Original Research Article

EUROPEAN STROKE JOURNAL

European Stroke Journal
2025, Vol. 10(3) 829–836

DOI: 10.1177/23969873251314340
journals.sagepub.com/home/eso
Sage

Automated extraction of post-stroke functional outcomes from unstructured electronic health records

Marta Fernandes[1], Kaileigh Gallagher[2], Niels Turley[2], Aditya Gupta[1], M Brandon Westover[2], Aneesh B Singhal[1] and Sahar F Zafar[1]

Abstract

Purpose: Population level tracking of post-stroke functional outcomes is critical to guide interventions that reduce the burden of stroke-related disability. However, functional outcomes are often missing or documented in unstructured notes. We developed a natural language processing (NLP) model that reads electronic health records (EHR) notes to automatically determine the modified Rankin Scale (mRS).

Method: We included consecutive patients (⩾18 years) with acute stroke admitted to our center (2015–2024). mRS scores were obtained from the Get With the Guidelines registry and clinical notes (if documented), and used as the gold standard to compare against NLP-generated scores. We used text-based features from notes, along with age, sex, discharge status, and outpatient follow-up to train a logistic regression for prediction of good (0–2) versus poor (3–6) mRS, and a linear regression for the full range of mRS scores. The models were trained for prediction of mRS at hospital discharge and post-discharge. The models were externally validated in a dataset of patients with brain injuries from a different healthcare center.

Findings: We included 5307 patients, 5006 in train and test and 301 in validation; average age was 69 (SD 15) and 65 (SD 17) years, respectively; 47% female. The logistic regression achieved an area under the receiver operating curve (AUROC) of 0.94 [CI 0.93–0.95] (test) and 0.94 [0.91–0.96] (validation), and the linear model a root mean squared error (RMSE) of 0.91 [0.87–0.94] (test) and 1.17 [1.06–1.28] (validation).

Discussion and Conclusion: The NLP-based model is suitable for use in large-scale phenotyping of stroke functional outcomes and population health research.

Keywords

Modified Rankin Scale, functional outcomes, electronic health records, phenotyping, natural language processing, machine learning

Date received: 7 October 2024; accepted: 5 January 2025

Introduction

Stroke is a leading cause of disability worldwide.[1] Stroke related disability poses a significant burden on stroke survivors, their families and caregivers, with associated economic costs.[2,3] Disparities in access to acute and post-acute care across age groups and geographic locations augment the impact of stroke related disability, making it a public health priority.[1] Population level tracking of stroke outcomes is critical for quality and process improvement, and to guide policies directed at reducing the burden of stroke related disability.[1] Electronic health records (EHR) are a rich source of data for tracking outcomes.[4] The Joint Commission recommends documentation of the modified Rankin Scale (mRS), a functional disability score, in patients with stroke as a national quality measure.[5] Unfortunately the mRS score is rarely documented in the EHR,[6] although

[1]Department of Neurology, Massachusetts General Hospital (MGH), Boston, MA, USA
[2]Department of Neurology, Beth Israel Deaconess Medical Center (BIDMC), Boston, MA, USA

Corresponding author:
Marta Fernandes, Department of Neurology, Massachusetts General Hospital (MGH), 55 Fruit Street, Boston, MA 02114, USA.
Email: mbentofernandes@mgh.harvard.edu

neurological deficits and functional status are commonly reported in non-structured clinical notes.[7] Manual abstraction of mRS from unstructured electronic records for population level studies is infeasible, labor intensive and time-consuming.[7] To address this challenge, we aimed to develop natural language processing (NLP) models to phenotype stroke outcomes in the real-world setting from existing clinical data in EHR. The NLP models we developed automatically read clinical notes combined with demographics, discharge status (deceased vs alive) and status of post-discharge follow-up visits, to predict the mRS score of patients with acute stroke at hospital discharge and post-discharge.

Patients and methods

Study population

We included consecutive patients (⩾18 years old) with acute ischemic or hemorrhagic stroke, admitted to our center between January 2015 and January 2024. Patients were identified using the American Heart Association's (AHA) Get With The Guidelines (GWTG)—Stroke Registry.[8] The external validation cohort included consecutive patients with brain injuries (ischemic and hemorrhagic stroke, traumatic brain injury, anoxic brain injury, neuroinfectious diseases, seizures, and toxic metabolic encephalopathy) admitted to a different healthcare system between June 2023 and November 2024.

This study consists of retrospective data analysis and is reported in accordance with the Artificial Intelligence Transparent Reporting of a multivariable model for Individual Prognosis Or Diagnosis TRIPOD statement.[9]

Clinical data

The AHA-GWTG-Stroke registry and validation cohort had prospectively collected data on patient demographics, stroke type, discharge status (alive vs deceased), and disposition (home self-care, home-health services, acute care facility—skilled nursing, rehab and hospice medical facilities, death and others). For training and external validation cohorts, we extracted EHR free text notes from the inpatient hospitalization and post-discharge outpatient visits. We included physical therapy (PT) and occupational therapy (OT) notes, discharge summaries (DS), and any other note types. Other note types included progress notes, hospital courses, history and physicals, nursing notes, procedures, operative notes, consults, emergency department documentation, discharge instructions, outpatient clinic notes, and telephone encounters. We excluded visits with less than 300 words, as these were frequently administrative notes with minimal medical information.

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 (bigrams) or 3 (trigrams) words) in the notes (see Supplemental File).

Clinical outcomes

Our outcome was discharge and post-discharge modified Rankin Scale. The gold standard scores were obtained from the local EHR-linked AHA GWTG Registry[8] and from clinic notes (if documented) using regexes (see Supplemental File). The mRS scores for the external validation cohort were obtained from a prospectively maintained RedCap registry. Notes were manually reviewed for expert validity by a neurologist/neurointensivist. In any case of discrepancy between the gold standard and clinical documentation, the notes were removed.

Modeling design

We selected each note type (PT/OT, DS, and other) from the calendar date closest to the time of gold standard mRS measurement. Models predicting discharge mRS used notes closest to the day of discharge, and models predicting post discharge mRS used notes documented closest to the day of post-discharge gold standard measurement. The data from our center was split into train (70%) and test (30%) sets, with unique patients in each set. With the train set we developed a logistic regression model for prediction of good (mRS 0–2) versus poor (mRS 3–6) mRS and a linear regression model for prediction of the full range of mRS 0–6.

Both models used the least absolute shrinkage and selection operator (LASSO) to select informative text-based features, age, sex, patient discharge status, and outpatient follow-up flag (yes/no) to predict the mRS scores. Age values were normalized using min–max normalization. For each model, we performed 100 iterations of fivefold cross validation in the training data to determine the best regularization parameter (see Supplemental Methods). The relative importance of the variables was assessed via their regression coefficients.

Our final model had three-stages: (stage 1) for patients with a discharge status of deceased we automatically assigned mRS 6 as the predicted score; (stage 2) for any encounter where mRS was documented by clinicians, regular expressions (see Supplemental Methods) were used to extract the score; (stage 3) for all other encounters (patients alive at discharge and those without mRS documentation) LASSO models were used for prediction.

We present performance of the three-stage model on the held-out test set from our center and on the external validation set. We also present results of the model without a staged approach that is, the LASSO model applied directly for prediction of mRS.

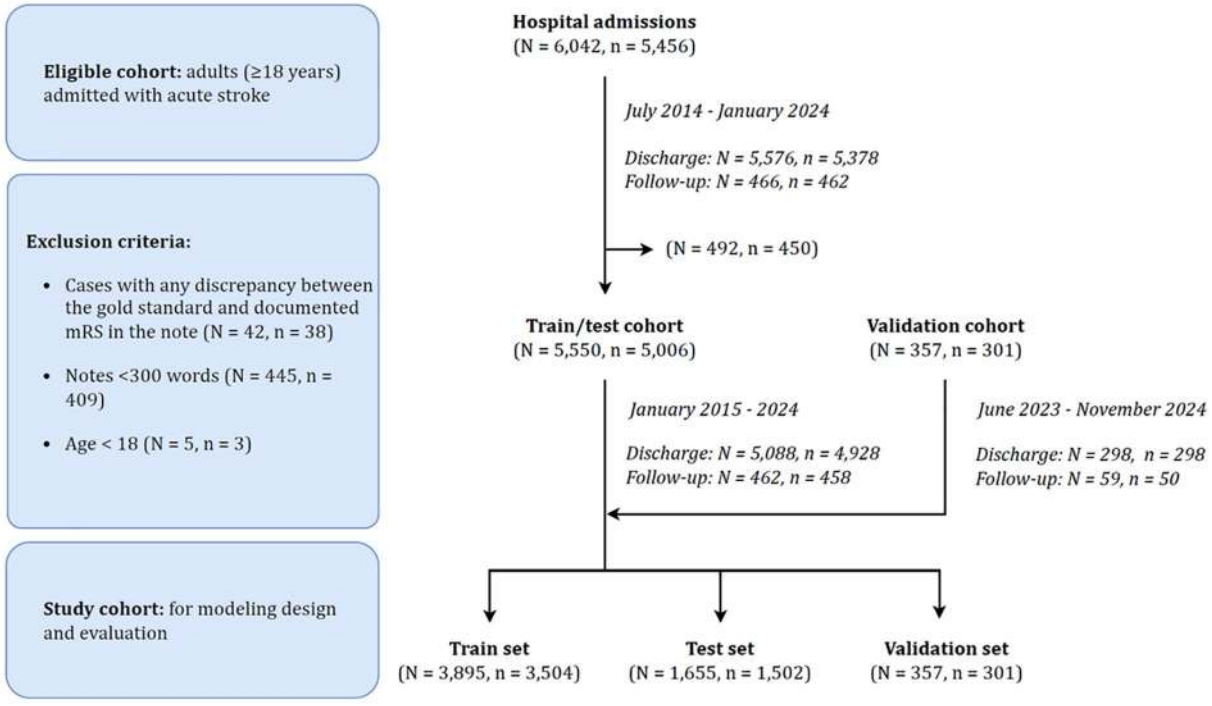

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) chart.
GWTG: Get With The Guidelines Stroke registry; mRS: modified Rankin Scale.
The number of patients is represented by "*n*" and the number of visits by "*N*."

Model performance metrics

The logistic regression model was evaluated using area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, sensitivity (or recall), and specificity. For classification (good vs poor mRS), we selected a threshold that converted the model probabilities into binary decisions, based on the training data, by maximization of the F1 score (harmonic mean of precision and recall; F1 = [precision × recall]/[precision + recall]), which is suited for imbalanced data.[10] We present the macro average performance for both poor and good functional outcomes. That is, we calculate each metric with good outcome as the target class, and with poor outcome as the target class, and compute the average. The linear regression model was evaluated using the root mean squared error (RMSE) and Spearman correlation.

We computed 95% confidence intervals using 1000 iterations of bootstrap random sampling with replacement. We also examined how model performance changed as time from calendar date of gold standard measurement and calendar date of availability of notes increased.

Results

Patients characteristics

Our study cohort (train and test) included 5006 patients with 5550 visits at our center (Figure 1). Most patients were White (74%), older adults with average age 69 years (Table 1). The majority had ischemic stroke or transient ischemic attack (98%). The most common discharge dispositions were acute care facility (46%) or home (44%). The validation cohort had similar demographic and disposition characteristics. The most common diagnosis in the validation cohort were brain injuries and seizures.

We present the distribution of availability of gold-standard mRS measurements at discharge and post-discharge for both cohorts (Figures S1 and S2). The most frequent discharge mRS was 4 (32%), whereas the post-discharge mRS scores were more evenly distributed, with mRS 0 (26%) being the most frequent.

Model performance

The logistic regression model performance is shown in Table 2. Results for the three-stage model are similar to those of the LASSO model applied to all visits. For the remaining analyses we present results of the three-stage model. The AUROC and AUPRC curves are presented in Figure S3 and the confusion matrices in Figure 2. Results for the linear model are also presented in Figure 2 (with additional details in Table S2 and Figure S4). The three-stage linear model achieved RMSE (95% CI) of 0.91 [0.87, 0.94] and 1.17 [1.06, 1.28] and Spearman correlation between predicted and gold standard mRS scores of 0.85

Table 1. Characteristics of the study population.

Characteristic	Train/Test dataset ($n = 5006$)	Validation dataset ($n = 301$)
Age (years), mean (SD)	68.5 (15.0)	64.8 (16.8)
Male, *n* (%)	2678 (53.5)	159 (52.8)
Race, *n* (%)		
White	3707 (74.0)	141 (46.8)
Black or African American	381 (7.6)	53 (17.6)
Asian	238 (4.8)	11 (3.7)
Other[a]	680 (13.6)	96 (31.9)
Hispanic or Latino, *n* (%)	352 (7.0)	6 (2.0)
Hospital admissions, *N*	5088	298
Discharge disposition, *N* (%)		
Acute care facility[b]	2335 (45.9)	116 (38.9)
Home[c]	2238 (44.0)	100 (33.6)
Expired	475 (9.3)	73 (24.5)
Other	40 (0.8)	9 (3.0)
Admission diagnosis, *N* (%)		
Ischemic	4713 (92.6)	44 (14.8)
ICH	276 (5.4)	24 (8.0)
SAH	48 (0.9)	19 (6.4)
Unspecified stroke	51 (1.1)	—
Traumatic brain injury	—	33 (11.1)
Seizures and status epilepticus	—	54 (18.1)
Other[d]	—	124 (41.6)
Follow-up visits, *N*	462	59
Follow-up (days), median [IQR]	94 [83, 100]	154 [96.5, 183.5]

ICH: intracerebral hemorrhage; IQR: interquartile range [25th, 75th] percentiles; SAH: subarachnoid hemorrhage.
The number of patients is represented by *n* and the number of visits is represented by *N*.
[a]Includes American Indian or Alaska Native and unknown race.
[b]Includes skilled nursing facility, rehabilitation facility, hospice medical facility.
[c]Includes home self-care and home-health services.
[d]Includes anoxic brain injury, brain tumors, neuroinfectious diseases, toxic metabolic encephalopathy, and non-specified altered mental status.

Table 2. Macro average performance for prediction of binary modified Rankin Scale, in 95% confidence intervals.

Model type	AUROC [95% CI]	Accuracy [95% CI]	Recall, Specificity [95% CI]	AUPRC [95% CI]
Test				
Three-stage	0.94 [0.93, 0.95]	0.88 [0.86, 0.89]	0.87 [0.86, 0.89]	0.94 [0.93, 0.95]
LASSO[a]	0.94 [0.93, 0.95]	0.87 [0.86, 0.89]	0.87 [0.85, 0.89]	0.94 [0.93, 0.95]
Validation				
Three-stage	0.94 [0.91, 0.96]	0.89 [0.86, 0.92]	0.86 [0.82, 0.90]	0.96 [0.94, 0.97]
LASSO[a]	0.94 [0.90, 0.96]	0.89 [0.86, 0.92]	0.86 [0.82, 0.90]	0.95 [0.93, 0.97]

AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; CI: confidence intervals; mRS: modified Rankin Scale.
[a]LASSO applied directly for prediction in all visits without performing stage 1 and stage 2, that is, without utilization of discharge status or extraction of mRS from notes using regexes.

[0.84, 0.87] and 0.85 [0.80, 0.89] in test and validation, respectively.

When notes were available on the same day as the gold standard mRS measurement, the performance of both logistic regression and linear models was higher (Figure 3; Table S3 and Figure S4). The majority of patients had notes available on the same day as the mRS was measured.

We also assessed the availability of each note type (Table S4). For discharge visits notes were available on the day of discharge, and in the case of PT/OT notes, up to 2 days prior

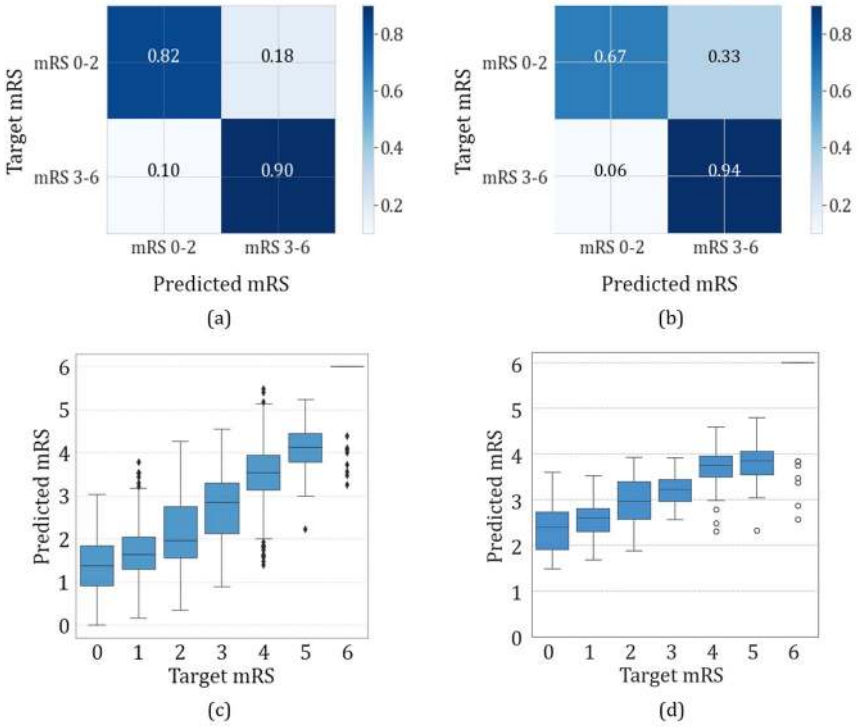

Figure 2. Three-stage logistic regression confusion matrix normalized by recall in (a) test and (b) validation and linear model modified Rankin Scale predicted scores versus target scores in (c) test and (d) validation.

to discharge. Post discharge visits had mostly other types of notes available on the follow-up day.

Feature importance

The training vocabulary consisted of 4475 combinations of unigrams, bigrams and trigrams, which was reduced to 374 features using LASSO regularization for binary classification and to 307 using linear modeling (Table S5). We present the relative importance of the top 20 binary modeling features in Figure 4.

For prediction of poor outcome (mRS 3–6), deceased status at discharge was the most important feature. Older age, "walk without assist" (can indicate that the patient is either able or unable to walk without assistance), discharge to rehabilitation, and "no movement" were also positively correlated with poor functional outcomes. For prediction of good (mRS 0–2) outcome, having a post-discharge follow up was an important feature, followed by discharge to home, "all activ" (can indicate that the patient is either able to perform all activities or may have a slight disability, and is unable to carry out all previous activities, but able to look after own affairs without assistance—mRS in the range 0–2) and independent ambulation ("ambul independ").

The importance estimates of the linear model features were in line with those found for the binary model (Figure S5). For prediction of the full range of mRS scores, deceased status at discharge was selected as the most important feature. Discharge to rehabilitation, requiring

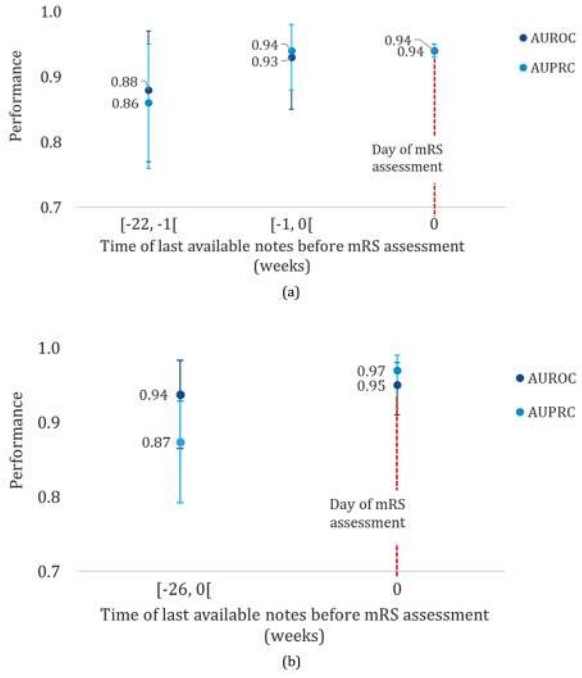

Figure 3. Three-stage model macro performance based on time interval between available note and gold standard mRS measurement in (a) test and (b) validation.
AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating curve.
X-axis legend: 0: Notes available same day as the gold standard mRS measurement (reference); [−1, 0[—notes available 1–7 days from the reference; [−22, −1[—notes available 1–22 weeks from the reference; [−26, 0[—notes available in the first week to 22 weeks from the reference.

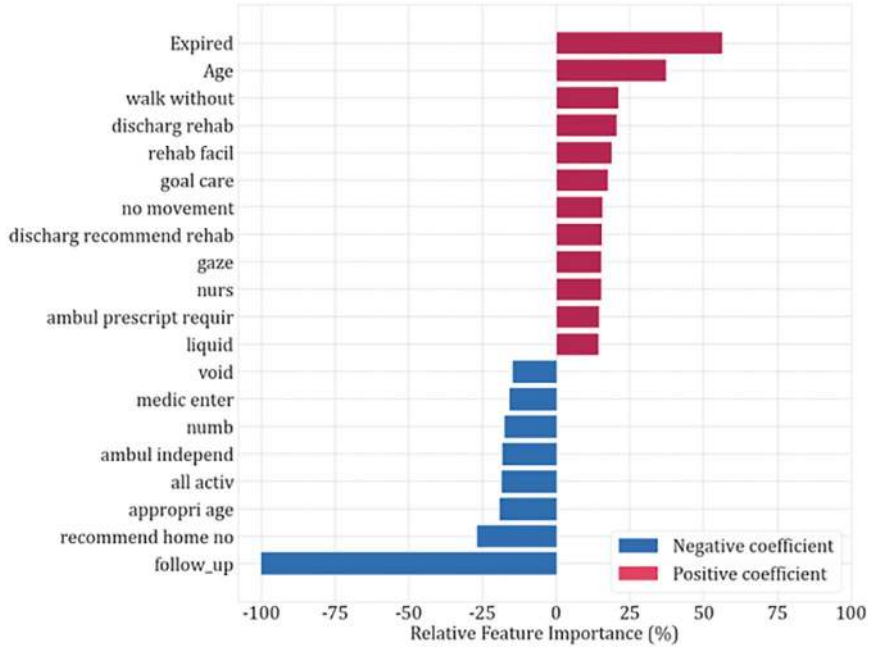

Figure 4. Top 20 features with highest relative importance estimates for the logistic regression model. Positive coefficient favors mRS 3–6.

Apart from "Expired," "Age" and "follow_up," the remaining text-based features are a result of stemming.[11] Expired: deceased at discharge; Age: age at admission; "walk without": (un)able to walk without (assistance); "discharg rehab": discharge to rehabilitation; "rehab facil": rehabilitation facility; "goal care": goals of care; "no movement": no movement/no response to stimulus; "discharg recommend rehab": recommendation of discharge to rehabilitation; "gaze": gaze (e.g., normal, right/left gaze preference); "nurs": nurse or nursing (e.g., skilled nursing facility, home nursing, nursing services/evaluation/notes); "ambul prescript requir": combination of "ambulance" and "prescriptions required" (e.g., patients traveling via ambulance with narcotic prescriptions - required on all patients being transferred to a skilled nursing facility); "liquid": liquid diet (e.g., associated with dysphasia); "void": related with urine output or renal function (patient voids (without difficulty) or renal function: voids); "medic enter": medications entered (e.g., home medications to lower risk of another stroke); "numb": numb(ness) (e.g., right hand numbness); "ambul independ": ambulates independently; "all activ": all activities (e.g., (un)able to use involved upper extremities for all activities, (un)able to carry out all activities); "appropri age": appropriate for age (e.g., speech is clear, appropriate for age); "recommend home no": recommendation to discharge home, no occupational therapy or home-health services needed; "follow_up": follow-up patient visit.

follow-up services and "no movement" were also associated with higher mRS scores. Having a post-discharge follow-up was the most important feature related to lower mRS scores, as well as independent ambulation and discharge to home.

Discussion

In this study, we developed a highly accurate NLP model that automates the extraction of mRS scores from unstructured electronic health records. This model represents an advance in stroke outcomes research with its ability to predict mRS scores across various post-stroke time points. By leveraging a diverse range of clinical notes, including physical and occupational therapy notes, discharge summaries, and other clinical documentation, our approach provides a more comprehensive and nuanced assessment of functional outcomes than previous methods. The high performance of our model, with an AUROC of 0.94 for binary classification and an RMSE of 0.91 for full range mRS prediction, surpasses existing tools. Our models enable automated, accurate, and scalable tracking of post-stroke functional outcomes, which is crucial for guiding interventions, informing health policies, and ultimately reducing the burden of stroke-related disability. We demonstrate the generalizability of our model in an external validation set from a different healthcare institution with a more heterogenous population.

Manual chart review for extraction of stroke outcomes is time consuming and resource intensive as it requires personnel for review and abstraction. With resource and personnel constraints, manual chart review is typically limited to small cohorts. The methods developed here have the potential to save time and resources. Our code for notes pre-processing and algorithm for extraction of mRS can directly be applied to new datasets that have a discharge summary, PT/OT assessments, or other documentation capturing discharge details and nursing assessments, as evidenced by our validation cohort.

Several prior studies have applied machine learning models for prediction of mRS scores. A few studies[12–15] used NLP for text representation, others used structured

data such as patient demographics, medical history, medications, laboratory findings, clinical scores, and other stroke-related information, including imaging features.[16–25] Some focused on predicting the full range of mRS scores 0–6 either using ordinal models, linear models or multiclassification, others focused on predicting poor functional outcome (mRS 3–6). However, prior work was either limited to prediction of mRS at a specific time point only or within specific disease subgroups, and had lower performance.

In one prior study,[12] NLP was applied to notes, however the distribution of scores was limited, and the model was only developed for prediction of mRS scores at discharge. In another study the authors developed NLP models using unstructured data from the history of presenting illness and computed tomography notes to predict 90-day mRS 3–6.[13] The authors found the NLP models performed well, however their performance metrics in both the internal and external validation data set were lower than ours (AUROC 0.82 and 0.79, respectively). This may be due to limitation of models to history and physical and radiology notes which have less information on post-stroke outcomes.[13]

Prior binary prediction models using NLP[14,15] had lower performance compared to our models (AUROCs ranging from 0.86 to 0.91[14] and from 0.78 to 0.80[15]). Performance of these models may be lower due to restriction to a smaller cohort of ICH patients,[14] and restriction to analysis of radiology reports.[15] Our study overcomes several limitations of prior work by including a larger cohort, both ischemic and hemorrhagic stroke, mRS at multiple time points, and a variety of unstructured note types.

Studies using structured data[16–23] to predict binary mRS (mRS >2[16–22] and mRS $\geqslant 5$[23]) had AUROCs ranging from 0.64 to 0.91, with majority of studies having lower performance than our study and other works using NLP. Reasons for the variable and lower performance include selection of variables only available at admission,[16] limitation to specific cohorts of patients such as those with stroke due to large vessel occlusion only[19] and smaller sample sizes.[19,20] Prior models using structured data to predict the full range of mRS scores[24,25] also had lower performance than our work, demonstrating the additional value of NLP and use of unstructured data.

One limitation of our study is that availability of notes close to the time of gold standard assessment impacts performance, and a small proportion of our patients had follow up visits limiting the notes available for model development. Despite this limitation, even for encounters with a larger delta between note availability and gold standard measurement, the model performed better than some prior NLP and structured data-based models. Our external validation dataset was small. However, the validation cohort included multiple diagnostic etiologies, in addition to stroke, demonstrating the generalizability of our model across healthcare systems and across disease diagnostic categories. With regards to note types, our model relies on discharge summaries and PT/OT assessments, or notes that have details on the discharge exam and patients' ability to perform activities of daily living independently versus with supervision. Therefore, these note types, or notes with appropriate clinical detail need to be available for application of the algorithm in other datasets. Finally, while mRS 0–1 is considered excellent outcome, due to the smaller proportion of mRS 0–1 at discharge, our binary model was designed to distinguish mRS 0–2 versus 3–6. However, to this goal, we have developed the linear model for predicting each level of mRS, which also had high performance. In our validation cohort we found that the linear performance for mRS 0–1 was lower (i.e., our model predicted higher mRS scores for these patients). On detailed evaluation of these errors, we found that majority of these patients were those admitted for seizures, and while they were independent and discharged home, the nursing documentation on the day of discharge included phrases such as "assistance from bed to chair", "supervision" due to the implementation of seizure safety precautions.

Conclusion

The models developed in this work automatically extract the mRS scores from electronic health records even if the scores are not documented. The models can enable real-world EHR stroke outcome phenotyping, and thereby may enable population research and quality improvement initiatives for stroke outcomes and disparities in stroke care. Future directions include applying and finetuning these models in additional EHRs, and using the outputs in comparative effectiveness research.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: 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 co-founder, 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.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was supported by NIH R01NS131347 (PI Sahar F. Zafar).

Ethical approval

The study was approved by the Mass General Brigham Institutional Review Board.

Informed consent

A waiver of informed consent was obtained for this observational study.

Guarantor

The first author accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish.

Contributorship

Conceptualization, Investigation, Methodology: M.F., A.B.S., S.F.Z.; Data curation, Software: M.F., K.G., N.T., A.G.; Formal analysis, Visualization: M.F.; Validation: M.F., S.F.Z.; Funding acquisition: S.F.Z.; Project administration, Supervision, Resources: S.F.Z.; Writing—original draft: M.F.; Writing—review & editing: M.B.W., A.B.S., and S.F.Z.

ORCID iDs

Marta Fernandes https://orcid.org/0000-0002-7203-2832
Kaileigh Gallagher https://orcid.org/0009-0005-2802-9790
Niels Turley https://orcid.org/0009-0009-4806-578X
Aditya Gupta https://orcid.org/0000-0002-5243-368X
Aneesh B Singhal https://orcid.org/0000-0002-2641-5277

Data availability

Code for notes preprocessing and modeling along with de-identified data are publicly available in a GitHub repository: https://github.com/mpriscila88/NLP_postStroke_mRS.

Supplemental material

Supplemental material for this article is available online.

References

1. Feigin VL, Brainin M, Norrving B, et al. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. *Int J Stroke Off J Int Stroke Soc* 2022; 17: 18–29.
2. Renedo D, Acosta JN, Leasure AC, et al. Burden of ischemic and hemorrhagic stroke across the US from 1990 to 2019. *JAMA Neurol* 2024; 81: 394–404.
3. Rajsic S, Gothe H, Borba HH, et al. Economic burden of stroke: a systematic review on post-stroke care. *Eur J Health Econ* 2019; 20: 107–134.
4. Kruse CS, Stein A, Thomas H, et al. The use of electronic health records to support population health: a systematic review of the literature. *J Med Syst* 2018; 42: 214.
5. Transmission Chapter TJC (v2023B) [Internet], https://manual.jointcommission.org/releases/TJC2023B/TransmissionChapterTJC.html (2023, accessed 17 June 2024)
6. Coleman CI, Concha M, Koch B, et al. Derivation and validation of a composite scoring system (SAVED2) for prediction of unfavorable modified Rankin scale score following intracerebral hemorrhage. *Front Neurol* 2023; 14: 1112723.
7. Quinn TJ, Ray G, Atula S, et al. Deriving Modified Rankin scores from medical case-records. *Stroke* 2008; 39: 3421–3423.
8. Reeves MJ, Smith EE, Fonarow GC, et al. Variation and trends in the documentation of National Institutes of Health Stroke Scale in GWTG-Stroke Hospitals. *Circ Cardiovasc Qual Outcomes* 2015; 8(6 Suppl 3): S90–S98.
9. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; 385: e078378.
10. Saito T and Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 2015; 10: e0118432.
11. Porter MF. An algorithm for suffix stripping. *Program* 1980; 14: 130–137.
12. Fernandes MB, Valizadeh N, Alabsi HS, et al. Classification of neurologic outcomes from medical notes using natural language processing. *Expert Syst Appl* 2023; 214: 119171.
13. Sung S, Chen C, Pan R, et al. Natural language processing enhances prediction of functional outcome after acute ischemic stroke. *J Am Heart Assoc* 2021; 10: e023486.
14. Hung LC, Su YY, Sun JM, et al. Clinical narratives as a predictor for prognosticating functional outcomes after intracerebral hemorrhage. *J Neurol Sci* 2023; 453: 120807.
15. Heo TS, Kim YS, Choi JM, et al. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. *J Pers Med* 2020; 10: 286.
16. Monteiro M, Fonseca AC, Freitas AT, et al. Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM Trans Comput Biol Bioinform* 2018; 15: 1953–1959.
17. Heo J, Yoon JG, Park H, et al. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke* 2019; 50: 1263–1265.
18. Herzog L, Kook L, Hamann J, et al. Deep learning versus neurologists: functional outcome prediction in LVO stroke patients undergoing mechanical thrombectomy. *Stroke* 2023; 54: 1761–1769.
19. Brugnara G, Neuberger U, Mahmutoglu MA, et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. *Stroke* 2020; 51: 3541–3551.
20. Xie Y, Jiang B, Gong E, et al. Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. *Am J Roentgenol* 2019; 212: 44–51.
21. Alaka SA, Menon BK, Brobbey A, et al. Functional outcome prediction in ischemic stroke: a comparison of machine learning algorithms and regression models. *Front Neurol* 2020; 11: 889.
22. van Os HJA, Ramos LA, Hilbert A, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. *Front Neurol* 2018; 9: 784.
23. Ramos LA, Kappelhof M, van Os HJA, et al. Predicting poor outcome before endovascular treatment in patients with acute ischemic stroke. *Front Neurol* 2020; 11: 580957.
24. Zhang MY, Mlynash M, Sainani KL, et al. Ordinal prediction model of 90-day modified rankin scale in ischemic stroke. *Front Neurol* 2021; 12: 727171.
25. Asadi H, Dowling R, Yan B, et al. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS ONE* 2014; 9: e88225.