0.82	0.75	0.79	0.76
Clinical+imaging features	0.84	0.75	0.82	0.84	0.80
Deep feed forward network					
Clinical+imaging features	0.86	0.88	0.84	0.88	0.85

AQ, aphasia quotient; LASSO, least absolute shrinkage and selection operator.
*Performance was calculated using the correlation coefficients between the actual and predicted Korean version of the Western Aphasia Battery (K-WAB) scores.

Supplementary Table 2. Performance of logistic regression and deep learning methods depending on the input features in the test set

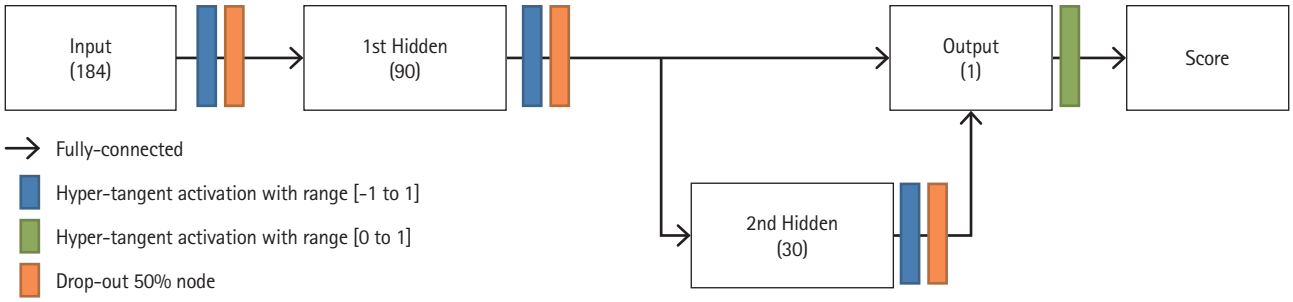
49 Performance in test set* (n=49)	AQ score	Spontaneous speech	Comprehension	Repetition	Naming
Logistic regression with LASSO					
Clinical feature only (6 features)	0.70	0.61	0.70	0.62	0.66
Imaging feature only (178 features)	0.54	0.64	0.54	0.53	0.50
Clinical+imaging features	0.63	0.67	0.65	0.59	0.59
Deep feed forward network					
Clinical+imaging features	0.72	0.75	0.71	0.65	0.71

AQ, aphasia quotient; LASSO, least absolute shrinkage and selection operator.
*Performance was calculated in the correlation coefficients between actual and predicted Korean version of the Western Aphasia Battery (K-WAB) scores.

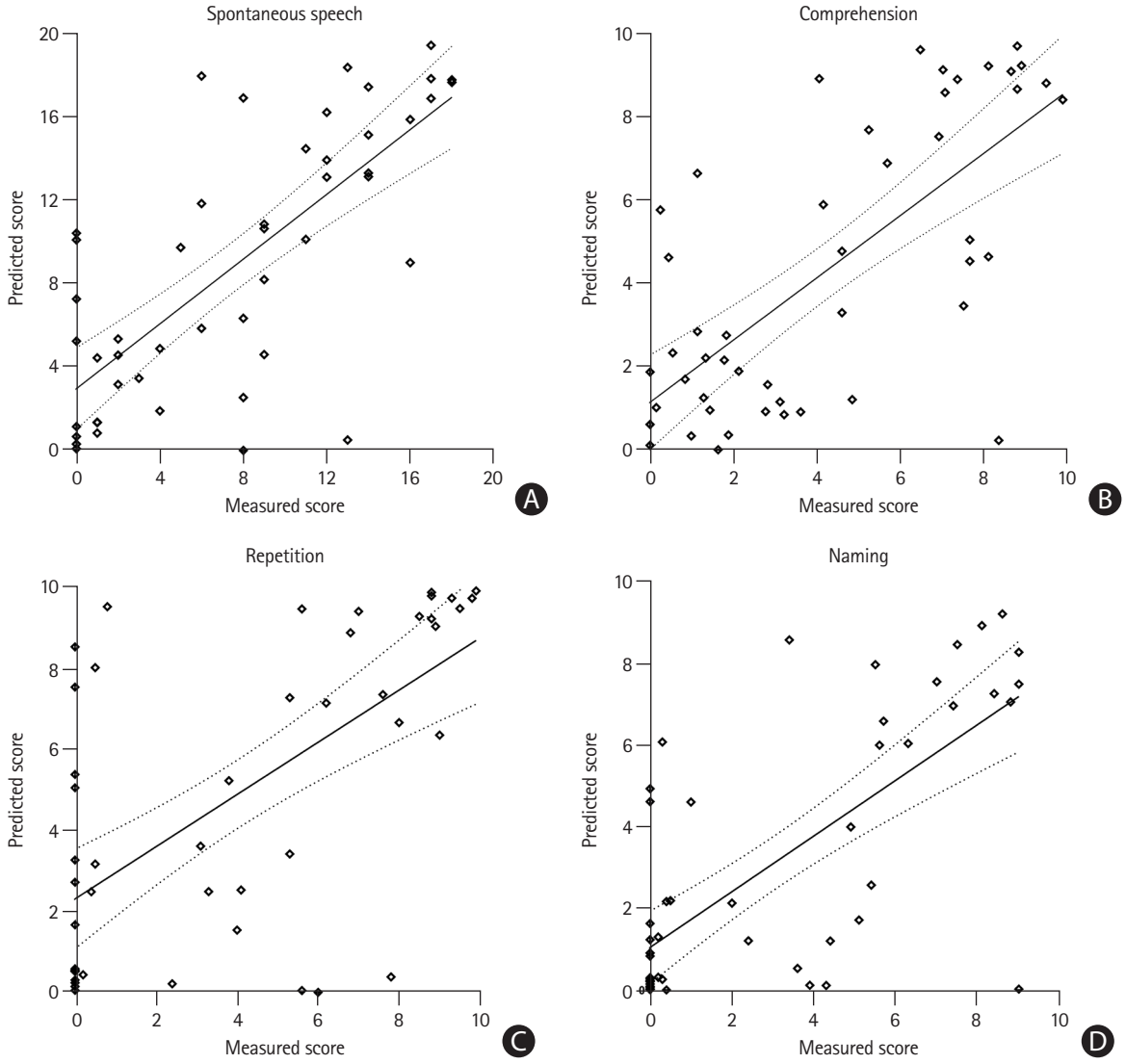
Supplementary Table 3. Contingency table between the true and model-predicted scores in patients who underwent acute intervention

	True score				Total
	Very severe	Severe	Moderate	Mild	
Model-predicted score					
Very severe	11	4	1	1	17
Severe	1	1	3	0	5
Moderate	0	0	2	2	4
Mild	0	2	1	7	10
Total	12	7	7	10	36
Accuracy	11/12 (92)	1/7 (14)	2/7 (29)	7/10 (70)	21/36 (58)

Values are presented as number (%). Cohen's weighted kappa, κ=0.60 (95% confidence interval, 0.41 to 0.78; P<0.001).

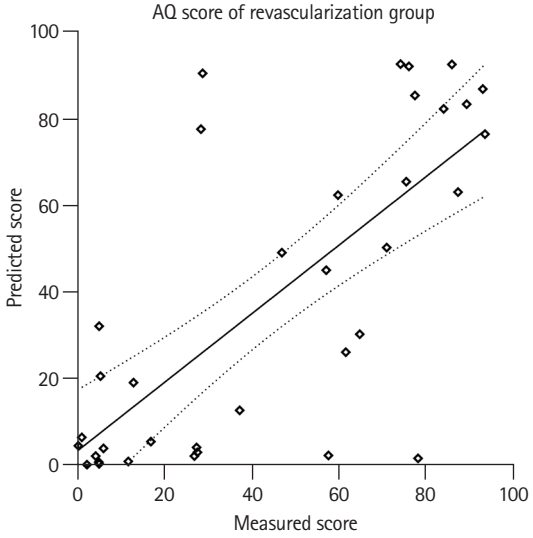


Supplementary Figure 1. Structure of the deep learning (DL) model. The number of nodes in each layer was noted in the box. For the input layer, 178 lesion occupying ratio features (left hemisphere associated regions from various atlases) and six clinical features (age, sex, Korean version of the Western Aphasia Battery [K-WAB] evaluation days from magnetic resonance imaging [MRI], MRI hours from onset, education years, and lesion volume) were used. Using the DL model, the final score was predicted to range from 0 to 1. The true score of the K-WAB was fed into the model by normalizing scores in the range of 0 to 1.



Supplementary Figure 2. Correlation analysis of sub-scores of aphasia quotient in the test set. (A) The correlation coefficient of spontaneous speech was 0.75 (95% confidence interval [CI], 0.59 to 0.85; $P<0.001$). (B) The correlation coefficient of repetition was 0.65 (95% CI, 0.44 to 0.78; $P<0.001$). (C) The correlation coefficient of comprehension was 0.71 (95% CI, 0.54 to 0.83; $P<0.001$). (D) The correlation coefficient of naming was 0.71 (95% CI, 0.54 to 0.83; $P<0.001$) (solid black line, regression line; dotted black line, 95% confidence limit).

https://doi.org/10.5853/jos.2021.02061



Supplementary Figure 3. Correlation analysis between true aphasia quotient (AQ) score and model-predicted AQ score of stroke patients undergone acute intervention. The correlation coefficient was 0.73 (95% confidence interval, 0.53 to 0.85; $P<0.001$).