Epilepsy & Behavior 157 (2024) 109835

Contents lists available at ScienceDirect

Epilepsy & Behavior

journal homepage: www.elsevier.com/locate/yebeh

Research Paper

Can machine learning predict late seizures after intracerebral hemorrhages? Evidence from real-world data

Alain Lekoubou [a,*], Justin Petucci [b,c], Temitope Femi Ajala [d], Avnish Katoch [c], Jinpyo Hong [e], Souvik Sen [f], Leonardo Bonilha [f], Vernon M. Chinchilli [g], Vasant Honavar [b,c,h,i,j]

[a] Department of Neurology, Milton S. Hershey Medical Center and Department of Public Health, Pennsylvania State University, USA
[b] Institute for Computational and Data Sciences, USA
[c] Clinical and Translational Sciences Institute, USA
[d] Alabama Department of Public Health, USA
[e] College of Medicine, Penn State University, Hershey, PA, USA
[f] University of South Carolina, Department of Neurology, USA
[g] Department of Public Health Sciences, Pennsylvania State University, USA
[h] Data Sciences Program, USA
[i] College of Information Sciences and Technology, USA
[j] Center for Artificial Intelligence Foundations and Scientific Applications, USA

ARTICLE INFO

Keywords:
Machine learning
Late seizures
Prediction
TriNetX
Intracerebral hemorrhage

ABSTRACT

Introduction: Intracerebral hemorrhage represents 15 % of all strokes and it is associated with a high risk of post-stroke epilepsy. However, there are no reliable methods to accurately predict those at higher risk for developing seizures despite their importance in planning treatments, allocating resources, and advancing post-stroke seizure research. Existing risk models have limitations and have not taken advantage of readily available real-world data and artificial intelligence. This study aims to evaluate the performance of Machine-learning-based models to predict post-stroke seizures at 1 year and 5 years after an intracerebral hemorrhage in unselected patients across multiple healthcare organizations.

Design/methods: We identified patients with intracerebral hemorrhage (ICH) without a prior diagnosis of seizures from 2015 until inception (11/01/22) in the TriNetX Diamond Network, using the International Classification of Diseases, Tenth Revision (ICD-10) I61 (I61.0, I61.1, I61.2, I61.3, I61.4, I61.5, I61.6, I61.8, and I61.9). The outcome of interest was any ICD-10 diagnosis of seizures (G40/G41) at 1 year and 5 years following the first occurrence of the diagnosis of intracerebral hemorrhage. We applied a conventional logistic regression and a Light Gradient Boosted Machine (LGBM) algorithm, and the performance of the model was assessed using the area under the receiver operating characteristics (AUROC), the area under the precision-recall curve (AUPRC), the F1 statistic, model accuracy, balanced-accuracy, precision, and recall, with and without seizure medication use in the models.

Results: A total of 85,679 patients had an ICD-10 code of intracerebral hemorrhage and no prior diagnosis of seizures, constituting our study cohort. Seizures were present in 4.57 % and 6.27 % of patients within 1 and 5 years after ICH, respectively. At 1-year, the AUROC, AUPRC, F1 statistic, accuracy, balanced-accuracy, precision, and recall were respectively 0.7051 (standard error: 0.0132), 0.1143 (0.0068), 0.1479 (0.0055), 0.6708 (0.0076), 0.6491 (0.0114), 0.0839 (0.0032), and 0.6253 (0.0216). Corresponding metrics at 5 years were 0.694 (0.009), 0.1431 (0.0039), 0.1859 (0.0064), 0.6603 (0.0059), 0.6408 (0.0119), 0.1094 (0.0037) and 0.6186 (0.0264). These numerical values indicate that the statistical models fit the data very well.

Conclusion: Machine learning models applied to electronic health records can improve the prediction of post-hemorrhagic stroke epilepsy, presenting a real opportunity to incorporate risk assessments into clinical decision-making in post-stroke care clinical care and improve patients' selection for post-stroke epilepsy research.

* Corresponding author at: Department of Neurology, Penn State University, Hershey Medical Center, Hershey, PA, USA.
E-mail addresses: alekouboulooti@pennstatehealth.psu.edu (A. Lekoubou), jmp579@psu.edu (J. Petucci), akatoch@pennstatehealth.psu.edu (A. Katoch), jhong3@pennstatehealth.psu.edu (J. Hong), souvik.sen@uscmed.sc.edu (S. Sen), leonardo.bonilha@uscmed.sc.edu (L. Bonilha), vchinchilli@pennstatehealth.psu.edu (V.M. Chinchilli), vuh14@psu.edu (V. Honavar).

https://doi.org/10.1016/j.yebeh.2024.109835
Received 26 February 2024; Received in revised form 8 May 2024; Accepted 8 May 2024
Available online 30 May 2024

1. Introduction

Intracerebral hemorrhage is the most devastating and least treatable form of stroke, affecting one in six stroke patients [1,2]. It results in severe disability or death in nearly 60 % of patients [2]. Long-term effects of stroke are frequent. Late seizures, i.e., post-stroke epilepsy, are frequent complications of intracerebral hemorrhage affecting up to 11 % of patients after a mean follow-up of 9 months. Studies suggest an independent association between late seizures and increased mortality and poor functional outcome [3,4]. Late seizures are also associated with worse cognitive outcomes and dementia [5]. Adequately managing seizures after intracerebral hemorrhage could potentially avoid complications and improve the quality of life of survivors of intracerebral hemorrhage. Identifying patients at risk of late seizures is an essential step to improve outcomes after intracerebral hemorrhage by targeting seizures. Independent predictors of late seizures include the involvement of the cortical regions, hematoma volume, intraventricular extension, stroke severity, younger age, and early seizures. Age, stroke severity [1], and atrial fibrillation [2] have also been identified as risk factors for seizures after stroke. Combining these individual risk factors into predictive models is more likely to predict the individual risk of developing late seizures after intracerebral hemorrhage than considering each factor individually. Risk scores to predict late seizures after intracerebral hemorrhage have been developed, including the CAVE score [3], the CAVS score [4], and the LANE score [5]. They had an average to good performance with an area under the receiver operating curve/c-statistics ranging from 0.69 to 0.83. However, the scores used clinical and imaging variables from selected patients in specialized units. Real-world data are increasingly available with hundreds of thousands of patients' data collected across healthcare organizations. These readily available data could be used for model development. As the data are readily available, models could be incorporated into electronic health records and provide real-time individual risk for predefined outcomes. Large electronic health records have been seldom used to predict seizures after intracerebral hemorrhage. Using powerful computational methods is more likely to handle large volume records than traditional logistic regression models alone. In this study, we applied machine learning to predict late seizures taking advantage of TriNetX Diamond Network, a large network of 71 healthcare organizations collecting data of nearly 106 million patients. We hypothesized applying machine learning to this large network, we could develop models to predict late seizures with good performance in an unselected heterogeneous population of patients with intracerebral hemorrhage.

2. Methods

2.1. Study design and data source

This was a retrospective cohort analysis using data obtained from TriNetX Research Network, a network of 71 Healthcare organization electronic health records comprising data of 106 million patients (September 2022).

2.2. Study population

Our study population included adult patients (age ≥18 years) with intracerebral hemorrhage, identified using the *International Classification of Diseases, tenth Revision* (*ICD-10*) I61, from January 1, 2015, through August 9, 2022. We excluded all participants with a diagnosis of seizures before the stroke, identified using any of the ICD-10 codes G40 and G41. A total of 85,679 had an ICD-10 code of intracerebral hemorrhage and no prior diagnosis of seizures, constituting our study cohort.

Assessment of outcome: The time at risk was 1-year and 5-year after the index stroke event. Seizures were identified using any of the ICD-10 codes G40 and G41. These codes are specific for epilepsy (late seizures) unlike the code R56, which is a nonspecific ICD-10 code for unspecified convulsions [6].

Covariates: Demographic variables included age (continuous variable), sex assigned at birth (male vs. female), and race/ethnicity. Race and ethnicity were grouped into four categories: Non-Hispanic White (NHW), Non-Hispanic Black (NHB), Hispanic, and others. Clinical variables included the following: diagnosis or history of hypertension, diagnosis or history of diabetes mellitus, diagnosis or history of atrial fibrillation, history of smoking, history of alcohol use, and stroke severity (assessed using a combination of factors and variables described in the appendix). Anti-seizure drugs use was identified using RxNorm, a unified medical language system developed by the National Library of Medicine that provided normalized names for clinical drugs [7]. Patients with traumatic brain injury, benign brain neoplasms, malignant brain neoplasms, unspecified brain neoplasms, severe intracranial infection, bacterial meningitis, encephalitis, and those who had decompressive craniotomy were excluded. We used ICD-10 and CPT codes to identify these variables (see supplemental materials).

Machine learning model methods: We applied a 5-fold nested cross-validation (CV) with non-overlapping training set (for training the model) and validation set (for hyperparameter tuning) and test set (for model evaluation). This approach was important for developing a generalizable model. First, we stratified the dataset into 5 disjoint subsets, or folds. Second, we iterated each fold over, serving once as the test set (red) while the remaining folds comprised the training set (blue). Third, within this training set, we conducted an inner cross-validation by dividing it into 5 further folds. In this crucial step for hyperparameter optimization, each parameter combination was trained on 4 folds (gray) and validated on the remaining fold (green), cycling through all 5 folds to determine the best-performing hyperparameters. Fourth, we used these optimal parameters to train a new model on the full training set of the outer loop. Fifth, we assessed the model's predictive performance on the outer test set, ensuring each data point was used for testing just once. Finally, after completing all 5 iterations, the performance metrics across all 5 outer test sets were aggregated to produce a comprehensive evaluation of the model's generalization capability (Fig. 1). Classification models we explored included the following: logistic regression, decision tree, random forest, LightGBM, AdaBoost, support vector machine, k-nearest neighbors, discriminant analysis, and Gaussian naïve Bayes [8–14]. We used Scikit-learn to train and evaluate all models [15], except LightGBM, where Microsoft's LightGBM library was employed[16] . We obtained the best generalized performance results using the LightGBM model. In LightGBM, the hyperparameter optimization consisted of a grid search (within the nested cross-validation) over the tree depth, learning rate, and ensemble size. We used a cost-sensitive learning approach to account for the class balance between patients who developed seizures and those who did not. In the cost-sensitive approach, the objective/cost function was modified to yield a stronger penalty for incorrectly predicting the minority class, i.e., those who developed seizures (by an amount proportional to the imbalance) using LightGBM's *'class_weight='balanced'* option. We used LightGBM and a set of feature importance scores derived from trained LightGBM models, and Shapley values to determine the features most important in predicting seizures [17]. We used Shapley values and partial dependence plots (PDP) to investigate the relationship between predictors and seizures. PDPs show only the average effect of the input variable, hence neglecting the impact of feature interactions, which can be present with tree-based models such as LightGBM.

Model performance was evaluated using the following metrics: area under the receiver operating characteristics (AUROC), the area under the precision-recall curve (AUPRC), the F1 statistic, model accuracy, balanced accuracy, precision, and recall. Model performance was assessed separately, including then excluding patients on anti-seizure drugs.

Standard protocol approvals, registrations, and patient consent: This study protocol was submitted to the Pennsylvania State College of Medicine institutional review board and was not considered human

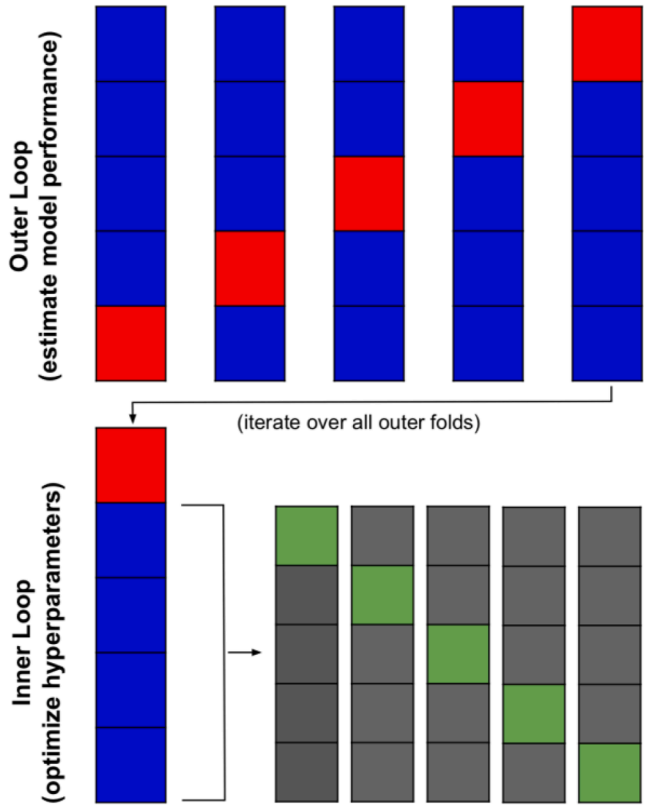

Fig. 1. 5- fold Nested Cross Validation. A visual representation of the nested 5-fold cross validation procedure. The outer loop partitions the data into 5 folds, where each fold serves as a unique and non-overlapping test set (red) once, while the remaining data forms the training set (blue). Within each outer training set, an inner 5-fold cross-validation is conducted, further dividing the data into 5 new folds. In this inner loop, one-fold is used as the validation set (green) for hyperparameter tuning each iteration, and the other folds act as the training set (gray).

subject research. All records contained within the database were fully de-identified. Thus, informed consent was waived.

Data availability: We used data from The TriNetX Research Network: health care organizations (de-identified claims data), 106 million patients, which are available to researchers from participating centers.

3. Results

A total of 85,679 patients had an ICD-10 codes of intracerebral hemorrhage and no prior diagnosis of seizures, constituting our study cohort. Seizures were present in 4.57 % (3915 patients) and 6.27 % (5372 patients) of patients within 1 and 5 years after ICH, respectively (Table 1). Patients who developed seizures were younger than those who did not at 1 year (52.4 ± 23.2 years vs. 57.4 ± 22.3 years, p-value < 0.001) and 5 years (51.2 ± 23.6 years vs. 57.4 ± 22.3, *p-value < 0.001)*. Sex distribution was similar in those who developed seizure and those who did not develop seizures at 1 year; however more male than female participants developed seizures at 5 years. Black individuals, smokers, and those who have a history of alcohol use were more likely to develop seizures than their counterparts at 1 and 5 years after ICH, while atrial fibrillation was more frequent among those who developed seizures than those who did not at 5-year post ICH ($p < 0.001$) only. Levetiracetam was the most frequently prescribed antiseizure medication both 1- and 5-years post-ICH. Patients who develop seizures were also more likely to be on electroencephalogram for both post-ICH time frames of seizure development.

Table 1
Demographic and Clinical Characteristics.

Characteristics	Total (%) 85,679	*Seizure Incidence following intracerebral hemorrhage after 1 year.* 3,915 (4.57 %)	*Seizure Incidence following intracerebral hemorrhage after 5 years.* 5,372 (6.27 %)
Age			
Mean ± (SD), [a]	57.2 ± 22.3	52.4 ± 23.2	51.2 ± 23.6
Median	66.2	57.9	57.9
Sex			
Male	48,751 (56.9)	2,268 (57.9)	3,133 (58.2)
Female	36,928 (47.5)	1,647 (42.1)	2,239 (41.7)
Race[a]			
White	52,462 (61.2)	2,327 (59.4)	3,179 (59.2)
Black/African American	13,644 (15.9)	737 (18.8)	1,038 (19.3)
Others or Unknown	19,573 (22.8)	851 (21.7)	1,155 (21.5)
Stroke Risk			
Smoking [a]	21,150 (24.7)	1,079 (27.6)	1,458 (27.1)
Hypertension [a]	46,674 (54.5)	2,178 (55.6)	2,901 (54.0)
Diabetes	17,308 (27.9)	802 (20.5)	1,037 (19.3)
Alcohol Use [a]	7,954 (4.7)	462 (11.8)	645 (12.0)
Atrial fibrillation [a]	12,105 (14.1)	530 (13.5)	670 (12.5)
Hyperlipidemia	24,297 (28.3)	1,152 (29.4)	1,493 (27.8)
ICH Location[a]			
Hemisphere, subcortical	10,929 (12.8)	500 (12.8)	688 (12.8)
Hemisphere, cortical	10,207 (11.9)	844 (21.6)	1,066 (1.8)
Hemisphere, unspecified	2,161 (2.5)	147 (3.8)	190 (10.0)
Brainstem	2,614 (3.1)	54 (1.4)	74 (1.4)
Cerebellum	4,392 (5.1)	108 (2.8)	156 (2.9)
Intraventricular	14,102 (16.5)	660 (16.9)	931 (17.3)
Multiple, localized	874 (1.0)	51 (1.3)	68 (1.3)
Unspecified	34,195 (39.9)	1,193 (30.5)	1,733 (32.3)
Antiplatelet Therapy			
Aspirin	19,169 (22.4)	863 (22.0)	1,168 (21.7)
Clopidogrel [a]	5,038 (13.6)	205 (5.2)	277 (5.2)
Ticagrelor	620 (0.7)	24 (0.6)	30(0.6)
Prasugrel	152 (0.2)	3 (0.1)	6 (0.2)
Electroencephalograms [a]			
Continuous EEG 2–12	378 (0.4)	40 (1.0)	40 (0.7)
Continuous EEG 2–26	662 (0.2)	72 (1.8)	79 (1.5)

(*continued on next page*)

Table 1 (*continued*)

Characteristics	Total (%) 85,679	*Seizure Incidence following intracerebral hemorrhage after 1 year.* 3,915 (4.57 %)	*Seizure Incidence following intracerebral hemorrhage after 5 years.* 5,372 (6.27 %)
Large artery Atherosclerotic Disease	6,338 (12.5)	316 (8.1)	427 (7.9)
Intravenous Tissue plasminogen activator	1,862 (2.2)	73 (1.9)	100 (1.9)
Stroke severity			
Aphasia dysphagia	12,247 (9.2)	860 (22.0)	1,121 (20.9)
Unspecific side hemiplegia	1,926 (1.2)	141 (3.6)	172 (3.2)
Non-dominant side hemiplegia	6,383 (7.4)	393 (10.0)	512 (9.5)
Aphasia	8,874 (10.4)	668 (17.1)	846 (15.7)
Aphasia-Cerebrovascular	1,236 (1.4)	95 (2.4)	125 (2.3)
Dysarthria	7,152 (8.3)	402 (10.3)	519 (9.7)
Speech Disturbance	5,062 (5.9)	301 (7.7)	386 (7.2)
Facial weakness	9,001 (10.5)	501 (12.8)	654 (12.2)
Cerebral Intracranial edema	1,604 (1.9)	127 (7.7)	166 (3.1)
Dominant left side hemiplegia	5,779 (6.7)	377 (9.6)	483 (9.0)
Altered mental status	15,139 (17.7)	1017 (26.0)	1,288 (24.0)
Neurological neglect syndrome	2,625 (3.1)	168 (4.3)	210 (3.9)
Respiratory failure	14,672 (17.1)	962 (24.6)	1,193 (22.2)
DNR	8,608 (10.0)	187 (4.8)	221 (4.1)
Palliative care consult	7,240 (8.5)	201 (5.1)	224 (4.2)
Mechanical thrombectomy	975 (1.1)	34 (0.9)	51 (0.9)
CE 1	9,452 (11.0)	557 (14.7)	757 (14.1)
CE 2	20,442 (23.9)	1,120 (28.6)	1,457 (27.1)
CE 3	3675 (4.3)	207 (5.3)	278 (5.2)
Respiratory ventilation procedure	11,848 (13.8)	659 (16.8)	842 (15.7)
Insertion of endotracheal tube	6,136 (7.2)	359 (9.2)	459 (8.5)
Insertion of feeding device	1,052 (1.2)	97 (2.5)	135 (2.5)
Intravenous thrombosis	1,862 (2.2)	73 (0.1)	100 (1.9)
Hemicraniectomy	38 (0.0)	0.5(0.1)	5 (0.1)
Antiseizure drug [a]			
Carbamazepine	204 (0.2)	18 (0.5)	24 (0.4)
Clobazam	17 (0.0)	5 (0.1)	5 (0.1)
Clonazepam	1415 (1.7)	82 (2.1)	115 (2.1)
Gabapentin	7303 (8.5)	367 (9.4)	499 (9.3)
Lacosamide	418 (0.5)	68 (1.7)	79 (1.5)
Lamotrigine	430 (0.5)	23 (0.6)	37 (0.7)
Levetiracetam	20,816 (24.3)	1674 (42.8)	2127 (39.6)
Oxcarbazepine	176 (0.2)	21 (0.5)	24 (0.4)
Perampanel	8 (0.0)	1 (0.0)	1 (0.0)
Phenobarbital	747 (0.9)	112 (2.9)	146 (2.7)
Phenytoin	882 (1.0)	72 (1.8)	101 (1.9)
Pregabalin	1135 (1.3)	48 (1.2)	70 (1.3)
Primidone	169 (0.2)	6 (0.2)	10 (0.2)
Topiramate	712 (0.8)	45 (1.1)	68 (1.3)
Valproate	882 (1.0)	86 (2.2)	107 (2.0)
Vigabatrin	2 (0.0)	1 (0.0)	1 (0.0)
Zonisamide	73 (0.1)	10 (0.3)	15 (0.3)

Foot Note:
SS_CE1(Stroke Severity Clinical Encounter 1): Describes a detailed interval history; A detailed examination; Medical decision making of high complexity.
SS_CE2 (Stroke Severity Clinical Encounter 2): Critical care, evaluation and management of the critically ill or critically injured patient; first 30–74 min.
SS_aphasia: Stroke Severity Aphasia.
ICH: Denotes patient with subsequent intracerebral hemorrhage ICD-10 codes.
AS: Antiseizure drugs.
[a] These characteristics are significantly associated with seizure at both one-year and five years after a stroke. *(p-value < 0.05)*.

Seven metrics were deployed to assess the model performances using LGBM algorithm to predict the risk of seizures at 1 year and 5 years, including the area under the receiver operating characteristics (AUROC), the area under the precision-recall curve (AUPRC), the F1 statistic, model accuracy, balanced-accuracy, precision, and recall, with and without seizure medication use in the models to allow an independent interpretation by the reader. These metrics were used simultaneously to account for the importance of classifying seizures and non-seizure patients, and the heavily imbalance sample. At 1-year, the AUROC, AUPRC, F1 statistic, accuracy, balanced-accuracy, precision, and recall were respectively 0.7051 (standard error: 0.0132), 0.1143 (0.0068), 0.1479 (0.0055), 0.6708 (0.0076), 0.6491 (0.0114), 0.0839 (0.0032), and 0.6253 (0.0216). Corresponding metrics at 5 years were 0.694 (0.009), 0.1431 (0.0039), 0.1859 (0.0064), 0.6603 (0.0059), 0.6408 (0.0119), 0.1094 (0.0037) and 0.6186 (0.0264), respectively (Table 2 and Figure 2). Out of 15 important features identified for LGBM model, age was identified as the most important feature in seizure risk prediction, while DNR, altered mental status, and aphasia followed in terms of subsequent features of importance (Figs. 3 and 4).

4. Discussion

In this retrospective analysis of nearly 90,000 patients with intracerebral hemorrhage from 71 healthcare organizations, the performance of machine learning models to predict seizures at 1 year and 5 years was good.

Several models have been developed to predict seizures after intracerebral hemorrhage. Arguably, the most widely used is the CAVE score, which combined four variables (age, hematoma volume, cortical involvement, and early seizures) to predict seizures at 1 year and 5 years. The model, which was developed in Finland had a good performance in the development and validation cohort with c-statistic ranging

Table 2
Model performances for all patients.

Metric (standard error)	*LGBM Update All features at 1 year*	*LGBM Update All features at 5 years*
AUROC	0.7051 (0.0132)	0.694 (0.009)
AUPRC	0.1143 (0.0068)	0.1431 (0.0039)
F1	0.1479 (0.0055)	0.1859 (0.0064)
Accuracy	0.6708 (0.0076)	0.6603 (0.0059)
Balance-Acc	0.6491 (0.0114)	0.6408 (0.0119)
Precision	0.0839 (0.0032)	0.1094 (0.0037)
Recall	0.6253 (0.0216)	0.6186 (0.0264)

from 0.69 to 0.81. Various models have been developed to predict late seizures after intracerebral hemorrhage in different populations. For instance, the CAVS model was developed using granular clinical data from a diverse US population, and the LANE model was developed specifically for Chinese patients. Both models have shown good performance similar the CAVE model. These findings demonstrate that risk models can predict late seizures using clinical data from selected population of patients with intracerebral hemorrhage.

Our study contributes to the prediction of late seizures after intracerebral hemorrhage. Our model's performance is comparable to previously developed models. Our study has four originalities. First, we did not use granular clinical data but relied on demographic variables and administrative codes to identify features that could predict late seizures after intracerebral hemorrhage. We confirmed that young age, cortical location, and surrogate of stroke severity such as the present of aphasia and altered mental status were important independent features contributing to late seizure prediction after intracerebral hemorrhage. Our study therefore provide evidence for the first time that data collected during routine clinical activity and available in electronic health record could be leveraged to predict late seizures after intracerebral hemorrhage. Second, unlike previous models based on relatively small cohorts of patients from specialized stroke units, we took advantage of a large network of shared data across several organizations in the United States, hence representing a heterogeneous population of stroke patients. The implication of including a heterogeneous population of patients with intracerebral hemorrhage from unselected healthcare organization and clinical settings across the United States is the enhanced generalizability of our model. Besides, the dataset was very large allowing for the identification of important predictors that could have been overlooked with smaller and selected patient populations. We were able to identify several important features contributing to the model, including some not previously reported. For example, 17.5 % of the risk of late seizures was explained by the presence of a do-not-resuscitate order; patients who had a do-not-resuscitate order were less likely to have seizures, suggesting that those patients could have died before developing seizures or that resources utilized to identified seizures such as electroencephalograms were sparely used when a do-not-resuscitate order was present. Third, we used machine learning for the purpose of predicting late seizures after intracerebral hemorrhage. Machine learning has been used in neurology to predict various outcomes [18–21]. With regards to seizure prediction, one study developed machine learning models to predict early seizures after intracerebral hemorrhage. Early seizures have different underlying pathophysiologic mechanisms and courses than late seizures. Early seizures are thought to result from transient cellular biochemical dysfunctions and have a 10-year risk of seizure recurrence of approximately 20 % whereas late seizures result from gliotic scaring and persistent neuronal excitability changes with a 10-year recurrence of 60 % [22,23], hence meeting the

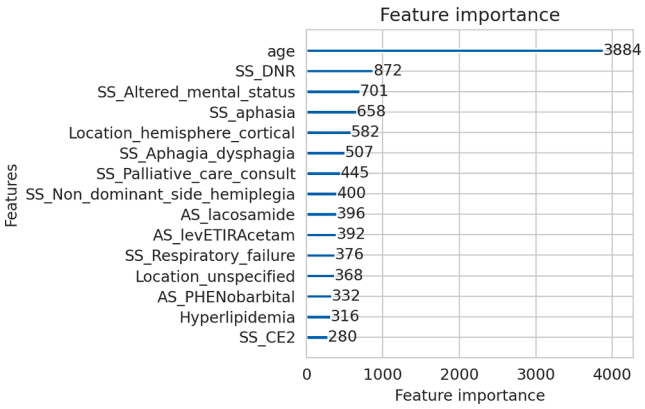

Fig. 3. Top 15 important features for LGBM model. Foot Note: SS_CE1 (Stroke Severity Clinical Encounter 1): Describes a detailed interval history; A detailed examination; Medical decision making of high complexity. SS_CE2 (Stroke Severity Clinical Encounter 2): Critical care, evaluation and management of the critically ill or critically injured patient; first 30–74 min. SS_aphasia: Stroke Severity Aphasia. ICH: Denotes patient with subsequent intracerebral hemorrhage ICD-10 codes. AS: Antiseizure drugs.

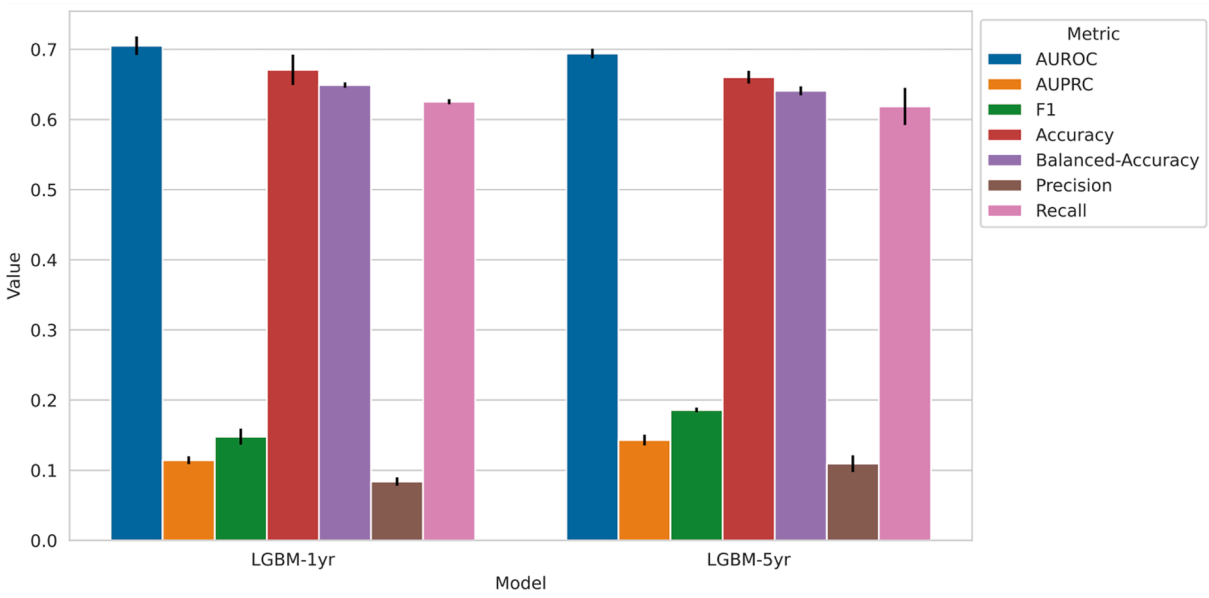

Fig. 2. Visual Model Performance.

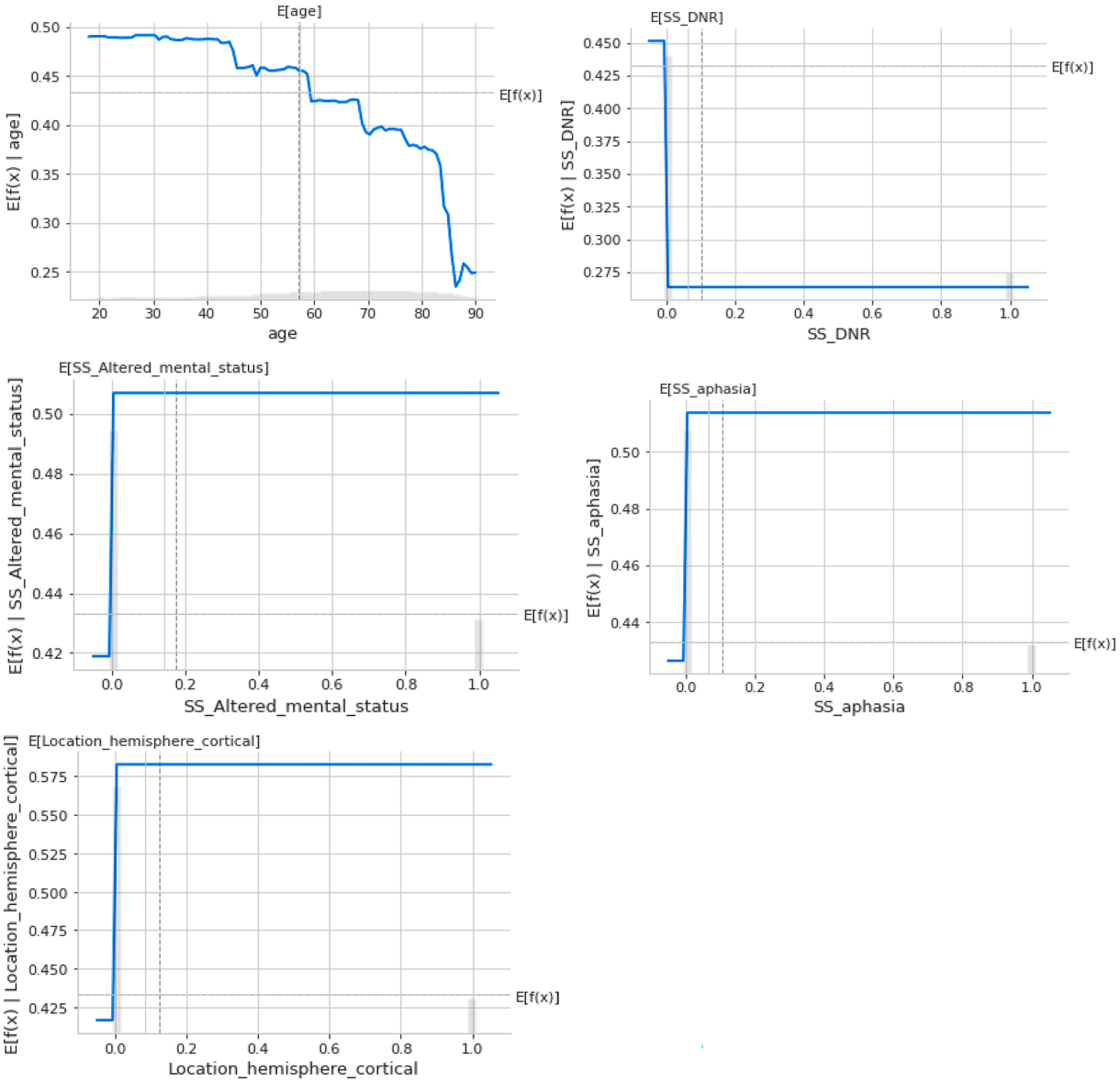

Fig. 4. Partial Dependent Plot for the five most important features.

new ILAE definition of epilepsy [24]. We are not aware of any previous use of machine learning models to predict late seizures after intracerebral hemorrhage. Ultimately, because the variables used in our model were readily available in electronic data across several institutions, the model has the potential to be incorporated into electronic health records and provide an instantaneous individual patient's risk of developing late seizures without interfering with patients' care, which could facilitate discussions between providers and patients/caregivers regarding their risk of late seizures. It is also possible that the identification of patients for potential inclusion in clinical trials based on their risk of late seizures could be facilitated.

5. Limitations

We used administrative ICD-10 diagnoses and procedures; therefore, we could not verify the accuracy of reporting these diagnoses in TriNetX. Despite relying on administrative ICD-10 code diagnoses and procedures, the performance of our model, i.e. AUC was similar to studies that did not rely on these codes. Although stroke severity was not assessed using standard severity scales such as the National Institute of Health Stroke Scale or Glasgow Coma Scale, all proxies of stroke severity used in the current analysis were associated with an increased risk of seizures, suggesting the validity of our approach. We did not have access to granular data such as brain imaging and electroencephalogram recording, which could have yielded additional predictors and improved the model's performance. Finally, machine learning models to predict late seizures in this study were not validated in external cohorts, i.e., non-US cohorts; however, we believe that such an external validation would not be necessary for two reasons: first, patients were recruited from 71 healthcare organizations across the US and included unselected patients, suggesting generalizability of our results. Second, we mitigated the need to externally validate the models by applying a 5-fold nested cross-validation (CV) with non-overlapping training set (for training the model) and validation set (for hyperparameter tuning) and test set (for model evaluation), which is important in generalizing machine learning models.

Despite these limitations, the use of large datasets from unselected patients across multiple healthcare organizations and the 5-fold nested cross-validation suggest that our model is generalizable to US patients with intracerebral hemorrhage. Our model could be easily integrated into electronic health records with little disruption of clinical flow in very busy hospital settings.

6. Conclusion

Electronic health records can be leveraged to predict late seizures after intracerebral hemorrhage, using machine learning. This could enhance clinical decision-making and prospective planning.

CRediT authorship contribution statement

Alain Lekoubou: Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Conceptualization. **Justin Petucci:** Writing – review & editing, Software, Methodology, Formal analysis. **Temitope Femi Ajala:** Writing – review & editing, Writing – original draft. **Avnish Katoch:** Writing – review & editing, Methodology, Data curation. **Jinpyo Hong:** Writing – review & editing, Writing – original draft. **Souvik Sen:** . **Leonardo Bonilha:** Writing – review & editing, Supervision. **Vernon M. Chinchilli:** Writing – review & editing, Supervision. **Vasant Honavar:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yebeh.2024.109835.

References

[1] Ferlazzo E, Gasparini S, Beghi E, Sueri C, Russo E, Leo A, et al. Epilepsy in cerebrovascular diseases: review of experimental and clinical data with meta-analysis of risk factors. Epilepsia 2016;57(8):1205–14.
[2] Zelano J, Redfors P, Asberg S, Kumlien E. Association between poststroke epilepsy and death: A nationwide cohort study. Eur Stroke J 2016;1(4):272–8.
[3] Haapaniemi E, Strbian D, Rossi C, Putaala J, Sipi T, Mustanoja S, et al. The CAVE score for predicting late seizures after intracerebral hemorrhage. Stroke 2014;45(7):1971–6.
[4] Kwon SY, Obeidat AZ, Sekar P, Moomaw CJ, Osborne J, Testai FD, et al. Risk factors for seizures after intracerebral hemorrhage: Ethnic/Racial Variations of Intracerebral Hemorrhage (ERICH) Study. Clin Neurol Neurosurg 2020;192:105731.
[5] Wang Y, Li Z, Zhang X, Chen Z, Li D, Chen W, et al. Development and validation of a clinical score to predict late seizures after intracerebral hemorrhage in Chinese. Epilepsy Res 2021;172:106600.
[6] Jette N, Beghi E, Hesdorffer D, Moshe SL, Zuberi SM, Medina MT, et al. ICD coding for epilepsy: past, present, and future–a report by the International League Against Epilepsy Task Force on ICD codes in epilepsy. Epilepsia 2015;56(3):348–55.
[7] Unified Medical Language System. RxNorm [Internet]. 2022 [.
[8] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[9] Cortes CV, V.; Saitta, L. Support-vector networks. Mach Learn. 1995;20(3):273-97.
[10] Cover TMH, P.E. Nearest Neighbor Pattern Classification. IEEE Trans Inf Theory. 1967;13(1):21-7.
[11] Efron B. The efficiency of logistic regression compared to normal discriminant analysis. Am Stat Assoc 1975;70(352):892–8.
[12] Freud YS, R.E. Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comut Syst Sci. 1997;55(1):119-39.
[13] Quinlan JR. Induction of Decision Trees. Mach Learn 1986;1(1):81–106.
[14] LightGBM - Microsoft Research. Accessed December 9, 2023 [Available from: https://www.microsoft.com/en-us/research/project/lightgbm/.
[15] Pedregosa FVG, Gramfort AL, Michel V, Thirion B, Grisel O, Blondel M, et al. Scikit-learn: Machine Learning in Python Journal of Machine Learning Research. 2011;12(85):2825−30.
[16] Ke GMQ, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst [Internet]. 2017:[3149-57 pp.]. Available from: https://github.com/Microsoft/LightGBM.
[17] Lundberg SML, SI, Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. Advances in Neural Information Processing Systems. 3220p. 4765-74.
[18] Devinsky O, Dilley C, Ozery-Flato M, Aharonov R, Goldschmidt Y, Rosen-Zvi M, et al. Changing the approach to treatment choice in epilepsy using big data. Epilepsy Behav 2016;56:32–7.
[19] Seccia R, Romano S, Salvetti M, Crisanti A, Palagi L, Grassi F. Machine learning use for prognostic purposes in multiple sclerosis. Life (Basel) 2021;11(2).
[20] Bunney G, Murphy J, Colton K, Wang H, Shin HJ, Faigle R, et al. Predicting early seizures after intracerebral hemorrhage with machine learning. Neurocrit Care 2022;37(Suppl 2):322–7.
[21] Li J, Huang Y, Hutton GJ, Aparasu RR. Assessing treatment switch among patients with multiple sclerosis: A machine learning approach. Explor Res Clin Soc Pharm 2023;11:100307.
[22] Hesdorffer DC, Benn EK, Cascino GD, Hauser WA. Is a first acute symptomatic seizure epilepsy? Mortality and risk for recurrent seizure. Epilepsia 2009;50(5):1102–8.
[23] Menon B, Shorvon SD. Ischaemic stroke in adults and epilepsy. Epilepsy Res 2009;87(1):1–11.
[24] Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, et al. ILAE official report: a practical clinical definition of epilepsy. Epilepsia 2014;55(4):475–82.