Journal of the American Heart Association

ORIGINAL RESEARCH

Automated Extraction of Stroke Severity From Unstructured Electronic Health Records Using Natural Language Processing

Marta Fernandes, PhD; M. Brandon Westover, MD, PhD; Aneesh B. Singhal, MD, MBBS; Sahar F. Zafar, MD, MSc

BACKGROUND: Multicenter electronic health records can support quality improvement and comparative effectiveness research in stroke. However, limitations of electronic health record–based research include challenges in abstracting key clinical variables, including stroke severity, along with missing data. We developed a natural language processing model that reads electronic health record notes to directly extract the National Institutes of Health Stroke Scale score when documented and predict the score from clinical documentation when missing.

METHODS AND RESULTS: The study included notes from patients with acute stroke (aged ≥18 years) admitted to Massachusetts General Hospital (2015–2022). The Massachusetts General Hospital data were divided into training/holdout test (70%/30%) sets. We developed a 2-stage model to predict the admission National Institutes of Health Stroke Scale, obtained from the GWTG (Get With The Guidelines) stroke registry. We trained a model with the least absolute shrinkage and selection operator. For test notes with documented National Institutes of Health Stroke Scale, scores were extracted using regular expressions (stage 1); when not documented, least absolute shrinkage and selection operator was used for prediction (stage 2). The 2-stage model was tested on the holdout test set and validated in the Medical Information Mart for Intensive Care (2001–2012) version 1.4, using root mean squared error and Spearman correlation. We included 4163 patients (Massachusetts General Hospital, 3876; Medical Information Mart for Intensive Care, 287); average age, 69 (SD, 15) years; 53% men, and 72% White individuals. The model achieved a root mean squared error of 2.89 (95% CI, 2.62–3.19) and Spearman correlation of 0.92 (95% CI, 0.91–0.93) in the Massachusetts General Hospital test set, and 2.20 (95% CI, 1.69–2.66) and 0.96 (95% CI, 0.94–0.97) in the MIMIC validation set, respectively.

CONCLUSIONS: The automatic natural language processing–based model can enable large-scale stroke severity phenotyping from the electronic health record and support real-world quality improvement and comparative effectiveness studies in stroke.

Key Words: electronic health records ■ machine learning ■ National Institutes of Health Stroke Scale ■ phenotyping ■ stroke

Real-world studies using multicenter electronic health records (EHRs) can pave the way for understanding patterns and practice variation in stroke care that can support process improvement and treatment decisions.[1] EHRs can be particularly useful for quality-of-care–improving efforts,[2,3] investigating and addressing disparities in health care,[4,5] understanding gaps in health care delivery,[6] and designing effective measures to improve patient outcomes.[7,8] However, limitations of EHR-based research include challenges

Correspondence to: Marta Fernandes, PhD, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114. Email: mbentofernandes@mgh.harvard.edu

This manuscript was sent to Michelle H. Leppert, MD, MBA, Associate Editor, for review by expert referees, editorial decision, and final disposition.

Preprint posted on MedRxiv March 11, 2024. doi: https://doi.org/10.1101/2024.03.08.24304011.

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.124.036386

For Sources of Funding and Disclosures, see page 10.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVES

What Is New?

- The automatic natural language processing model presented enables automatic retrieval of National Institutes of Health Stroke Scale scores from unstructured data, overcoming key limitations of prior models that use administrative data or natural language processing for prediction of stroke severity.

What Are the Clinical Implications?

- Our model can enable large-scale stroke severity phenotyping from electronic health records and quality improvement studies to address process improvement, outcomes research, and health disparities in stroke care.

Nonstandard Abbreviations and Acronyms

GWTG	Get With The Guidelines
LASSO	least absolute shrinkage and selection operator
MGH	Massachusetts General Hospital
MIMIC-III	Medical Information Mart for Intensive Care III
NIHSS	National Institutes of Health Stroke Scale
NLP	natural language processing
RMSE	root mean squared error

in abstracting key clinical variables, such as stroke severity, from nonstructured data at scale.[9] This is further compounded by missing data.[10] Accurately measuring stroke severity from EHRs is critical for large-scale comparative effectiveness research and quality improvement.[11,12]

The National Institutes of Health Stroke Scale (NIHSS) is the gold standard for measuring stroke severity in the clinical environment.[9] The Joint Commission requires the NIHSS to be performed and documented on all patients who receive thrombolytics or undergo thrombectomy and those presenting within 12 hours of symptom onset.[13] Unfortunately, the NIHSS is not always documented in the EHR, and when documented, it is frequently in clinical notes and not as structured data.[9] The lack of standardized documentation creates limitations for performing population-level research or conducting quality improvement, particularly in the case of patients with acute ischemic stroke not meeting the criteria for acute treatments or interventions,[14] patients seen at smaller community or nonstroke centers[15–17] or admitted to non–Joint Commission accredited centers.[17] An additional limitation is that the NIHSS is not always documented for patients with hemorrhagic stroke. While scores including the Intracerebral Hemorrhage and Hunt and Hess scores are used for patients with hemorrhagic stroke, the NIHSS can be used to predict death and correlates with hemorrhage volume[18,19] and serves as a common score across all stroke subtypes for population-level research.[20,21]

While the NIHSS can be abstracted from clinical notes by chart review, this is labor intensive.[22] Prior studies have developed machine learning models using structured data, such as *International Classification of Diseases* (*ICD*) codes and Current Procedural Terminology codes to measure stroke severity.[23] However, models restricted to only structured data, besides not considering missing data,[10] leave out the extensive data available in unstructured notes that would allow for more accurate phenotyping. Existing natural language processing (NLP) models can be applied only to notes with a documented NIHSS score or its subcomponents and therefore are not applicable for missing data and have not been validated in other data sets.[24]

The lack of standardized documentation and reporting of the NIHSS in EHR databases, along with challenges in extracting the score particularly when the score itself or its subcomponents are missing, precludes the advancement of population-level stroke-related comparative effectiveness research.[25,26] Here, we aimed to address these challenges by developing an NLP model that automatically reads EHR clinical notes to accurately predict the NIHSS score of patients presenting with acute stroke. Our objective was to train an algorithm that can extract the score directly when documented, use additional clinical observations to predict the score when missing from the notes, and does not rely on documentation of the subcomponent parts. We used regional EHRs linked to the American Heart Association's (AHA) GWTG (Get With The Guidelines) stroke registry[27] as the gold standard data set and validated the model on an external independent data set. This model is intended to enable large-scale EHR stroke severity phenotyping even when the NIHSS or its subcomponents are not documented or recorded.

METHODS

Study Population

We included adult patients (aged ≥18 years) with acute stroke (ischemic and hemorrhagic). The study was approved by the Mass General Brigham Institutional Review Board; a waiver of informed consent was

obtained for this observational study. This study consists of retrospective data analysis and is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology statement.[28] Anonymized data and materials have been made publicly available at the StrokeNIHSS_NLP GitHub repository and can be accessed at: https://github.com/mpriscila88/StrokeNIHSS_NLP.

Data Sets

Our stroke cohort was derived from 2 sources: (1) Patients admitted with acute stroke at Massachusetts General Hospital (MGH) between March 2015 and December 2022 were identified through the American Heart Association's GWTG stroke registry linked with EHRs[27]; (2) stroke admissions from the publicly available deidentified database Medical Information Mart for Intensive Care III (MIMIC-III) version 1.4,[29] which contains medical records for intensive care unit admissions at Beth Israel Deaconess Medical Center between 2001 and 2012. Patients with acute stroke in the MIMIC data set were identified using the *ICD* coding system.[24] Further details on cohort derivation are provided in Data S1.

Clinical Notes

The EHR data in our study comprised free text admission notes (MGH) and discharge summaries (MIMIC), which consist of semistructured text written by physicians. The MGH notes were extracted for the first and second dates of admission, given our goal to measure and predict admission stroke severity. All available notes for these dates were extracted, including, among others, any neurology notes, emergency department, assessment and plan, history and physical, operative, procedures, consults, discharge instructions, discharge summaries, hospital course, progress, transfer, nursing, physical therapy, and occupational therapy notes. For model external validation, we used discharge summaries from the MIMIC data set, because the admission NIHSS scores are mainly stored in this type of notes. The discharge summaries include the patients' initial history and physical and admission clinical examinations. The notes from each patient were preprocessed (see Table S1) and converted into a structured format consisting of binary variables. Each variable indicated the presence or absence of an n-gram (single word [unigram], or sequence of 2 [bigram] or 3 [trigram] words) in the notes (see Data S1).

Outcomes and Gold Standard

Our outcome was the initial (admission) total NIHSS score. For the MGH cohort, the gold standard scores were obtained from the American Heart Association's GWTG—stroke registry.[27] For the MIMIC cohort, gold standard scores were obtained by applying rule-based regular expressions (regexes) to the notes for extraction of the scores, followed by manual note review for expert validity by a neurologist/neurointensivist (S.F.Z.) (see Data S1). After chart review, for MGH patients, notes with any discrepancy between the GWTG gold standard and the neurological examination documented in the note or the NIHSS, if also documented in the note, were removed. For MIMIC patients, any notes that did not have a documented NIHSS were removed.

Statistical Analysis

We split the MGH data randomly into a training set (70%) and a holdout test set (30%), as in previous studies.[30–32] With the training data, we developed a linear regression model using the least absolute shrinkage and selection operator (LASSO) that used the text-based features from the notes to predict the patients' NIHSS scores, from 0 to 42. We performed 100 iterations within the training data of 5-fold cross-validation to determine the best regularization parameter (see Data S1). The importance of the variables was assessed by magnitude of the regression coefficients.

We then created a 2-stage model, applied on the MGH holdout test set and externally validated on the MIMIC validation set: (1) In stage 1, notes were checked for the NIHSS and, if detected, the score was directly extracted. This stage used simple hard-coded regular expressions. (2) In stage 2, for notes in which the NIHSS score was not detected/documented, we applied the LASSO model to estimate the NIHSS from information contained in the note.

To evaluate the linear regression model, we used the root mean squared error (RMSE) and Spearman correlation. We performed 1000 iterations of bootstrap random sampling with replacement to calculate 95% CIs.

We report overall results for the 2-stage model on all notes in the MGH holdout test set and MIMIC validation set (those that contain extractable NIHSS scores and those that do not). We also report results separately for stage 1 (for notes with extractable NIHSS) and stage 2 (for notes in which the NIHSS scores were not documented and the LASSO model was used to predict the score).

We also developed an ordinal logistic regression model within the training data for each of 4 classes (NIHSS scores): minor stroke (0–4), moderate stroke (5–15), moderate to severe stroke (16–20), and severe stroke (21–42). We fitted a parallel adjacent category probability model with logit link, using the ordinalNet

R package.[33] We selected this type of model given that it focuses on comparisons of adjacent categories, comparing a response category to the next response category above it, which is suitable for prediction of probabilities in a severity scale. Because this was a parallel model, the coefficients of each predictor were identical for each stroke severity class except for the intercepts. Thus, the importance of the variables was assessed by magnitude of the regression coefficients. For each patient visit, we selected the maximum probability of the 4 probabilities from the ordinal regression model. Similar to the linear regression model, the ordinal regression used LASSO regularization utilizing the text-based features from the notes to predict the patients' NIHSS scores for each of the four classes. We also performed 100 iterations within the training data of 5-fold cross-validation to determine the best regularization parameter and then created the 2-stage model, as described above.

To evaluate the ordinal regression model, we used the area under the receiver operating characteristic curve, area under the precision recall curve, accuracy, and sensitivity (or recall) and specificity. We present the micro average performance, suited for imbalanced classes, which calculates these metrics globally by counting the total true positives, false negatives, and false positives. We present the detailed results for ordinal regression model in Data S1.

The codes for notes preprocessing and modeling along with deidentified data are publicly available in a GitHub repository.

RESULTS

Patients' Characteristics

Our study cohort for analysis (Figure 1) included a total of 4163 patients (MGH, n=3876; MIMIC, n=287) with an average age of 69 (SD, 15) years, majority men (53%), White individuals (72%), and presenting with ischemic stroke (90%). There were no observable differences in age, sex, or race between training, test, and external validation sets (Table 1). The median NIHSS score was 5 versus 13 for the MGH versus MIMIC data sets. When looking at the distribution of NIHSS scores (Figure 2), we observe that the majority of patients in the MGH cohort had a score <5, while the MIMIC cohort displays a higher frequency of scores between 6 and 23. For approximately the entire MIMIC cohort (99%), the admissions were of emergency type, while for the MGH cohort the admissions included emergency (85%), urgent (14%), and elective (1%) types.

Model Performance

The 2-stage model was able to predict the NIHSS score with an RMSE of 2.89 (95% CI, 2.62–3.19) in the MGH holdout test set and of 2.20 (95% CI, 1.69–2.66) in the MIMIC external validation set, as presented in Table 2. For both, the Spearman correlation was at least 90%. The performance for the extraction of the documented scores with regexes versus the gold standard (model stage 1) yielded an RMSE <1

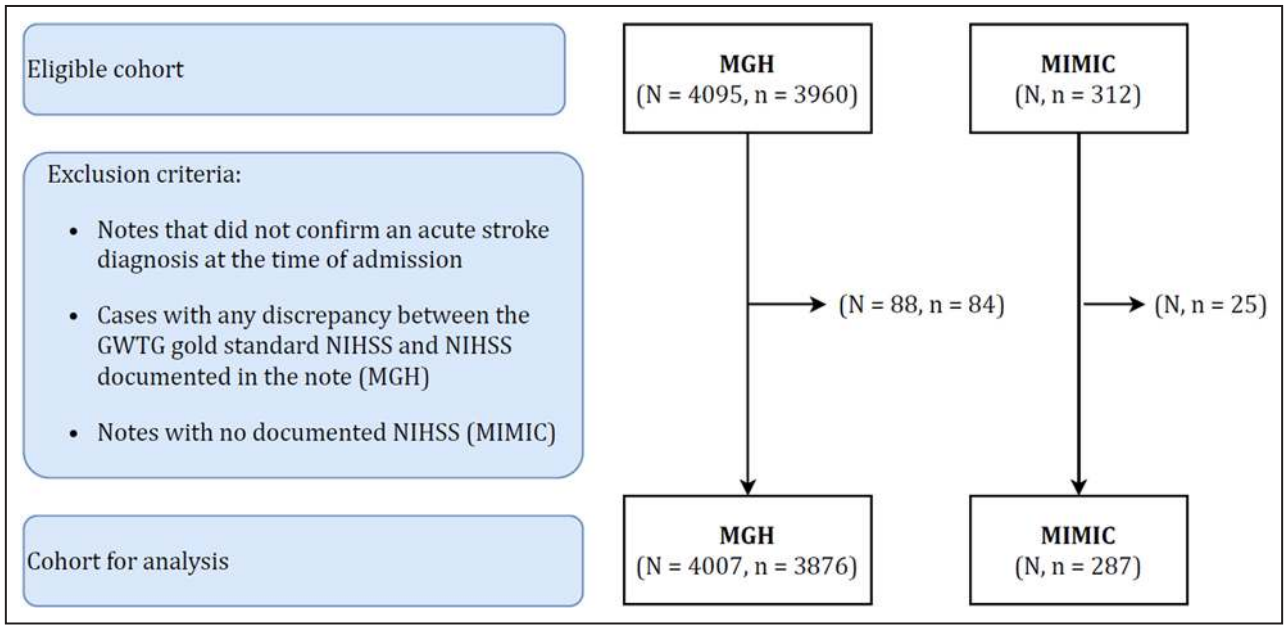

Figure 1. Consolidated Standards of Reporting Trials chart.
The number of patients is represented by n and the number of visits by N. GWTG indicates Get With The Guidelines Stroke registry; MGH, Massachusetts General Hospital; MIMIC, Medical Information Mart for Intensive Care; and NIHSS, National Institutes of Health Stroke Scale.

Table 1. Characteristics of the Study Population Stratified by Training, Test, and External Validation Sets

Characteristic	MGH			MIMIC
	Training set (n=2713)	Holdout test set (n=1163)	Full set (n=3876)	Validation set (n=287)
Age, y, mean±SD	68.4±15.3	69.1±14.7	68.6±15.1	70.3±15.9
Men, n (%)	1467 (54.1)	605 (52.0)	2072 (53.5)	148 (51.6)
Race, n (%)				
White	1958 (72.2)	826 (71.0)	2784 (71.8)	219 (76.3)
Black	201 (7.4)	82 (7.1)	283 (7.3)	23 (8.0)
Other*	554 (20.4)	255 (21.9)	809 (20.9)	45 (15.7)
Hispanic or Latino, n (%)	191 (7.0)	68 (5.8)	259 (6.7)	8 (2.8)
Hospital admissions, N	2814	1193	4007	287
NIHSS, median (IQR)	4 (1–12)	5 (2–12)	5 (2–12)	13 (7–19)
Admission type, N (%)				
Emergency	2404 (85.4)	1016 (85.2)	3420 (85.4)	285 (99.3)
Urgent	390 (13.9)	163 (13.7)	553 (13.8)	2 (0.70)
Elective	19 (0.7)	14 (1.2)	33 (0.8)	0 (0.0)
Type of stroke, N (%)				
Ischemic	2530 (89.9)	1085 (90.9)	3615 (90.2)	254 (88.5)
ICH	162 (5.8)	69 (5.8)	231 (5.8)	51 (17.8)
SAH	28 (1.0)	13 (1.1)	41 (1.0)	8 (2.8)
Other unspecified	94 (3.3)	26 (2.2)	120 (3.0)	6 (2.1)

The number of patients is represented by n and the number of visits is represented by N. ICH indicates intracerebral hemorrhage; IQR, interquartile range (25th–75th) percentiles; MGH, Massachusetts General Hospital; MIMIC, Medical Information Mart for Intensive Care; NIHSS, National Institutes of Health Stroke Scale; and SAH, subarachnoid hemorrhage.

* Includes Asian, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, and unknown race.

(0.84 [95% CI, 0.53, 1.19] for the MGH test set and 0.27 [95% CI, 0.00–0.47] for the MIMIC validation set) and Spearman correlation of ≈99%. The statistics regarding the number of documents with NIHSS used for the first stage of the model for both sets are presented in Table S2. We observed that the number of visits with only 1 NIHSS documented was superior for the external validation set (76%), compared with the holdout test set (44%). MIMIC discharge summaries have more succinct narratives, as compared with MGH admission notes that include narratives from >1 department and from multiple health care professionals.

The 2-stage ordinal regression model achieved an area under the receiver operating characteristic curve and area under the precision recall curve of 0.97 (95% CI, 0.96–0.97) and 0.92 (95% CI, 0.91–0.93), respectively, in the MGH holdout test set; and area under the receiver operating characteristic curve and area under the precision recall curve of 0.99 (95% CI, 0.98–1,00) and 0.97 (95% CI, 0.96–0.99), respectively, in the MIMIC validation set (Table S3).

We present the area under the receiver operating characteristic curve and area under the precision recall curve curve for the ordinal regression model in Figures S1 and S2, for both the MGH holdout test set and MIMIC validation set. Confusion matrices for the MGH test set are presented in Figure 3 and for MIMIC in Figure S3. We observed that the model presented higher error for the moderate to severe stroke class (NIHSS score, 16–20), in both test and validation sets.

Error Analysis

The distribution of the 2-stage model predicted versus target NIHSS is presented in Figure 4. We performed a manual review of the model errors to better understand the dispersion and identify the main types of misclassifications (see Table S4).

First, we identified cases with fluctuating symptoms, that is, improvement or worsening of the neurological examination. These changes in the examination occurred, for example, between teleconsultation/outside hospital presentation and arrival at the stroke center emergency department, or between initial emergency department presentation and neurological consultation or admission to the intensive care unit. In such cases, the model either predicted 1 of the scores, an average, or a score within the range documented. For example, in a case where the GWTG gold standard NIHSS score was 30, and the note documented improvement in the NIHSS score from 30 to 22, the model predicted an NIHSS score of 21. This case yielded an absolute error of 9 compared with the first documented score;

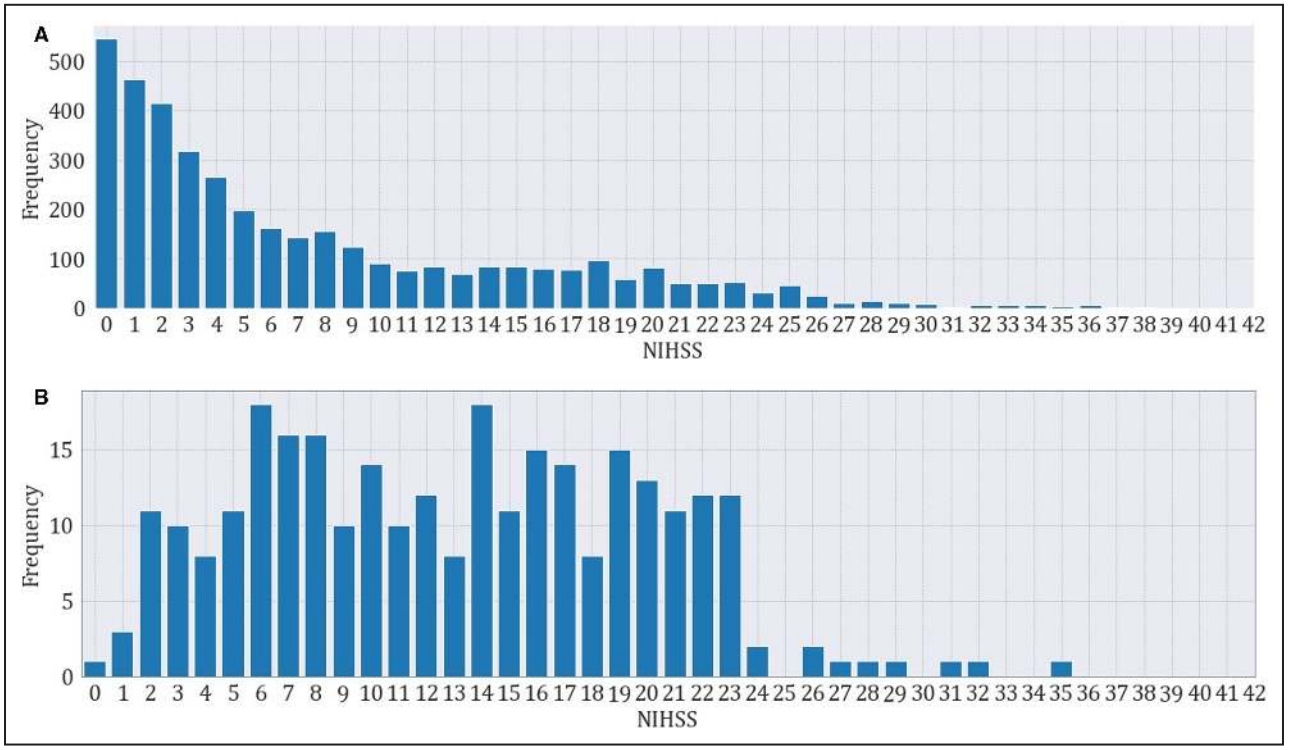

Figure 2. NIHSS score distribution for (A) MGH and (B) MIMIC cohorts.
MGH indicates Massachusetts General Hospital; and MIMIC, Medical Information Mart for Intensive Care.

however, the model prediction score of 21 was close to the improved score of 22.

We also identified cases in which the note indicated a range of scores but not the exact target score documented in the GWTG registry. An illustration of this scenario is a case in which the NIHSS documented in the note was "15±3," the GTWG gold standard was 14, and the model prediction was 15.71. Even though in this example the error was relatively small, it demonstrated that in certain cases a model error up to 3 points might be acceptable.

Finally, we observed that our model performance decreased for NIHSS scores >30. This is likely the result of a smaller number of cases with NIHSS scores >30, as depicted in Figure 2, and therefore not enough cases for model development and learning. Types of errors for greater scores included cases with fluctuating symptoms, as already illustrated, and cases with fewer clinical details documented in the examination sections of the notes, with many of the notes documenting the patients were "comatose" or "unresponsive," and fewer details on motor examination.

Features Importance

The initial number of variables in the training vocabulary was 7565, which was reduced to 347 by LASSO regularization (see Table S5). We assessed the degree of missingness of these variables in the data before and after feature selection for each data set (Table S6). Of the initial set of 7565 variables, on average each patient had ≈30%, 28%, and 6% of the variables present in the notes per visit, in the training, testing, and validation sets, respectively. After feature selection, of the reduced set of 347 variables, the percentages increased to 39%, 37%, and 15% in the training, testing, and validation sets, respectively.

Table 2. Performance of the 2-Stage Model and Each Individual Model Stage

	2-stage model		Stage 1 (regexes)		Stage 2 (LASSO)	
Cohort	RMSE (95% CI)	SC (95% CI)	RMSE (95% CI)	SC (95% CI)	RMSE (95% CI)	SC (95% CI)
MGH holdout test set	2.89 (2.62–3.19)	0.92 (0.91–0.93)	0.84 (0.53–1.19)	0.99 (0.98–0.99)	3.78 (3.43–4.14)	0.88 (0.87–0.90)
MIMIC validation set	2.20 (1.69–2.66)	0.96 (0.94–0.97)	0.28 (0.00–0.47)	1.00 (0.99–1.00)	4.49 (3.70–5.33)	0.82 (0.71–0.90)

MGH indicates Massachusetts General Hospital; MIMIC, Medical Information Mart for Intensive Care; regexes, regular expressions; RMSE, root mean squared error; and SC, Spearman correlation.

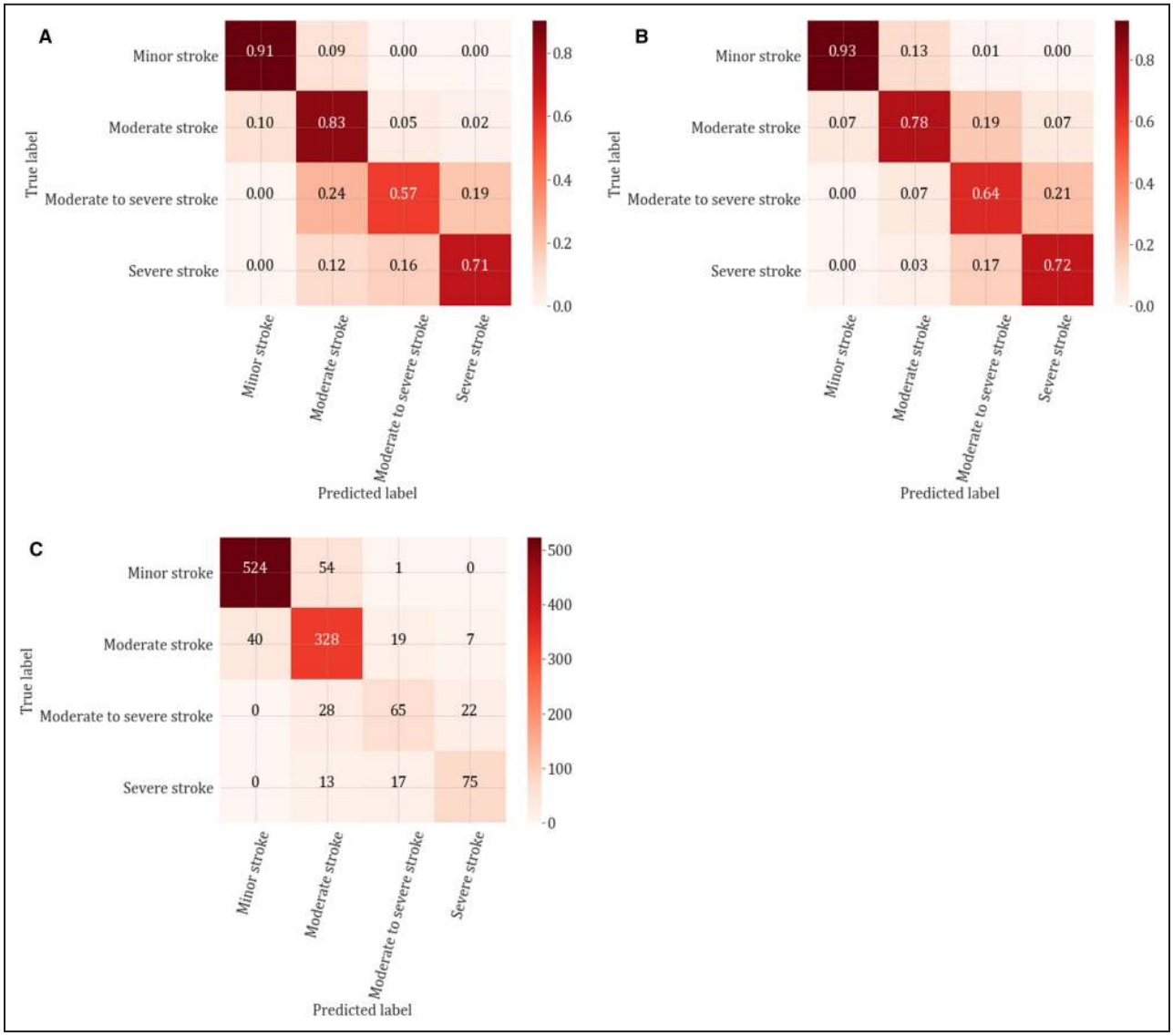

Figure 3. Confusion matrices for the 2-stage ordinal regression model in the Massachusetts General Hospital holdout test set, (A) normalized by recall; (B) normalized by precision; (C) without normalization.

The importance of the top 20 modeling variables is presented in Figure 5. We observe that indication of lack of movement and not answering questions or following commands are all associated with higher stroke severity scores. Gaze deviation, use of urinary catheter, a middle cerebral artery stroke, facial palsy, and paralysis are also associated with higher scores. On the other hand, "drift" contributed the most to prediction of lower scores and was associated with lower scores for the NIHSS motor subcomponents (arms and legs). A level of consciousness of "alert and responsive," as well as the word *deni* (denies), indicating a patient is able to communicate, also contributed to lower scores.

The relative importance of features for ordinal regression was also assessed (Figure S4). Overall, the importance estimates of features was similar to those of the linear model. Similar to features associated with higher scores in the linear model, we observed that indication of lack of movement, not following commands, and gaze deviation were considered important features in the ordinal model. Similar to the linear model, "drift" and "alert" were associated with lower scores in the ordinal model.

DISCUSSION

In this work, we developed an NLP model that automatically "reads" EHR clinical notes to determine the NIHSS score of adult patients presenting with acute stroke. The linear model for prediction of scores from 0 to 42 achieved good performance with an RMSE of 2.89 (95% CI, 2.62–3.19) and 2.20 (95% CI, 1.69–2.66) and Spearman correlation of 0.92 (95% CI, 0.91–0.93)

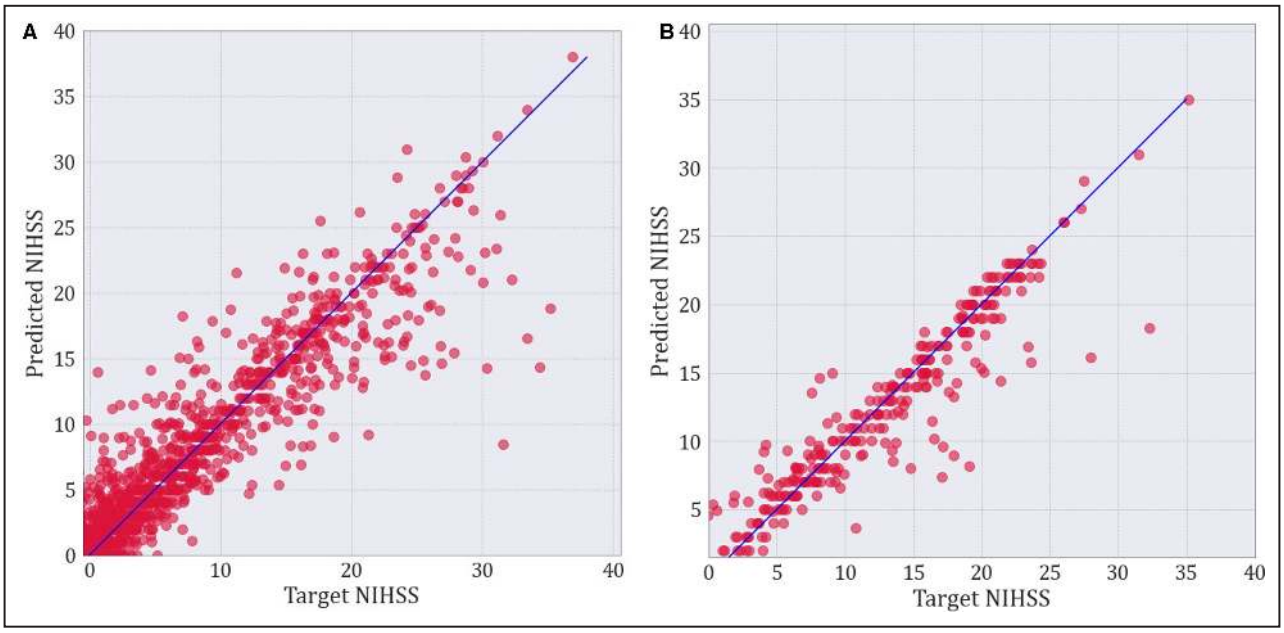

Figure 4. Two-stage model predicted versus target NIHSS scores in the (A) MGH holdout test set and (B) MIMIC external validation set.
MGH indicates Massachusetts General Hospital; MIMIC, Medical Information Mart for Intensive Care; and NIHSS, National Institutes of Health Stroke Scale.

and 0.96 (95% CI, 0.94–0.97) in test and validation, respectively. The modeling variables most indicative of higher scores were related to patients being unresponsive, with middle cerebral artery stroke, urinary catheter, and other neurological deficits, such as facial palsy, paralysis, or gaze deviation. The variables contributing the most to lower scores were related to the patients being alert and responsive, with lower scores in the motor subcomponents of the scale. The ordinal regression model for prediction of 4 classes—namely, minor, moderate, moderate to severe, and severe stroke—also achieved good performance with areas under the curve >90% for both data sets. Our findings suggest that automated EHR phenotyping of stroke is feasible in large and diverse cohorts.

NIHSS scores are not always available as structured information in EHR systems, and often when available they are documented in notes rather than in structured data.[9] NIHSS scores are also less commonly documented for patients presenting beyond 12 hours of symptoms onset, or those who do not receive thrombolytics and thrombectomies.[13] Automated EHR phenotyping algorithms that leverage large-scale nonstructured data to measure NIHSS scores can enable health services and population-level research, including prognostication and resource allocation in stroke care.

A prior study[23] used administrative data including *ICD* and Current Procedural Terminology codes, demographics, prescriptions/medications, hospital visit information, and comorbidities to predict the NIHSS score. Using machine learning to assess stroke severity, the authors found that the main predictors included death within the same month as stroke occurrence, length of hospital stay following stroke occurrence, aphagia/dysphagia diagnosis, hemiplegia diagnosis, and whether a patient was discharged to home or self-care. Comparing the imputed NIHSS scores with the gold standard on the holdout test set yielded an RMSE of 4.5 and Pearson correlation of 0.76. Based on the higher performance of our model, we hypothesize that using NLP and unstructured text notes can more accurately measure NIHSS score, compared with using administrative data alone.

Another study[24] combined the BERT (Bidirectional Encoder Representations from Transformers)-BiLSTM (bidirectional Long Short-Term Memory)-CRF (conditional random field) and a random forests model for the task of NIHSS item and score recognition, achieving an F1 score of 0.90, which outperformed their rule-based method (regexes) with an F1 score of 0.81. The NIHSS item extraction showed best performance when using the rule-based method (regexes) with a precision of 1.00, recall of 0.95, and F1 score of 0.97. The study was, however, focused only on extraction of the scores from notes when documented and therefore cannot be applied when data are missing. On the other hand, our work focused both on extraction of the scores, when documented (stage 1) and also on developing a model to predict the scores (stage 2) and evaluate these against a gold standard. Our model can therefore be used to predict NIHSS from notes, even when missing or not documented.

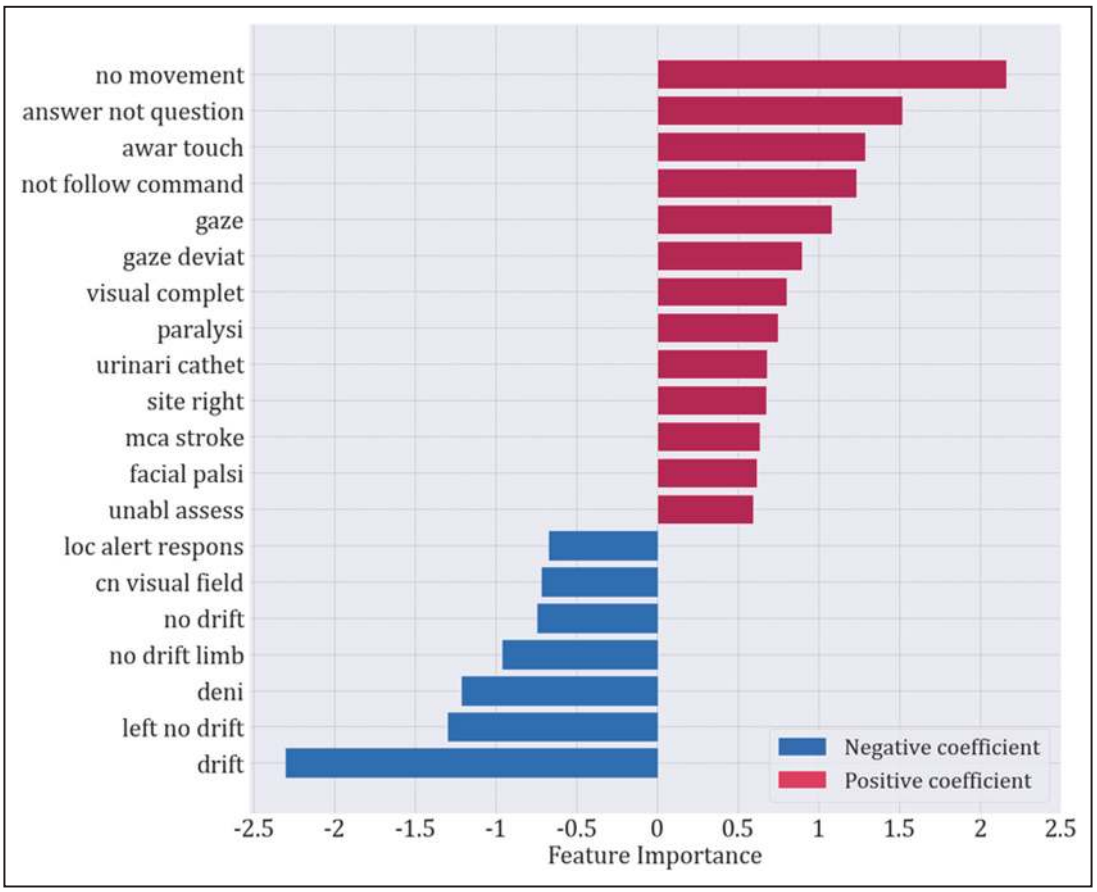

Figure 5. Feature importance given by the LASSO model coefficients for the top 20 features.
LASSO indicates least absolute shrinkage and selection operator.

Other studies have predicted stroke severity by approaching the data as a binary task problem[32,34] and thus are not directly comparable to ours. In one study,[34] a 3-dimensional convolutional neural network was trained to predict low (NIHSS score <5) versus high (NIHSS score ≥5) on the basis of preprocessed diffusion-weighted imaging images. The NIHSS category was predicted at admission and on day 7 of hospitalization, achieving an area under the receiver operating characteristic curve of 0.85 and 0.90, respectively. However, limiting to a binary classification limits the utility of the model, especially in light of an NIHSS score of 6 being put in the same category as an NIHSS score of 42. In a different study,[32] the authors developed a random forests model that achieved a recall, precision, and F1 score of 0.91 to predict improvement versus worsening of NIHSS progression as a binary outcome from hospital admission until discharge, using baseline data within the first 72 hours of admission. By predicting NIHSS score as a continuous variable and enabling the prediction of NIHSS score even when it is not documented or is missing from the EHR, our model overcomes the limitations of these prior works.

Our study has some limitations. While we validated our model in data from a different hospital, both hospitals are located in the same geographic region (Boston, MA). Nevertheless, the hospitals have different EHR systems, providers, and note types that increase generalizability. Future studies are required to assure generalizability of the model in other US and non-US populations. In future studies we will utilize the model across different hospitals and EHR systems in the United States. Our model performance decreased for NIHSS scores >30 due to fewer cases with high scores. Therefore, the model was not able to fully learn the patterns or main drivers for prediction of scores in that range. However, on closer examination of these cases, the patients' clinical examinations were poor, with fewer clinical details documented for motor examination. Other cases included patients with fluctuating symptoms in the notes, making it difficult for the model to provide an accurate prediction. Future studies should investigate the model performance when developing the model with cohorts including higher NIHSS scores, using the methodology presented here. Also, the model was trained with data from admission notes, and it was externally validated in discharge summary notes. Nonetheless, we did assess performance in a holdout test set of admission notes

and observed that the error was approximately the same. Additionally, the discharge summaries included the initial history and physical along with the admission examination. Another limitation consisted of the relatively small sample size of the MIMIC cohort, with a different distribution of NIHSS scores, compared with that of the MGH cohort. Thus, in future work we aim to validate the model in other cohorts with more diverse score distributions. Finally, the goal of our study was to extract or predict NIHSS score from the EHR, rather than deriving a new score. Thus, we cannot directly apply the NIHSS subcomponent scores to our model features to compute the score.

The automatic NLP model presented herein enables automatic retrieval of NIHSS scores from unstructured data, thereby enabling large-scale stroke severity phenotyping from EHRs. This work overcomes key limitations of prior models that use administrative data or NLP for prediction of stroke severity. Our model can enable real-world research and quality improvement studies to address process improvement, outcomes research, and health disparities in stroke care. Future directions include validating this model in additional electronic health data sets.

ARTICLE INFORMATION

Received May 7, 2024; accepted August 26, 2024.

Affiliations

Department of Neurology, Massachusetts General Hospital (MGH), Boston, MA (M.F., A.B.S., S.F.Z.); and Department of Neurology, Beth Israel Deaconess Medical Center (BIDMC), Boston, MA (M.B.W.).

Sources of Funding

M.B.W. was supported by grants from the National Institutes of Health (RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119), and the National Science Foundation (2014431). S.F.Z. was supported by the National Institutes of Health (K23NS114201, R01NS126282, R01AG082693, R01NS131347).

Disclosures

Dr Zafar is a clinical neurophysiologist for Corticare, received speaking honoraria from Marinus, and received royalties from Springer Publishing, unrelated to this work. Dr Westover is a cofounder, scientific advisor, and consultant to Beacon Biosignals and has a personal equity interest in the company. He receives royalties for authoring *Pocket Neurology* from Wolters Kluwer and *Atlas of Intensive Care Quantitative EEG* by Demos Medical. None of these interests played any role in the present work. The remaining authors have no disclosures to report.

Supplemental Material

Data S1
Tables S1–S6
Figures S1–S4
References 35–38

REFERENCES

1. Fonarow GC, Kapral MK, Schwamm LH. Future of quality and outcomes research in stroke. *Circ Cardiovasc Qual Outcomes*. 2015;8:S66–S68. doi: 10.1161/CIRCOUTCOMES.115.002309
2. Payne TH. The electronic health record as a catalyst for quality improvement in patient care. *Heart Br Card Soc*. 2016;102:1782–1787. doi: 10.1136/heartjnl-2015-308724
3. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, Specchia ML. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. *Eur J Pub Health*. 2016;26:60–64. doi: 10.1093/eurpub/ckv122
4. Smith MA, Gigot M, Harburn A, Bednarz L, Curtis K, Mathew J, Farrar-Edwards D. Insights into measuring health disparities using electronic health records from a statewide network of health systems: a case study. *J Clin Transl Sci*. 2023;7:e54. doi: 10.1017/cts.2022.521
5. Rumball-Smith J, Bates DW. The electronic health record and health IT to decrease racial/ethnic disparities in care. *J Health Care Poor Underserved*. 2018;29:58–62. doi: 10.1353/hpu.2018.0006
6. Adane K, Gizachew M, Kendie S. The role of medical data in efficient patient care delivery: a review. *Risk Manag Healthc Policy*. 2019;12:67–73. doi: 10.2147/RMHP.S179259
7. Uslu A, Stausberg J. Value of the electronic medical record for hospital care: update from the literature. *J Med Internet Res*. 2021;23:e26323. doi: 10.2196/26323
8. Wani D, Malhotra M. Does the meaningful use of electronic health records improve patient outcomes? *J Oper Manag*. 2018;60:1–18. doi: 10.1016/j.jom.2018.06.003
9. Lyden P. Using the National Institutes of Health stroke scale. *Stroke*. 2017;48:513–519. doi: 10.1161/STROKEAHA.116.015434
10. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Med Inform*. 2018;6:e8960. doi: 10.2196/medinform.8960
11. Patient-Centered Outcomes Research Institute. National Priorities for Research and Research Agenda. 2012. Accessed February 6, 2024. https://www.pcori.org/assets/PCORI-National-Priorities-and-Research-Agenda-2012-05-21-FINAL1.pdf. Adopted by PCORI Board of Governors.
12. Committee on Comparative Effectiveness Research Prioritization. *Initial National Priorities for Comparative Effectiveness Research*. National Academies Press; 2009.
13. Leifer D, Bravata DM, Connors JJ, Hinchey JA, Jauch EC, Johnston SC, Latchaw R, Likosky W, Ogilvy C, Qureshi AI, et al. Metrics for measuring quality of Care in Comprehensive Stroke Centers: detailed follow-up to brain attack coalition comprehensive stroke center recommendations. *Stroke*. 2011;42:849–877. doi: 10.1161/STR.0b013e318208eb99
14. Jovin TG, Albers GW, Liebeskind DS; STAIR IX Consortium. Stroke treatment academic industry roundtable: the next generation of endovascular trials. *Stroke*. 2016;47:2656–2665. doi: 10.1161/STROKEAHA.116.013578
15. Williams LS, Yilmaz EY, Lopez-Yunez AM. Retrospective assessment of initial stroke severity with the NIH stroke scale. *Stroke*. 2000;31:858–862. doi: 10.1161/01.STR.31.4.858
16. Kasner SE, Chalela JA, Luciano JM, Cucchiara BL, Raps EC, McGarvey ML, Conroy MB, Localio AR. Reliability and validity of estimating the NIH stroke scale score from medical records. *Stroke*. 1999;30:1534–1537. doi: 10.1161/01.STR.30.8.1534
17. Song S, Fonarow GC, Olson DM, Liang L, Schulte PJ, Hernandez AF, Peterson ED, Reeves MJ, Smith EE, Schwamm LH, et al. Association of get with the guidelines-stroke program participation and clinical outcomes for medicare beneficiaries with ischemic stroke. *Stroke*. 2016;47:1294–1302. doi: 10.1161/STROKEAHA.115.011874
18. Specogna AV, Patten SB, Turin TC, Hill MD. The reliability and sensitivity of the National Institutes of Health stroke scale for spontaneous intracerebral hemorrhage in an uncontrolled setting. *PLoS One*. 2013;8:e84702. doi: 10.1371/journal.pone.0084702
19. You S, Zheng D, Yoshimura S, Ouyang M, Han Q, Wang X, Cao Y, Delcourt C, Song L, Arima H, et al. Optimum baseline clinical severity scale cut points for prognosticating intracerebral hemorrhage: INTERACT studies. *Stroke*. 2024;55:139–145. doi: 10.1161/STROKEAHA.123.044538
20. Finocchi C, Balestrino M, Malfatto L, Mancardi G, Serrati C, Gandolfo C. National Institutes of Health stroke scale in patients with primary intracerebral hemorrhage. *Neurol Sci*. 2018;39:1751–1755. doi: 10.1007/s10072-018-3495-y
21. Andersen KK, Olsen TS, Dehlendorff C, Kammersgaard LP. Hemorrhagic and ischemic strokes compared. *Stroke*. 2009;40:2068–2072. doi: 10.1161/STROKEAHA.108.540112

22. Kasner SE, Cucchiara BL, McGarvey ML, Luciano JM, Liebeskind DS, Chalela JA. Modified National Institutes of Health stroke scale can be estimated from medical records. *Stroke*. 2003;34:568–570. doi: 10.1161/01.STR.0000052630.11159.25
23. Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. *BMC Med Inform Decis Mak*. 2020;20:8. doi: 10.1186/s12911-019-1010-x
24. Yang L, Huang X, Wang J, Yang X, Ding L, Li Z, Li J. Identifying stroke-related quantified evidence from electronic health records in real-world studies. *Artif Intell Med*. 2023;140:102552. doi: 10.1016/j.artmed.2023.102552
25. Katzan IL, Spertus J, Bettger JP, Bravata DM, Reeves MJ, Smith EE, Bushnell C, Higashida RT, Hinchey JA, Holloway RG, et al. Risk adjustment of ischemic stroke outcomes for comparing hospital performance: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2014;45:918–944. doi: 10.1161/01.str.0000441948.35804.77
26. Fonarow GC, Alberts MJ, Broderick JP, Jauch EC, Kleindorfer DO, Saver JL, Solis P, Suter R, Schwamm LH. Stroke outcomes measures must be appropriately risk adjusted to ensure quality care of patients: a presidential advisory from the American Heart Association/American Stroke Association. *Stroke*. 2014;45:1589–1601. doi: 10.1161/STR.0000000000000014
27. Reeves MJ, Smith EE, Fonarow GC, Zhao X, Thompson M, Peterson ED, Schwamm LH, Olson D. Variation and trends in the documentation of National Institutes of Health stroke scale in GWTG-stroke hospitals. *Circ Cardiovasc Qual Outcomes*. 2015;8:S90–S98. doi: 10.1161/CIRCOUTCOMES.115.001775
28. Vandenbroucke JP, Elm E v, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Med*. 2007;4:e297. doi: 10.1371/journal.pmed.0040297
29. Johnson AE, Pollard TJ, Shen L, Lehman L, wei H, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. doi: 10.1038/sdata.2016.35
30. Mayampurath A, Parnianpour Z, Richards CT, Meurer WJ, Lee J, Ankenman B, Perry O, Mendelson SJ, Holl JL, Prabhakaran S. Improving prehospital stroke diagnosis using natural language processing of paramedic reports. *Stroke*. 2021;52:2676–2679. doi: 10.1161/STROKEAHA.120.033580
31. Guan W, Ko D, Khurshid S, Lipsanopoulos ATT, Ashburner JM, Harrington LX, Rost NS, Atlas SJ, Singer DE, McManus DD, et al. Automated electronic phenotyping of cardioembolic stroke. *Stroke*. 2021;52:181–189. doi: 10.1161/STROKEAHA.120.030663
32. Gkantzios A, Kokkotis C, Tsiptsios D, Moustakidis S, Gkartzonika E, Avramidis T, Tripsianis G, Iliopoulos I, Aggelousis N, Vadikolias K. From admission to discharge: predicting National Institutes of Health stroke scale progression in stroke patients using biomarkers and explainable machine learning. *J Pers Med*. 2023;13:1375. doi: 10.3390/jpm13091375
33. Wurm MJ, Rathouz PJ, Hanlon BM. Regularized ordinal regression and the ordinalNet R package. *J Stat Softw*. 2021;99(6). doi: 10.18637/jss.v099.i06
34. Zeng Y, Long C, Zhao W, Liu J. Predicting the severity of neurological impairment caused by ischemic stroke using deep learning based on diffusion-weighted images. *J Clin Med*. 2022;11:4008. doi: 10.3390/jcm11144008
35. Woodfield R, Grant I; Group UBSO, Group UBFU and OW, Sudlow CLM. Accuracy of electronic health record data for identifying stroke cases in large-scale epidemiological studies: a systematic review from the UK biobank stroke outcomes group. *PLoS One*. 2015;10:e0140533. doi: 10.1371/journal.pone.0140533
36. Mitchell JD, Collen JF, Petteys S, Holley AB. A simple reminder system improves venous thromboembolism prophylaxis rates and reduces thrombotic events for hospitalized patients1. *J Thromb Haemost*. 2012;10:236–243. doi: 10.1111/j.1538-7836.2011.04599.x
37. Porter MF. An algorithm for suffix stripping. *Program Electron Libr Inf Syst*. 1980;14:130–137. doi: 10.1108/eb046814
38. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58:267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x