RESEARCH ARTICLE

Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis

Yizhao Ni[1,2]*, Kathleen Alwell[3], Charles J. Moomaw[3], Daniel Woo[3], Opeolu Adeoye[4], Matthew L. Flaherty[3], Simona Ferioli[3], Jason Mackey[5], Felipe De Los Rios La Rosa[6], Sharyl Martini[7], Pooja Khatri[3], Dawn Kleindorfer[3], Brett M. Kissela[3]

1 Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States of America, **2** Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America, **3** Department of Neurology and Rehabilitation Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America, **4** Department of Emergency Medicine and Neurosurgery, University of Cincinnati, Cincinnati, Ohio, United States of America, **5** Department of Neurology, Indiana University, Indianapolis, Indiana, United States of America, **6** Baptist Health Neuroscience Center, Miami, Florida, United States of America, **7** Michael E. DeBakey VA Medical Center, Houston, Texas, United States of America

* yizhao.ni@cchmc.org

OPEN ACCESS

Citation: Ni Y, Alwell K, Moomaw CJ, Woo D, Adeoye O, Flaherty ML, et al. (2018) Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis. PLoS ONE 13(2): e0192586. https://doi.org/10.1371/journal.pone.0192586

Editor: Neil R. Smalheiser, University of Illinois-Chicago, UNITED STATES

Received: January 9, 2017

Accepted: January 26, 2018

Published: February 14, 2018

Data Availability Statement: The data from the Greater Cincinnati/Northern Kentucky Stroke Study (GCNKSS) were collected under the auspices of the NIH funding (grant number: R01NS30678). The data are hosted and managed by the GCNKSS team that is included in our author list. The study team is committed to data sharing as per NIH policy. Because the data are identified as necessary for the GCNKSS, the dataset is only available on request. The interested researchers can make specific requests to collaborate via the GCNKSS website (www.gcnkss.com). They can submit an

Abstract

Objective

1) To develop a machine learning approach for detecting stroke cases and subtypes from hospitalization data, 2) to assess algorithm performance and predictors on real-world data collected by a large-scale epidemiology study in the US; and 3) to identify directions for future development of high-precision stroke phenotypic signatures.

Materials and methods

We utilized 8,131 hospitalization events (ICD-9 codes 430–438) collected from the Greater Cincinnati/Northern Kentucky Stroke Study in 2005 and 2010. Detailed information from patients' medical records was abstracted for each event by trained research nurses. By analyzing the broad list of demographic and clinical variables, the machine learning algorithms predicted whether an event was a stroke case and, if so, the stroke subtype. The performance was validated on gold-standard labels adjudicated by stroke physicians, and results were compared with stroke classifications based on ICD-9 discharge codes, as well as labels determined by study nurses.

Results

The best performing machine learning algorithm achieved a performance of 88.57%/93.81%/92.80%/93.30%/89.84%/98.01% (accuracy/precision/recall/F-measure/area under ROC curve/area under precision-recall curve) on stroke case detection. For detecting stroke subtypes, the algorithm yielded an overall accuracy of 87.39% and greater than 85% precision on individual subtypes. The machine learning algorithms significantly outperformed the

ICD-9 method on all measures (P value<0.001). Their performance was comparable to that of study nurses, with better tradeoff between precision and recall. The feature selection uncovered a subset of predictive variables that could facilitate future development of effective stroke phenotyping algorithms.

Discussion and conclusions

By analyzing a broad array of patient data, the machine learning technologies held promise for improving detection of stroke diagnosis, thus unlocking high statistical power for subsequent genetic and genomic studies.

"abstract/manuscript preparation form" (https://www.gcnkss.com/forms/manuscript-preparation-intake-form) to describe their study hypothesis and request the corresponding data. The users can specific the year of the data, the types of strokes, and the clinical variables in the request form. All variables used in our study have been listed in the Supporting Information.

Funding: This study was supported by grant number 1R01LM012230 from National Library of Medicine, https://www.nlm.nih.gov, to YN; grant number 1U01HG008666, from National Human Genome Research Institute, https://www.genome.gov, to YN; grant number: R01NS30678, from National Institute of Neurological Disorders and Stroke, https://www.ninds.nih.gov, to KA, CM, DW, OA, MF, SF, JM, FDLRLR, SM, PK, DK, BK; and no grant number, from Cincinnati Children's Hospital Medical Center, https://www.cincinnatichildrens.org, to YN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Stroke is the fifth leading cause of death in the US and is a major cause of adult disability.[1] Patients with stroke require expensive long-term rehabilitation care, resulting in an annual cost of over $33 billion nationally.[1] The main pathological subtypes of stroke include ischemic stroke, hemorrhagic stroke, and transient ischemic attack (TIA). Understanding clinical causation of stroke and its subtypes is critical for the planning, implementation, and evaluation of patient treatments. In particular, it will enable development of stroke phenotypes, which is the first step toward more powerful genetic and genomic studies that can lead to a better understanding of stroke etiology.[2–4] However, determination of stroke and its subtypes requires integration of multiple demographic, clinical, diagnostic, and imaging features; consequently, there is great variability between individual patients.[5–12]

Previous efforts have been made to identify predictors associated with stroke diagnosis. Medical history of hypertension, hyperlipidemia, obesity, diabetes mellitus, and atrial fibrillation have been commonly recognized as risk factors associated with stroke.[6, 9, 13–16] Computed tomography (CT) and magnetic resonance imaging (MRI) are routinely used in the diagnostic work-up of stroke patients. As new technologies of image processing have been introduced over time, imaging patterns have been increasingly adopted as "image markers" to facilitate stroke diagnosis.[17–19] In addition to clinical characteristics, patient demographics, family history, and substance use behaviors are considered influential factors on their risk of stroke.[6, 11, 20] Despite these findings, no single factor or group of factors would make a definite diagnosis. Rule-based approaches have been developed to heuristically combine the predictors to identify stroke cases, but large variability in reported sensitivities and specificities exists for the assessments.[21, 22] To detect stroke subtypes, current studies typically rely on International Classification of Diseases (ICD) codes or death certificate data. However, they are limited by precisions ranging from 6% to 97% across study designs and stroke subtypes.[23–28] Physician review of patients' complicated medical records remains the gold-standard method of ascertaining stroke diagnosis, and the process is labor intensive and expensive.[29, 30]

Machine learning (ML) is a methodology of data analytics that utilizes computerized algorithms to identify the relation between, and make prediction on, sets of data. By iteratively learning from example inputs (i.e., training data), ML algorithms identify hidden insights of the data and generate predictions on unseen examples (i.e., test data). In the literature of stroke research, ML technologies have been applied to identify stroke cases,[15, 31] predict stroke outcomes (e.g., mortality and recurrent stroke),[32–34] and evaluate therapy outcomes.[35, 36] Nevertheless, most of the studies have been limited to small patient cohorts (fewer than

200 samples), explored limited predictors, and did not have statistical power to discover relationships among a larger set of risk factors. A handful of studies utilized larger datasets (about 3000) to develop stroke detection models.[37, 38] However, their optimal accuracy plateaued at less than 75%.[37] In particular, none of the studies investigated the detection of stroke subtypes. Because ascertainment of stroke subtypes requires integration of findings from multiple clinical assessments and diagnostic tests,[39–41] the complexity and accuracy in detecting individual subtypes can vary dramatically.[25, 26] Additional study is therefore required to evaluate the effectiveness of ML technologies on stroke subtype detection.

Epidemiology studies collect a tremendous amount of multi-site samples with corresponding demographic and clinical data.[5, 42–44] In particular, some studies utilize physician review of the electronic health record (EHR) data to confirm stroke diagnosis for improved ascertainment accuracy.[43] By utilizing a comprehensive list of clinical data collected from such population-based metropolitan study, we investigated ML methodology to detect stroke diagnosis.

Objective

Our long-term objective is to develop a phenotyping algorithm that retrospectively identifies stroke cases across institutions to support genetic and genomic research. Because genetic and genomic studies typically require a case cohort of high purity (represented with a precision of 95%), we aim to establish a ML approach to detect stroke diagnosis with high precision and adequate recall. The specific aims of this study are: 1) to develop a ML approach to detect stroke cases and subtypes based on a broad array of hospitalization data; 2) to assess algorithm performance and predictors on real-world data collected from a large-scale epidemiology study of stroke in the US; and 3) to identify directions for future development of stroke phenotypic signatures. The study is the first, known to us, to investigate detection of multiple stroke subtypes in a large-scale via ML technologies.

Materials and methods

We utilized all hospitalization events collected from the Greater Cincinnati/Northern Kentucky Stroke Study (GCNKSS), a large-scale, population-based epidemiology study that measures temporal trends in stroke incidence rates in a population of 1.3 million.[43] The study was approved by the institutional review boards of participating hospitals (University of Cincinnati, Tri-Health, the Jewish Hospital and Mercy Hospital System, the Christ Hospital, and the St. Elizabeth Healthcare) and a waiver of individual consent was authorized (Study ID: 2013–3959 04061501).

Fig 1 diagrams the overall processes of the study. We first extracted clinical variables and stroke adjudications for all hospitalization events from the GCNKSS data (processes 1 and 2 in Fig 1). ML technologies were then applied to build stroke detection models with three steps: 1) features were generated from the variables and were normalized (process 3), 2) feature selection was applied to select predictive features for model construction (process 4), and 3) a variety of ML algorithms were developed to detect stroke diagnosis based on the selected features (process 5). Finally, the performance of ML models was assessed and compared with that of ICD-9 method and human experts (process 6).

Stroke events and patient EHR data

The GCNKSS collected and ascertained all potential stroke events that occurred among residents of the study region in 2005 and 2010. The GCNKSS first identified hospitalization events with potential stroke-related diagnoses from all 19 regional hospitals using ICD-9-CM codes

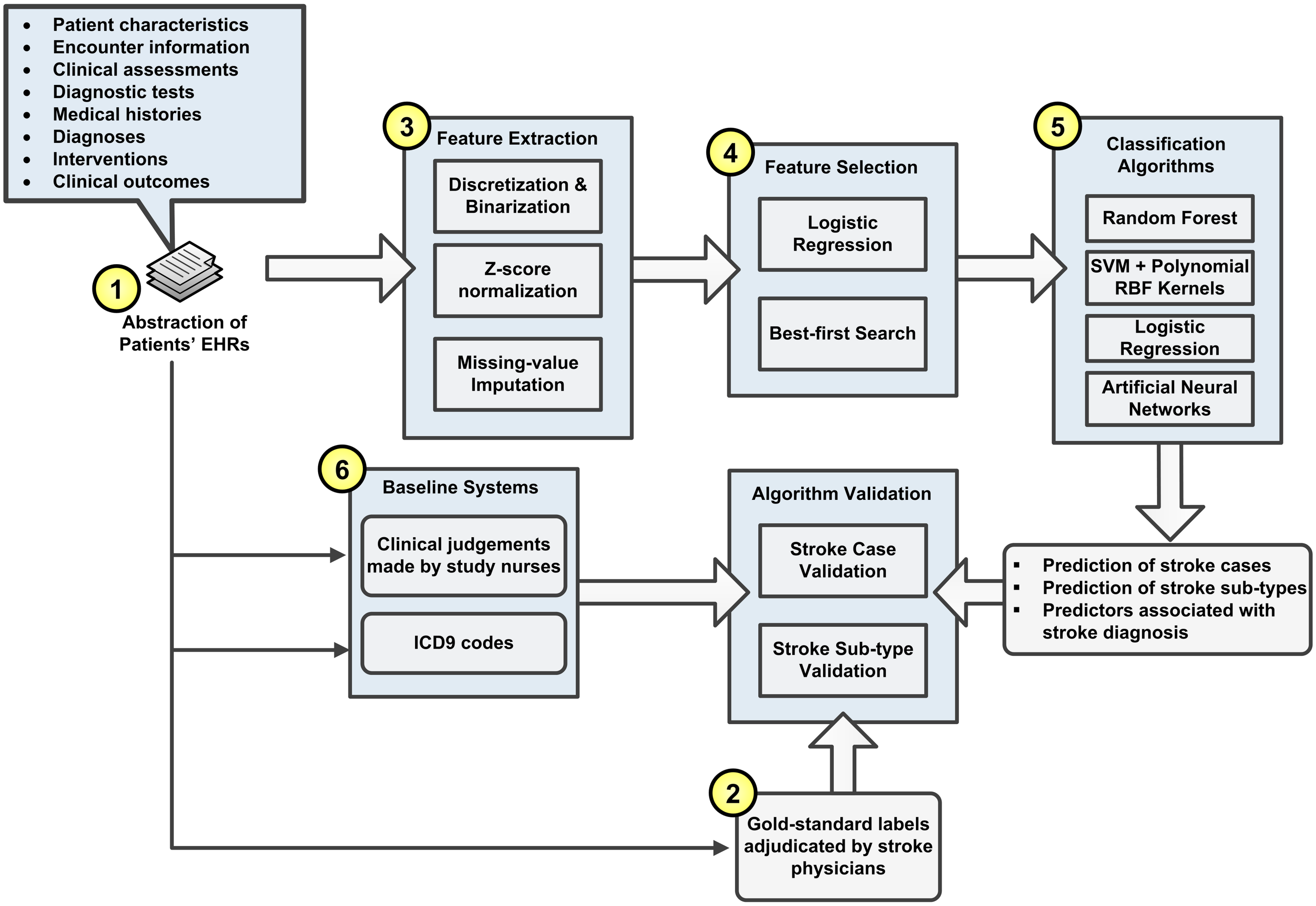


Fig 1. The overall processes of the study.

https://doi.org/10.1371/journal.pone.0192586.g001

(primary or secondary discharge diagnoses) of 430–438 that match the World Health Organization definition of stroke.[45] Detailed information from patients' EHRs was then abstracted for each event by trained research nurses (process 1 in Fig 1). We selected all 316 structured variables that specified patients' characteristics, encounter information, physiological status, diagnostic tests, medical histories, hospital diagnoses, interventions and clinical outcomes. The variables were categorized into 19 sets, which are summarized in Table 1. The description of each variable is presented in S1 Table. Because our goal was to retrospectively identify stroke cases, we leveraged all available information from a patient's hospitalization, including ICD discharge codes, interventions, and clinical outcomes.

Gold-standard stroke diagnosis

Ten stroke physicians were available to adjudicate study abstracts. Each abstract was reviewed by at least one stroke physician to determine whether the event was a stroke case and, if so, the stroke subtype (process 2 in Fig 1). Complicated events (35.1% of the collected events) were adjudicated by at least two physicians through group discussion to ensure the accuracy of diagnosis. The adjudicators had rigid criteria to determine stroke cases and subtypes,[43] but they were allowed to use their clinical judgment to clarify events (e.g., MRI negative for stroke but

Table 1. Summary of the variables used in the study.

Variable Category	Number of Variables	Description
DEMO	6	Patient demographics, including age, sex, race, ethnicity, marital status and employment status
SU	11	Patients' history of substance use (smoking, alcohol and street drugs)
VI	4	Visit information at time of admission (e.g., type of first medical contact, type of visited institution)
ED	13	Evaluations (e.g., blood pressure, Glasgow Coma Scale) performed in the emergency department
SE	29	Stroke-related evaluations (e.g., NIH stroke scale)
SS	20	Signs and symptoms that caused a patient to seek medical attention (e.g., weakness, headache, speech and vision)
CT/MRI	24	CT or MRI performed (Yes/No) and, if so, the findings (e.g., normal, acute infarct, intracerebral hemorrhage)
ANG	6	MRA, CTA, or cerebral angiography performed (Yes/No) and, if so, the findings (e.g., normal/abnormal)
CU	2	Carotid ultrasound performed (Yes/No) and, if so, the findings (e.g., normal/abnormal)
ECHO	19	Echocardiogram performed (Yes/No) and, if so, the findings (e.g., cardiomyopathy Yes/No)
EKG	16	Electrocardiogram performed (Yes/No) and, if so, the findings (e.g., normal/abnormal)
LAB	14	Laboratory results collected during hospitalization (e.g., white blood cell count, glucose level, total cholesterol)
MH	52	General medical history prior to hospitalization (e.g., history of hypertension Yes/No)
SH	18	History of stroke prior to hospitalization (e.g., ischemic stroke Yes/No)
ICD9	1	Primary and secondary ICD-9 codes on patients' discharge lists
DX	47	Complications and new diagnoses during hospitalization (e.g., pain, seizure, cardiac arrest Yes/No)
IT	13	Interventions performed (e.g., aneurysm clipping/coiling, clot evacuation Yes/No)
TH	15	Therapies performed (e.g., physical, occupational or speech therapy Yes/No)
OC	6	Clinical outcome of hospitalization (e.g., disposition at discharge, modified Rankin Scale)

https://doi.org/10.1371/journal.pone.0192586.t001

clinical symptoms and history consistent with stroke diagnosis could be called a case). For our study, we maintained the case criteria without exception: an event was labeled as a stroke case only if it met the case criteria. The event labels adjudicated by physicians were used to train and evaluate the ML algorithms.

Detecting stroke diagnosis with ML technologies

Feature extraction. We followed the methodology used in our earlier studies to process the clinical variables.[46–48] All nominal variables (e.g., sex, ICD-9 codes) were converted to binary features using dummy variable coding.[49] We then used two methods to discretize and normalize numerical variables. The National Institute of Health Stroke Scale (NIHSS) and Glasgow Coma Scale were discretized into categories based on clinical classification.[50, 51] For example, the NIHSS was discretized into no symptoms (0), mild (1–4), moderate (5–14), severe (15–24), and very severe (25–42) to stratify stroke severity.[50] The real-valued vital signs were discretized into "normal" and "out of normal range".[52] The remaining variables, including age and laboratory results, were normalized using z-score normalization.[53]

Because a patient might not take all diagnostic tests and assessments during hospitalization, the event samples could have missing values on certain variables. To alleviate the influence of missing data, we implemented unique-value imputation and grand mean and mode imputation based on their computational efficiency and performance in ML tasks.[54, 55] For each nominal variable, we created a unique category representing "unknown" for missing values. For a numerical variable, we replaced the missing values with the variable's mean (for continuous variables) or mode (for discrete variables) derived from the data.

Feature selection. The feature extraction yielded a large number of features for model construction. To reduce noise and avoid overfitting, we implemented a wrapper-based feature selection using logistic regression (LR) and best first search.[56, 57] The feature selection also provided a better insight of individual variables contributing to stroke diagnosis.

In each iteration, features from a category variable (Table 1) was added to the LR for training and testing to determine the top-performing category. The process was repeated until all 19 categories were added. The optimal feature set was chosen as the point at which additional features did not increase predictive performance. Note that some ML algorithms (e.g., random forest) inherently eliminate irrelevant features during model training, and they might not benefit from feature selection. As such, whether using the original or the optimal feature sets was tuned for individual ML algorithms based on the cross-validation performance.

Stroke case detection. We formatted detection of stroke cases as a binary-class classification and implemented four ML classifiers: 1) LR, a direct probability model that measures the linear relationship between features and stroke diagnosis;[58] 2) support vector machines with polynomial (SVM-P) and radial basis function (SVM-R) kernels, which construct hyperplanes in linear and non-linear feature spaces to classify stroke cases and non-stroke "controls";[59] 3) random forest (RF), which uses a multitude of decision trees to learn a highly irregular combination of features;[60] and 4) artificial neural networks (ANNs) that comprise three layers of LR models to learn non-linear patterns among features.[58] We chose these classifiers to allow for the possibility of linear and non-linear relationships between features and stroke diagnosis.

The classifiers output predictive values ($-\infty$, $+\infty$) to represent the possibility of stroke diagnosis. If a predictive value was positive, we assigned +1 to the output suggesting a stroke case. Otherwise, we assigned -1 suggesting a non-stroke "control". Given that the values output by ANNs ranged between 0 and 1, we set the threshold to 0.5 for ANNs.

Stroke subtype detection. We modeled stroke subtype detection as a task of four-class (ischemic stroke, hemorrhagic stroke, TIA, and non-stroke "control") classification. The RF and ANNs are natural multiclass classifiers, and they can predict the possibilities of classes simultaneously. The LR and SVM were extended to multiclass setting using the one-versus-all approach,[58] which trained a single classifier per class, with the samples of that class as cases and all other samples as controls. After training, it applied all classifiers to a test example and predicted the class for which the corresponding classifier output the highest predictive value.

Coping with imbalanced data. The distribution of stroke events in the real-world data was unbalanced, which could cause prediction bias and compromise the performance of ML algorithms.[61] Because the majority of abstracted events were stroke cases, the ML algorithms might predict all events as cases; this would achieve high accuracy, but would sacrifice other measures such as precision. To address this issue, we adopted adaptive synthetic sampling (ADA-SYN) to oversample minority class (e.g., non-stroke "control") in the training data.[62] The algorithm adaptively synthesized different numbers of samples from each minority example until the classes reached similar sizes. The balanced data were then used to train the ML algorithms. Similar to feature selection, the ADA-SYN sampling was integrated into the cross-validation process.

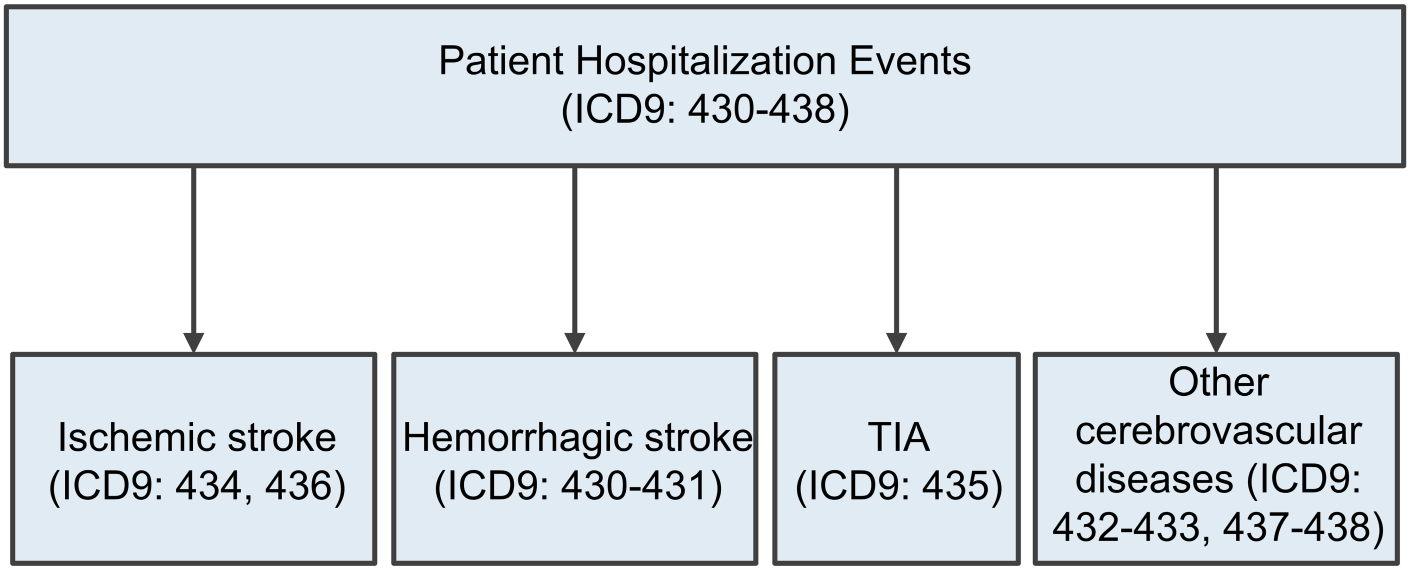


Fig 2. The ICD-9 coded baseline.

https://doi.org/10.1371/journal.pone.0192586.g002

Baseline systems

We implemented two baseline systems to compare with the ML algorithms (process 6 in Fig 1). The first baseline was an ICD9-coded method (denoted by ICD9) because phenotype algorithms frequently use ICD codes to identify qualified cases.[26, 63] The method was developed to identify high-precision stroke cohorts, and its logic is illustrated in Fig 2.[26] In the GCNKSS the abstractors (trained research nurses) also provided their judgments of stroke diagnosis for each event. We used these clinical judgments as the second baseline (denoted by CLIN) that simulated the decision-making of research nurses on stroke diagnosis.

Experiments

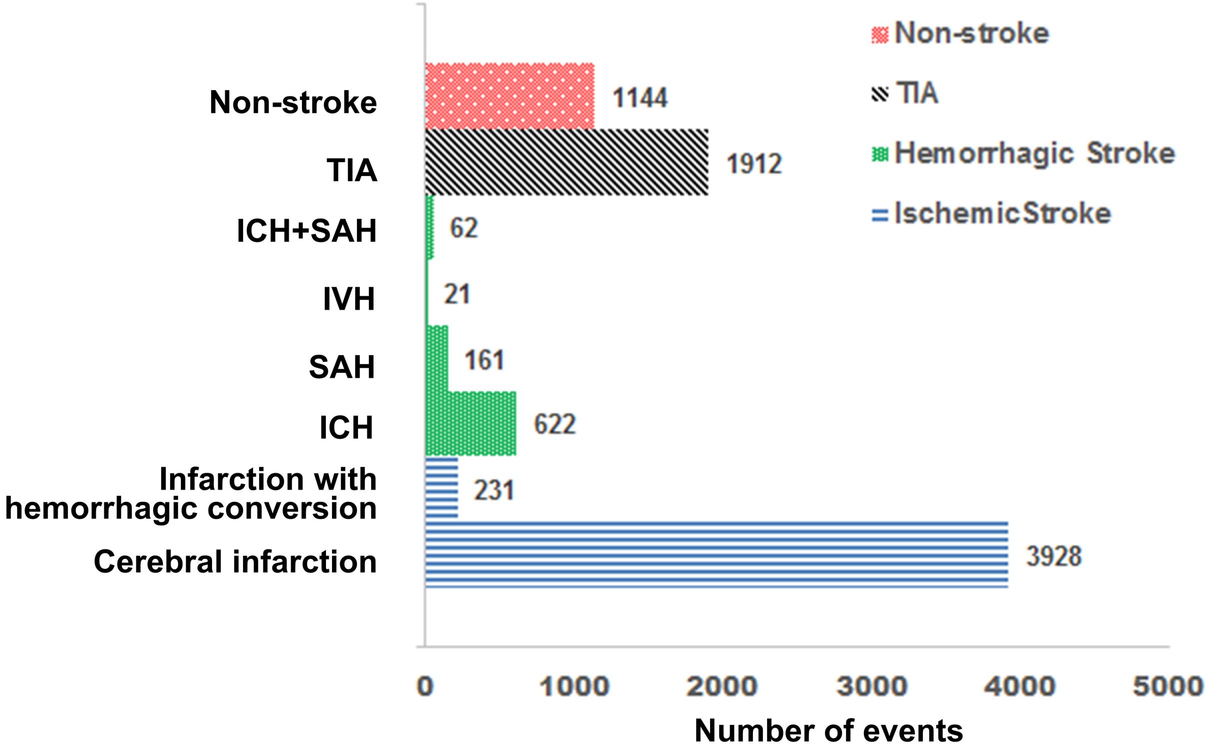
Evaluation metrics. We adopted five customary evaluation metrics to assess algorithm performance: 1) Accuracy = (True positives+True negatives)/Total events (denoted by ACC); 2) Precision = True positives/(True positives+False positives) (denoted by P); 3) Recall = True positives/(True positives+False negatives) (denoted by R); 4) F-measure = 2PxR/(P+R) (denoted by F); and 5) the area under receiver operating characteristics curves, which measures balance between recall and specificity (denoted by AUC).[64–66] Because the goal of this study was to identify high-precision stroke cohorts with adequate recall, we also generated precision-recall curves and measured the area under the curve (denoted by AUC-PR) to assess balance between precision and recall.[67]

Experiment setup. We performed a stratified random sampling based on number of events for each stroke subtype to split the data into two sets, 80% for training and development and 20% for testing and error analysis. Two iterations of ten-fold cross-validation were applied on the training set to select features and tune model parameters. Both cross-validation processes used the same data partition. The first cross-validation was applied to perform feature selection and generate the optimal feature set. The second cross-validation was used to tune hyper-parameters of the ML classifiers, including cost parameters (C) for LR, SVM-P, SVM-R and ANN (screened at 2 increments from 2^{-10} to 2^{16}), optimal degree for SVM-P (screened from 1 to 6), parameter γ for SVM-R (screened from 2^{-15} to 2^{5}), number of trees for RF (screened from 2^2 to 2^{11}), and number of neuros for ANNs (screened at 20 increments from 10 to 100). Whether using ADA-SYN sampling and the optimal feature set was also tuned during the second cross-validation process. Finally, the ML classifiers with optimal parameters were applied to the test data for performance comparison and error analysis.

For stroke case detection, events with a definite stroke diagnosis were labeled +1, and events without a stroke diagnosis were labeled -1. The event labels were then used to train and evaluate the ML algorithms. Feature selection was performed to identify predictive variables. All evaluation metrics were used, and we adopted the AUC-PR as the primary measure.

For stroke subtype detection, events were grouped into four categories (ischemic stroke, hemorrhagic stroke, TIA, and non-stroke "control", Fig 3) based on their subtypes. They were then labeled 1–4 to train and evaluate the algorithms. The optimal feature set was inherited from stroke case detection that captured informative variables for all subtypes. We did not perform feature selection in the multiclass setting because the small sample sizes of minority classes (e.g., hemorrhagic stroke) could cause overfitting during feature selection and propagate errors to the classifiers.[68, 69] For evaluation we reported overall accuracy, and precision, recall, and F-measure on each category. We also compared confusion matrices between different algorithms. The accuracy was adopted as the primary measure. We did not assess AUC and AUC-PR because they were primarily designed for binary classification.

Statistical analysis. Our primary outcome was to demonstrate that using the ML approach would detect stroke diagnosis more accurately, compared with baseline methods



TIA: Transient ischemic attack, ICH: Intracerebral hemorrhage, SAH: Subarachnoid hemorrhage, IVH: Intraventricular hemorrhage

Fig 3. The event distribution of stroke subtypes among the four categories.

https://doi.org/10.1371/journal.pone.0192586.g003

(ICD9 and CLIN). To this end, the statistical significance of the difference between systems' performances was assessed and reported using paired T-test.[70]

In our experiments the ML algorithms, evaluation metrics, and statistical analyses were implemented using MATLAB Version 2014a.[71]

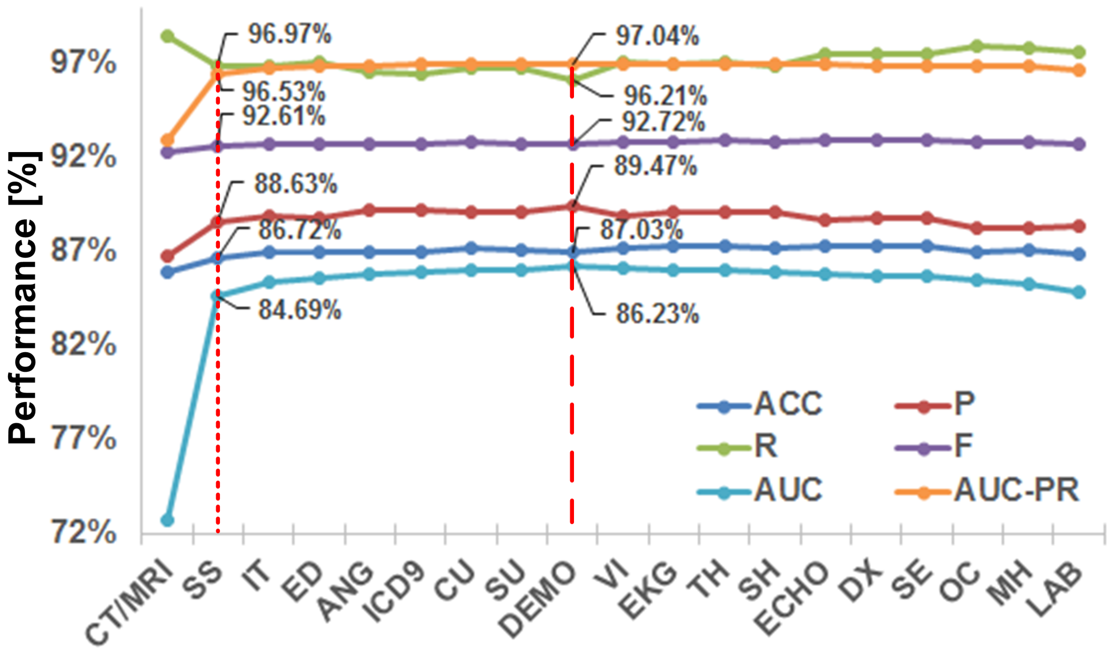
Results

Descriptive statistics of the data set

The study personnel reviewed a total of 8,131 events, of which 6,987 samples (85.9%) were adjudicated to be stroke cases. We excluded 50 samples (0.6%) due to undetermined stroke subtypes. Fig 3 depicts the event distribution of stroke subtypes among the four categories. After stratified sampling and feature extraction, the training set contained 6,463 samples (3,327/692/1,529/915 ischemic/hemorrhagic/TIA/non-stroke) with 1,071 unique features. The test set had 1,618 samples (832/174/383/229 ischemic/hemorrhagic/TIA/non-stroke) with 994 features. In total there were 1,091 features in the data set.

Results of feature selection

Fig 4 shows the performance curves on cross-validation for each incremental variable set addition. All performance measures except recall increased and then stabilized. The best AUC-PR and AUC achieved by feature selection were 97.04% and 86.23% respectively (dash line in Fig 4). The optimal feature set included CT/MRI findings (CT/MRI), signs and symptoms (SS), interventions (IT), ED assessments (ED), findings from angiography (ANG) and carotid ultrasound (CU) tests, ICD-9 codes (ICD9), substance use characteristics (SU), and demographics



ACC - Accuracy; P - Precision; R - Recall; F - F-measure; AUC - Area under ROC curve; AUC-PR – Area under precision-recall curve

Fig 4. The performance curves when adding the variable sets (Table 1).

https://doi.org/10.1371/journal.pone.0192586.g004

Table 2. Performance of different classification algorithms for stroke case identification.

Measure	Cross Validation Performance [%]						
	ICD9	CLIN	LR	SVM-P	SVM-R	RF	ANN
ACC	60.45	85.41	87.17	87.56	**88.07**	87.56	87.35
P	87.96	86.06	88.96	90.26	90.32	**92.75**	91.25
R	62.47	**99.05**	97.11	95.87	96.45	92.81	94.31
F	73.05	92.10	92.86	**92.97**	93.28	92.78	92.75
AUC	55.83	50.60	86.11	85.93	86.41	**88.02**	85.98
AUC-PR	87.91	83.29	97.15	96.81	96.86	**97.54**	96.74
Measure	**Test Set Performance [%]**						
	ICD9	**CLIN**	**LR**	**SVM-P**	**SVM-R**	**RF**	**ANN**
ACC	61.68	85.85	87.21	87.89	**88.38**	88.57	86.90
P	88.41	86.39	89.40	90.94	90.99	**93.81**	91.59
R	63.72	**99.14**	96.54	95.39	95.97	92.80	93.31
F	74.06	92.32	92.84	93.11	**93.41**	93.30	92.44
AUC	55.40	51.65	86.69	86.31	86.61	**89.84**	85.87
AUC-PR	88.29	83.51	97.23	97.19	97.22	**98.01**	96.89

Bold numbers indicate the best results.

https://doi.org/10.1371/journal.pone.0192586.t002

(DEMO). Using only CT/MRI and SS achieved an AUC-PR/AUC of 96.53%/84.69% (dotted line), which was close to the optimal performance.

Performance of stroke case detection

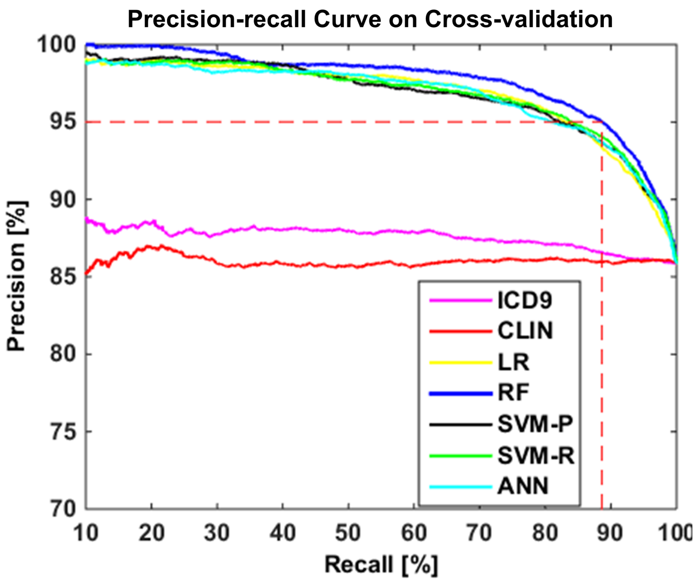
Table 2 shows the performance of different classification algorithms on detecting stroke cases. Compared with the ICD9 baseline, the ML classifiers performed significantly better on all measures (Table 3). They also outperformed research nurses (CLIN) on all measures except recall. Fig 5 plots precision-recall curves generated by the algorithms. The best curves were generated by the RF, with AUC-PR of 97.54% on cross-validation and 98.01% on the test set.

Table 3. Statistical significance tests (paired T-test) of the performance difference between the machine learning algorithms and the baselines on stroke case identification.

Baseline	Measure	P Values between the Machine Learning Algorithms and the Baselines				
		LR	SVM-P	SVM-R	RF	ANN
ICD9	ACC	1.04E-12*	1.63E-12*	6.82E-13*	1.65E-12*	9.61E-12*
	P	4.99E-4*	4.95E-6*	7.01E-6*	7.68E-9*	4.67E-7*
	R	3.47E-13*	1.40E-12*	2.50E-13*	9.98E-13*	4.34E-12*
	F	1.50E-12*	2.07E-12*	8.01E-13*	2.31E-12*	9.13E-12*
	AUC	4.40E-11*	3.46E-12*	3.11E-12*	8.43E-12*	1.58E-11*
	AUC-PR	8.28E-12*	1.25E-11*	1.01E-11*	2.46E-12*	1.02E-11*
CLIN	ACC	1.63E-4*	4.65E-5*	1.16E-5*	5.45E-5*	6.96E-4*
	P	1.53E-7*	6.62E-10*	1.43E-10*	1.04E-10*	1.49E-10*
	R	0.999	1.00	0.999	1.00	1.00
	F	1.00E-3*	7.85E-4*	1.13E-4*	3.90E-3*	1.48E-2*
	AUC	4.40E-11*	1.50E-11*	7.37E-12*	1.23E-11*	4.86E-11*
	AUC-PR	5.94E-9*	3.25E-9*	3.80E-9*	3.94E-9*	9.13E-9*

*indicates statistical significance (p value < 0.05).

https://doi.org/10.1371/journal.pone.0192586.t003



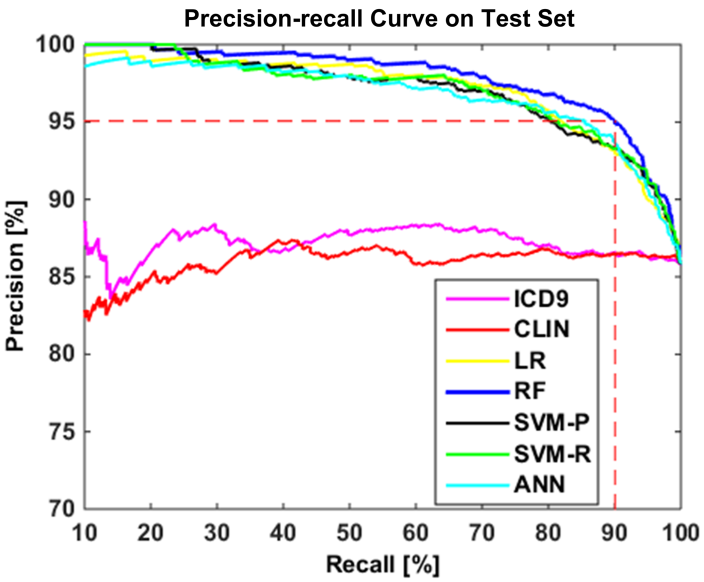


Fig 5. Precision-recall curves generated by the algorithms.

https://doi.org/10.1371/journal.pone.0192586.g005

Performance of stroke subtype detection

Table 4 shows the algorithm performance on stroke subtype detection, where the statistical significance tests were reported in Table 5 on overall accuracy and per-class precision that were of most interest to our study. The improvements of ML classifiers over ICD9 were statistically significant. The ML classifiers also outperformed CLIN significantly on accuracy, and on precisions for ischemic stroke, TIA, and non-stroke "control". The RF achieved the highest accuracy, and its improvements over the other classifiers were statistically significant (p value <0.05 under paired t-test). Fig 6 illustrates the confusion matrices generated by ICD9, CLIN, and the best-performing RF on the test set, in which an off-diagonal cell (i,j) numbers the events in category i that were misclassified into category j. A more diagonal matrix suggests a more accurate match between algorithm predictions and gold-standard labels.

Discussion

Despite being the most common approach for recording clinical conditions, the ICD-9 methods are sub-optimal for phenotyping diseases including stroke.[24] All ML algorithms performed better than ICD9 significantly for stroke case detection. The RF achieved the best performance in terms of the primary measure (Table 2). Its performance was also comparable to that of trained research nurses (CLIN), with better tradeoff between precision and recall (evidenced by the higher AUC-PRs achieved). Both ICD9 and CLIN did not achieve a precision of 95% (Fig 5), and thus their predictions could not be utilized directly to support genetic and genomic research.[72, 73] In comparison, the best-performing RF could provide approximately 90% of the cases with 95% precision (dotted line in Fig 5), which would allow high statistical power for subsequent studies without labor-intensive clinician review.

For stroke subtype detection, the precisions obtained by the algorithms varied between subtypes, with the best on hemorrhagic stroke, followed by ischemic stroke and TIA (Table 4). The variation of performance was in accordance with complexities in diagnosing these stroke subtypes: if a stroke is caused by hemorrhage, a CT scan can show evidence immediately.[40] However, a normal CT scan does not rule out the diagnosis of ischemic stroke and a MRI,

Table 4. Performance of different classification algorithms for stroke type identification.

Category	Measure	Cross Validation Performance [%]						
		ICD9	CLIN	LR	SVM-P	SVM-R	RF	ANN
Overall	ACC	68.67	84.05	86.66	85.63	86.57	**86.90**	85.75
Ischemic Stroke	P	80.64	83.36	89.37	**93.87**	92.22	92.58	90.31
	R	79.59	**97.57**	94.17	88.48	91.05	90.53	91.80
	F	80.11	89.91	**91.70**	91.08	91.62	91.54	91.04
Hemorrhagic Stroke	P	87.40	93.88	94.31	**94.68**	93.85	92.69	94.61
	R	82.80	**98.85**	96.97	94.36	96.24	97.98	94.51
	F	84.99	**96.27**	95.60	94.50	95.00	95.23	94.50
Transient Ischemic Attack	P	67.42	83.83	87.02	**88.67**	87.43	86.16	88.00
	R	72.47	**96.60**	94.36	89.19	93.18	96.07	91.48
	F	69.81	89.75	90.53	88.90	90.19	**90.84**	89.69
Non-stroke Control	P	12.52	31.23	**60.70**	52.30	57.00	59.18	54.67
	R	11.90	2.72	38.71	**62.64**	51.91	49.96	47.55
	F	12.18	4.98	47.10	**56.89**	54.20	54.15	50.75

Category	Measure	Test Set Performance [%]						
		ICD9	CLIN	LR	SVM-P	SVM-R	RF	ANN
Overall	ACC	67.68	84.12	86.40	85.17	87.08	**87.39**	87.08
Ischemic Stroke	P	79.50	82.43	89.91	**94.58**	93.28	93.60	91.16
	R	79.69	**97.00**	92.07	86.06	90.02	89.66	91.71
	F	79.59	89.12	90.97	90.12	**91.62**	91.59	91.43
Hemorrhagic Stroke	P	88.02	**97.18**	97.14	95.98	95.53	94.48	94.35
	R	84.48	**98.85**	97.70	95.98	98.28	98.28	95.98
	F	86.22	**98.01**	97.42	95.98	96.88	96.34	95.16
Transient Ischemic Attack	P	65.99	84.48	87.65	**89.58**	88.21	86.71	89.22
	R	67.36	96.61	94.52	89.82	93.73	**97.13**	95.04
	F	66.67	90.13	90.96	89.70	90.89	91.63	**92.04**
Non-stroke Control	P	11.95	50.00	56.18	50.09	56.77	**59.24**	58.67
	R	11.79	5.24	43.67	**65.54**	56.77	54.59	50.22
	F	11.87	9.49	49.14	56.78	56.77	**56.82**	54.12

Bold numbers indicate the best results.

https://doi.org/10.1371/journal.pone.0192586.t004

particularly diffusion-weighted imaging, is typically required to confirm the findings.[41] Finally, the MRI shows diagnostic findings in a low percentage of TIA cases.[39] Determining TIA additionally relies on a patient's ability to provide a history of transient stroke-like symptoms, and on a physician's ability to match these symptoms to the operational concept of TIA. Consequently, the clinical diagnosis of TIA is difficult and has limited inter-observer reliability.[74]

The experimental results (Fig 6 and Table 4) suggested that such complexities in stroke diagnosis affected the baselines and ML algorithms differently. Without comprehensive information from patient records, the ICD9 baseline was unable to distinguish among stroke subtypes accurately. The research nurses were capable of identifying hemorrhagic stroke, but they tended to overcall more complicated subtypes (as evidenced by its confusion matrix). Compared with humans, the confusion matrix made by RF showed fewer misclassifications between ischemic stroke, TIA and non-stroke "control". In fact, the RF showed comparable performance on detecting hemorrhagic stroke and significantly better

Table 5. Statistical significance tests (paired T-test) of the performance difference between the machine learning algorithms and the baselines on stroke type identification.

Baseline	Measure	P Values between the ML Algorithms and the Baselines				
		LR	SVM-P	SVM-R	RF	ANN
ICD9	Overall ACC	2.23E-10*	2.62E-10*	1.63E-9*	1.65E-10*	3.21E-10*
	P (Ischemic stroke)	6.07E-9*	9.81E-8*	5.24E-8*	1.02E-8*	4.40E-8*
	P (Hemorrhagic stroke)	1.15E-5*	4.73E-5*	2.27E-5*	4.87E-5*	7.92E-5*
	P (TIA)	2.13E-10*	2.17E-9*	7.82E-11*	3.72E-10*	8.84E-11*
	P (Non-stroke control)	2.63E-9*	8.46E-9*	1.61E-8*	8.88E-10*	2.19E-9*
CLIN	Overall ACC	2.40E-3*	1.89E-2*	1.64E-4*	7.57E-7*	2.60E-3*
	P (Ischemic stroke)	4.35E-11*	3.64E-9*	3.94E-9*	1.73E-10*	5.33E-9*
	P (Hemorrhagic stroke)	0.104	0.229	0.529	0.993	0.116
	P (TIA)	3.33E-5*	6.05E-5*	4.56E-6*	2.34E-5*	1.33E-4*
	P (Non-stroke control)	1.90E-3*	2.90E-3*	7.05E-4*	3.14E-4*	2.00E-3*

*indicates statistical significance (p value < 0.05).

https://doi.org/10.1371/journal.pone.0192586.t005

precisions on all other categories (Table 4). The findings suggested the strength of ML-based methods in capturing and weighing information from different aspects of patient data to detect stroke subtypes.

In addition, the feature selection process identified a subset of predictive variables that synthesizes a human-oriented conceptualization of stroke diagnosis. The majority of the variables were related to diagnostic tests for stroke (CT/MRI, ANG and CU in Fig 4), and patients' physiological characteristics during hospitalization (SS and ED).[75] Interventions (IT) such as carotid endarterectomies were used for stroke prevention and they could imply higher risk of stroke onset. Finally, patients' demographics (DEMO) and substance use behaviors (SU) were shown to be influential, which were consistent with the literature findings.[6, 11, 20] In particular, the CT/MRI and SS were the most predictive variables and they yielded more than 98% of the performance gain (Fig 4). The relative importance of these variables could help physicians weigh the information when chart reviewing a patient's record.

Our findings contribute to the body of knowledge in stroke research on several fronts. In the experiments the ML models were evaluated on a population from multiple hospitals, and the positive results suggested their generalizability in stroke detection. As such, the developed approach has potential to facilitate case identification for multi-site genomic studies.[72, 73] By leveraging a centralized dataset, a coordinating center could develop and disseminate ML models along with data abstraction protocols. The participating sites could then abstract site-specific data and apply the models to identify stroke cases. The feature selection uncovered a subset of predictive variables (CT/MRI and SS) that could facilitate the development of more effective phenotyping algorithms to reduce workload in data abstraction, which is an interesting direction of our future work. In addition, the ML approach has potential to generalize to other applications in stroke research. Currently, stroke epidemiology studies that utilize administrative databases suffer from misclassification bias by using only ICD discharge codes, whereas the studies involving manual inspection such as the GCNKSS are hindered by time required for data collection and adjudication. By calibrating the ML predictions with linear regression, we could estimate incidence rates of stroke in a study region with a high degree of confidence.[76] As such, the developed approach could provide great benefit for reducing time and effort for executing stroke-related epidemiology studies, allowing near real-time estimates of stroke incidence.

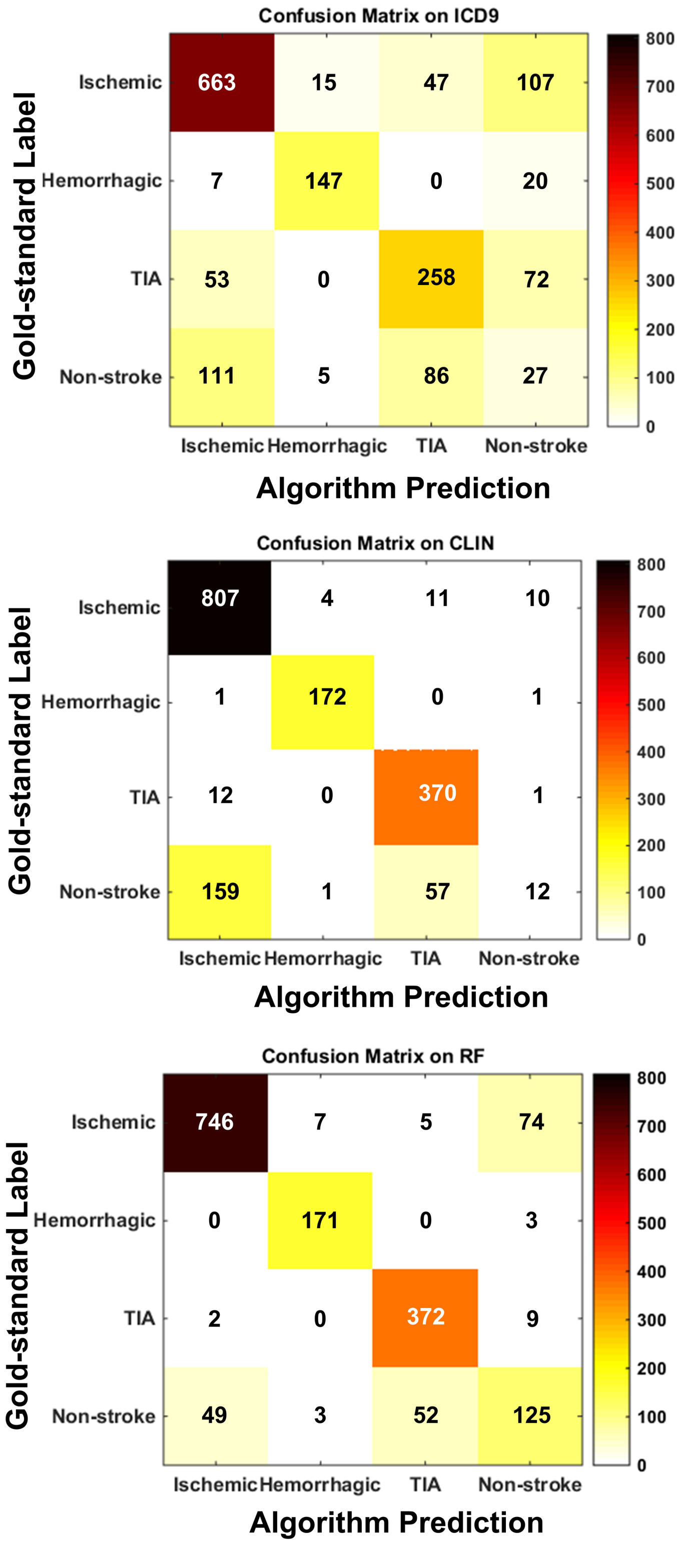


Fig 6. Confusion matrices generated by ICD9, CLIN, and RF on the test set.

https://doi.org/10.1371/journal.pone.0192586.g006

Table 6. Misclassification errors made by the RF algorithm on the test set.

ID	Gold-standard label	IS			HS	TIA		NS		
	Predicted label	HS	TIA	NS	NS	IS	NS	IS	HS	TIA
1	No focal symptoms and key diagnostic tests (CT/MRI findings) were not performed (16.67%)	0	0	2	0	0	0	12	0	20
2	Missing CT/MRI findings (e.g., "no acute intracranial abnormality") stored in textual data fields (11.27%)	0	0	6	1	0	1	1	0	14
3	Physicians used information not in the data (e.g., raw MRI images and clinical notes) to make the decisions (6.86%)	0	0	7	0	1	1	3	1	1
4	Missing information (e.g., MRI findings) due to ED or outpatient settings (6.86%)	0	1	4	1	0	2	2	0	4
5	Dilemma samples. Physicians determined as cases but the patients did not meet all inclusion criteria. The events were labeled as non-stroke "control" in our study (14.71%)	0	0	0	0	0	0	27	2	1
6	Complex cases. Ischemic stroke with hemorrhagic conversion (4.90%)	7	0	3	0	0	0	0	0	0
7	Undetermined etiology of cases. No focal symptoms or findings from diagnostic tests (12.25%)	0	4	16	1	0	4	0	0	0
8	Conflict findings between symptoms and diagnostic tests (21.08%)	0	0	33	0	1	1	4	0	4
9	Wrong predictions. Unidentified reason (5.39%)	0	0	2	0	0	0	1	0	8

IS: Ischemic stroke; HS: Hemorrhagic stroke; TIA: Transient ischemic attack; NS: non-stroke control. Percentage of errors for each category is presented in the bracket.

https://doi.org/10.1371/journal.pone.0192586.t006

Error analysis, limitations and future work

We performed error analysis for the RF algorithm on stroke subtype detection. The algorithm made 204 errors on the test data. By reviewing physicians' clarification on these events, we grouped the errors into nine categories. Table 6 shows the error categories and the numbers of misclassifications between gold-standard and predicted labels for each category. Approximately 42% of errors were due to missing information in the data (categories 1–4). For some events the CT/MRI tests were not performed, possibly due to that the patients did not have stroke symptoms and hence the tests were determined unnecessary by healthcare providers (category 1). The observation suggested that interaction between variables (e.g., no symptoms plus no diagnostic tests) could be informative for stroke diagnosis. In the future we will explicitly model variable connections with tensor product representation and see if it improves the accuracy in stroke detection.[77] In addition, some findings were stored in textual data fields and not used in the current study (category 2). Utilizing NLP algorithms to extract information from these fields is therefore another direction of our future work. Finally, important information was missed occasionally due to healthcare settings (e.g., outpatient) and difficulty of abstraction (e.g., missing subtle information from clinical notes) (categories 3–4). This observation could benefit the design of a more effective abstraction protocol, which however, is out of scope of this study.

Another 32% of errors were ascribed to the complexity of events (categories 5–8). The algorithm identified physicians' decisions well but did not capture more rigid inclusion criteria, hence misclassifying a noticeable amount of non-stroke "controls" into cases (category 5). It also confused between ischemic and hemorrhagic strokes when the events were ischemic strokes with hemorrhagic conversion (category 6). To solve these problems, we plan to implement knowledge-based post-processing to explicitly include structured inclusions and to adjust algorithm predictions when patients present both ischemic and hemorrhagic characteristics. In addition, the algorithm misclassified several ischemic events with unknown etiology, in which the patients did not present traditional stroke symptoms and diagnostic findings (category 7). Understanding the etiology of these events will help identify predictors for the ML-based methods, which warrants further investigation by neurologists.

Finally, approximately 21% of errors were caused by conflicts between patients' symptoms and diagnostic findings (mainly CT/MRI findings). If a patient had focal stroke symptoms but CT/MRI findings were normal, the study physicians often override the findings and considered the patient a case because the symptoms could be mild such that they did not show up on CT/MRI. In contrast, if there were multiple CT/MRI tests showing consistent findings, the physicians would weigh more on diagnostic results even if the patient did not have symptoms. Compared with stroke physicians, the RF algorithm always weighted more on CT/MRI findings and had less flexibility in balancing conflict variables, consequently misclassifying a notable amount of events in which the CT/MRI findings were normal. To alleviate this problem, we will develop advanced multi-layer classifiers in our future work to balance weights between different variable sets before aggregating them for stroke detection.[47]

One significant limitation of the study is that the variables used were abstracted by research nurses. Manual abstraction of clinical variables requires not only substantial subspecialty expertise, but also intensive manpower. Consequently, the limitation could hinder the dissemination of the developed approach across institutions. To alleviate this problem, variables should ideally be extracted from EHR data automatically. Recent studies have shown the feasibility of automating abstraction of stroke related risk factors from EHR data.[78, 79] Because the ML models could achieve competitive performance with a limited set of 44 variables (CT/MRI and SS), we anticipate that automated data abstraction for stroke detection is feasible with appropriate NLP and regular expression algorithms.

Another limitation is that we did not assess the inter-observer reliabilities among stroke physicians in the epidemiology study. Although each hospitalization event was reviewed by at least one clinical nurse and a stroke physician, and the complex events were adjudicated through group discussion, variability may exist in the final adjudications, particularly for TIA and stroke cases with negative diffusion-weighted imaging results. To address this limitation, we have initiated documentation of physician decisions in the ongoing GCNKSS, which allows for the evaluation of inter-observer reliabilities on future data. In addition, we grouped the stroke subtypes into four categories to avoid the problem of data sparseness. To improve the granularity of detection, we will continue collecting data from the GCNKSS to develop more powerful predictive models.

As a final limitation, the work was limited to reporting system performance on a population collected in a single epidemiology study. To assess its generalizability, project planning is in progress to evaluate the developed approach in a separate stroke population with different data collection and representation methods.

Conclusions

In this study we demonstrated the strength of ML technologies in identifying stroke cases and pathological subtypes. By analyzing a broad array of patient data, the ML models showed good capacity for detecting stroke diagnosis. The algorithms significantly outperformed the ICD-9 method that is commonly implemented in current studies. Their performance was comparable to that of trained research nurses, with better tradeoff between precision and recall. The feature selection uncovered a subset of predictive variables, which could facilitate future development of effective stroke phenotyping algorithms. The anticipated benefits of machine learning have potential to bring stroke phenotyping to the forefront of biomedical research, unlocking high statistical power for subsequent genetic and genomic studies.

Supporting information

S1 Table. Description of the variables used in the study.
(DOCX)

Author Contributions

Conceptualization: Yizhao Ni, Brett M. Kissela.

Data curation: Yizhao Ni, Kathleen Alwell, Charles J. Moomaw.

Formal analysis: Yizhao Ni, Kathleen Alwell, Daniel Woo, Opeolu Adeoye, Matthew L. Flaherty, Simona Ferioli, Jason Mackey, Felipe De Los Rios La Rosa, Sharyl Martini, Pooja Khatri, Dawn Kleindorfer, Brett M. Kissela.

Funding acquisition: Dawn Kleindorfer, Brett M. Kissela.

Investigation: Yizhao Ni, Kathleen Alwell, Charles J. Moomaw, Daniel Woo, Opeolu Adeoye, Matthew L. Flaherty, Simona Ferioli, Jason Mackey, Felipe De Los Rios La Rosa, Sharyl Martini, Pooja Khatri, Dawn Kleindorfer, Brett M. Kissela.

Methodology: Yizhao Ni, Charles J. Moomaw, Brett M. Kissela.

Software: Yizhao Ni.

Supervision: Brett M. Kissela.

Validation: Yizhao Ni.

Writing – original draft: Yizhao Ni.

Writing – review & editing: Yizhao Ni, Kathleen Alwell, Charles J. Moomaw, Daniel Woo, Opeolu Adeoye, Matthew L. Flaherty, Simona Ferioli, Jason Mackey, Felipe De Los Rios La Rosa, Sharyl Martini, Pooja Khatri, Dawn Kleindorfer, Brett M. Kissela.

References

1. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics-2016 update: A report from the american heart association. Circulation. 2016; 133(4):e38–60. PMID: 26673558
2. Black M, Wang W, Wang W. Ischemic stroke: From next generation sequencing and gwas to community genomics? OMICS. 2015; 19(8):451–60. https://doi.org/10.1089/omi.2015.0083 PMID: 26230531
3. Lindgren A. Stroke genetics: A review and update. J Stroke. 2014; 16(3):114–23. https://doi.org/10.5853/jos.2014.16.3.114 PMID: 25328870
4. Tonk M, Haan J. A review of genetic causes of ischemic and hemorrhagic stroke. J Neurol Sci. 2007; 257(1–2):273–9. https://doi.org/10.1016/j.jns.2007.01.037 PMID: 17328915
5. Wolfe CD, Tilling K, Beech R, Rudd AG. Variations in case fatality and dependency from stroke in western and central europe. The european biomed study of stroke care group. Stroke. 1999; 30(2):350–6. PMID: 9933270
6. O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the interstroke study): A case-control study. Lancet. 2010; 376(9735):112–23. https://doi.org/10.1016/S0140-6736(10)60834-3 PMID: 20561675
7. Jackson CA, Hutchison A, Dennis MS, Wardlaw JM, Lindgren A, Norrving B, et al. Differing risk factor profiles of ischemic stroke subtypes: Evidence for a distinct lacunar arteriopathy? Stroke. 2010; 41 (4):624–9. https://doi.org/10.1161/STROKEAHA.109.558809 PMID: 20150553
8. Shaikh Q, Ahmed B, Ahmed M, Mahar JH, Ahmad M, Ahmed A, et al. Left atrial volumes and associated stroke subtypes. BMC Neurol. 2013; 13:149. https://doi.org/10.1186/1471-2377-13-149 PMID: 24139054
9. Ozkul-Wermester O, Guegan-Massardier E, Triquenot A, Borden A, Perot G, Gerardin E. Increased blood-brain barrier permeability on perfusion computed tomography predicts hemorrhagic transformation in acute ischemic stroke. Eur Neurol. 2014; 72(1–2):45–53. https://doi.org/10.1159/000358297 PMID: 24853726
10. Arsava EM, Bas DF, Atalar E, Has AC, Oguz KK, Topcuoglu MA. Ischemic stroke phenotype in patients with nonsustained atrial fibrillation. Stroke. 2015; 46(3):634–40. https://doi.org/10.1161/STROKEAHA.114.006396 PMID: 25634003

11. Trivedi MM, Ryan KA, Cole JW. Ethnic differences in ischemic stroke subtypes in young-onset stroke: The stroke prevention in young adults study. BMC Neurol. 2015; 15:221. https://doi.org/10.1186/s12883-015-0461-7 PMID: 26515647

12. Kleindorfer D, Khoury J, Alwell K, Moomaw CJ, Woo D, Flaherty ML, et al. The impact of magnetic resonance imaging (mri) on ischemic stroke detection and incidence: Minimal impact within a population-based study. BMC Neurol. 2015; 15:175. https://doi.org/10.1186/s12883-015-0421-2 PMID: 26407627

13. Mitchell AB, Cole JW, McArdle PF, Cheng YC, Ryan KA, Sparks MJ, et al. Obesity increases risk of ischemic stroke in young adults. Stroke. 2015; 46(6):1690–2. https://doi.org/10.1161/STROKEAHA.115.008940 PMID: 25944320

14. Aslanyan S, Weir CJ, Lees KR, GAIN International Steering Committee and Investigators. Elevated pulse pressure during the acute period of ischemic stroke is associated with poor stroke outcome. Stroke. 2004; 35(6):E153–E5. https://doi.org/10.1161/01.STR.0000126598.88662.16 PMID: 15073388

15. Moons KG, Bots ML, Salonen JT, Elwood PC, Freire de Concalves A, Nikitin Y, et al. Prediction of stroke in the general population in europe (eurostroke): Is there a role for fibrinogen and electrocardiography? J Epidemiol Commun H. 2002; 56:I30–I6.

16. Hayden DT, Hannon N, Callaly E, Ni Chroinin D, Horgan G, Kyne L, et al. Rates and determinants of 5-year outcomes after atrial fibrillation-related stroke: A population study. Stroke. 2015; 46(12):3488–93. https://doi.org/10.1161/STROKEAHA.115.011139 PMID: 26470776

17. Tyan YS, Wu MC, Chin CL, Kuo YL, Lee MS, Chang HY. Ischemic stroke detection system with a computer-aided diagnostic ability using an unsupervised feature perception enhancement method. Int J Biomed Imaging. 2014; 2014:947539. https://doi.org/10.1155/2014/947539 PMID: 25610453

18. Tang FH, Ng DK, Chow DH. An image feature approach for computer-aided detection of ischemic stroke. Comput Biol Med. 2011; 41(7):529–36. https://doi.org/10.1016/j.compbiomed.2011.05.001 PMID: 21605853

19. Bentley P, Ganesalingam J, Carlton Jones AL, Mahady K, Epton S, Rinne P, et al. Prediction of stroke thrombolysis outcome using ct brain machine learning. Neuroimage Clin. 2014; 4:635–40. https://doi.org/10.1016/j.nicl.2014.02.003 PMID: 24936414

20. Knottnerus IL, Gielen M, Lodder J, Rouhl RP, Staals J, Vlietinck R, et al. Family history of stroke is an independent risk factor for lacunar stroke subtype with asymptomatic lacunar infarcts at younger ages. Stroke. 2011; 42(5):1196–200. https://doi.org/10.1161/STROKEAHA.110.602383 PMID: 21441152

21. Purrucker JC, Hametner C, Engelbrecht A, Bruckner T, Popp E, Poli S. Comparison of stroke recognition and stroke severity scores for stroke detection in a single cohort. J Neurol Neurosurg Psychiatry. 2015; 86(9):1021–8. https://doi.org/10.1136/jnnp-2014-309260 PMID: 25466259

22. Sanders LM, Srikanth VK, Blacker DJ, Jolley DJ, Cooper KA, Phan TG. Performance of the abcd2 score for stroke risk post tia: Meta-analysis and probability modeling. Neurology. 2012; 79(10):971–80. https://doi.org/10.1212/WNL.0b013e31825f9d02 PMID: 22700810

23. Goldstein LB. Accuracy of icd-9-cm coding for the identification of patients with acute ischemic stroke: Effect of modifier codes. Stroke. 1998; 29(8):1602–4. PMID: 9707200

24. Quan H, Li B, Saunders LB, Parsons GA, Nilsson CI, Alibhai A, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Services Research. 2008; 43(4):1424–41. https://doi.org/10.1111/j.1475-6773.2007.00822.x PMID: 18756617

25. Olson KL, Wood MD, Delate T, Lash LJ, Rasmussen J, Denham AM, et al. Positive predictive values of icd-9 codes to identify patients with stroke or TIA. American Journal of Managed Care. 2014; 20(2):E27–E34. PMID: 24738552

26. Woodfield R, Grant I, UK Biobank Stroke Outcomes Group, UK Biobank Follow-Up and Outcomes Working Group, Sudlow CL. 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(10):e0140533. https://doi.org/10.1371/journal.pone.0140533 PMID: 26496350

27. McCormick N, Bhole V, Lacaille D, Avina-Zubieta JA. Validity of diagnostic codes for acute stroke in administrative databases: A systematic review. PLoS One. 2015; 10(8):e0135834. https://doi.org/10.1371/journal.pone.0135834 PMID: 26292280

28. Chang TE, Lichtman JH, Goldstein LB, George MG. Accuracy of ICD-9-CM codes by hospital characteristics and stroke severity: Paul coverdell national acute stroke program. Journal of the American Heart Association. 2016; 5(6):e003056. https://doi.org/10.1161/JAHA.115.003056 PMID: 27247334

29. Coull AJ, Silver LE, Bull LM, Giles MF, Rothwell PM. Direct assessment of completeness of ascertainment in a stroke incidence study. Stroke. 2004; 35(9):2041–5. https://doi.org/10.1161/01.STR.0000137605.48864.2f PMID: 15256682

30. Feigin VL, Carter K. Editorial comment—stroke incidence studies one step closer to the elusive gold standard? Stroke. 2004; 35:2045–7. PMID: 15331801

31. Colak C, Karaman E, Turtay MG. Application of knowledge discovery process on the prediction of stroke. Comput Methods Programs Biomed. 2015; 119(3):181–5. https://doi.org/10.1016/j.cmpb.2015.03.002 PMID: 25827533

32. Tirschwell DL, Longstreth WT, Becker KJ, Gammans RE Sr, Sabounjian LA, Hamilton S, et al. Shortening the NIH stroke scale for use in the prehospital setting. Stroke. 2002; 33(12):2801–6. PMID: 12468773

33. Peng SY, Chuang YC, Kang TW, Tseng KH. Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. Eur J Neurol. 2010; 17(7):945–50. https://doi.org/10.1111/j.1468-1331.2010.02955.x PMID: 20136650

34. Ho KC, Speier W, El-Saden S, Liebeskind DS, Saver JL, Bui AA, et al. Predicting discharge mortality after acute ischemic stroke using balanced data. In AMIA Annu Symp Proc. 2014; 2014:1787–96.

35. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. PLoS One. 2014; 9(2):e88225. https://doi.org/10.1371/journal.pone.0088225 PMID: 24520356

36. Cheng CA, Lin YC, Chiu HW. Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. Studies in health technology and informatics. 2014; 202:115–8. PMID: 25000029

37. Konig IR, Malley JD, Pajevic S, Weimar C, Diener HC, Ziegler A. Patient-centered yes/no prognosis using learning machines. Int J Data Min Bioin. 2008; 2(4):289–341.

38. Linder R, Konig IR, Weimar C, Diener HC, Poppl SJ, Ziegler A. Two models for outcome prediction—a comparison of logistic regression and neural networks. Methods Inf Med. 2006; 45(5):536–40. PMID: 17019508

39. Schaefer PW, Grant PE, Gonzalez RG. Diffusion-weighted MR imaging of the brain. Radiology. 2000; 217(2):331–45. https://doi.org/10.1148/radiology.217.2.r00nv24331 PMID: 11058626

40. Smith SD, Eskey CJ. Hemorrhagic stroke. Radiol Clin North Am. 2011; 49(1):27–45. https://doi.org/10.1016/j.rcl.2010.07.011 PMID: 21111128

41. Yew KS, Cheng E. Acute stroke diagnosis. Am Fam Physician. 2009; 80(1):33–40. PMID: 19621844

42. Brown RD, Whisnant JP, Sicks JD, OFallon WM, Wiebers DO. Stroke incidence, prevalence, and survival—secular trends in rochester, minnesota, through 1989. Stroke. 1996; 27(3):373–80. PMID: 8610298

43. Broderick J, Brott T, Kothari R, Miller R, Khoury J, Pancioli A, et al. The greater cincinnati northern kentucky stroke study—preliminary first-ever and total incidence rates of stroke among blacks. Stroke. 1998; 29(2):415–21. PMID: 9472883

44. Jiang G, Li W, Wang D, Shen C, Ji Y, Zheng W. Epidemiological transition and distribution of stroke incidence in tianjin, China, 1988–2010. Public Health. 2016; 131:11–9.

45. Hatano S. Experience from a multicentre stroke register: A preliminary report. Bull World Health Organ. 1976; 54(5):541–53. PMID: 1088404

46. Ni Y, Beck AF, Taylor R, Dyas J, Solti I, Grupp-Phelan J, et al. Will they participate? Predicting patients' response to clinical trial invitations in a pediatric emergency department. J Am Med Inform Assoc. 2016; 23(4):671–80. https://doi.org/10.1093/jamia/ocv216 PMID: 27121609

47. Zhai H, Srikant I, Ni Y, Lingren T, Kirkendall E, Li Q, et al. Mining a large-scale ehr with machine learning methods to predict all-cause 30-day unplanned readmissions. In 2nd ASE International Conference on Big Data Science and Computing. Standford University: Standford, CA; May 27–31, 2014.

48. Zhai H, Brady P, Li Q, Lingren T, Ni Y, Wheeler DS, et al. Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. Resuscitation. 2014; 85(8):1065–71. https://doi.org/10.1016/j.resuscitation.2014.04.009 PMID: 24813568

49. Hardy M. Regression with dummy variables. Newbury Park, CA: Sage; 1993.

50. Brott T, Adams HP Jr., Olinger CP, Marler JR, Barsan WG, Biller J, et al. Measurements of acute cerebral infarction: A clinical examination scale. Stroke. 1989; 20(7):864–70. PMID: 2749846

51. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. The Lancet. 1974; 304 (7872):81–4.

52. Simel DL. Approach to the patient: History and physical examination. In: Goldman L, Schafer AI, editors. Goldman's cecil medicine (24th ed). Philadelphia, PA: Elsevier Saunders; 2011.

53. Isaac S, Michael WB. Handbook in research and evaluation for education and the behavioral sciences (3rd ed.): EDITS; 1995.

54. Twala BETH, Jones MC, Hand DJ. Good methods for coping with missing data in decision trees. Pattern Recogn Lett. 2008; 29(7):950–6.

55. Ding YF, Simonoff JS. An investigation of missing data methods for classification trees applied to binary response data. Journal of Machine Learning Research. 2010; 11:131–70.

56. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997; 97(1–2):273–324.

57. Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003; 3:1157–82.

58. Bishop CM. Pattern recognition and machine learning: Springer Science+Business Media, LLC; 2006.

59. Shawe-Taylor J, Christianini N. Kernel methods for pattern analysis: Cambridge University Press; 2004.

60. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32.

61. Haibo H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009; 21(9):1263–84.

62. Haibo H, Yang B, Garcia EA, Shutao L. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. 2008:1322–8.

63. Mo H, Thompson WK, Rasmussen LV, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc. 2015 Nov; 22(6):1220–30. https://doi.org/10.1093/jamia/ocv112 PMID: 26342218

64. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994; 308(6943):1552. PMID: 8019315

65. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994; 309(6947):102. PMID: 8038641

66. Rice JA. Mathematical statistics and data analysis (3rd ed.): Duxbury Advanced; 2006.

67. Davis J, Goadrich M. The relationship between precision-recall and roc curves. In proc. of the 23rd International Conference on Machine Learning; 2006; 2006:233–40.

68. Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997; 19(2):153–8.

69. Sima C, Dougherty ER. What should be expected from feature selection in small-sample settings. Bioinformatics. 2006; 22(19):2430–6. https://doi.org/10.1093/bioinformatics/btl407 PMID: 16870934

70. McDonald JH. Handbook of biological statistics (3rd ed.): Sparky House Publishing; 2014.

71. MathWorks. Matlab—the language of technical computing. 2017 [cited 2 January, 2017]; https://www.mathworks.com/products/matlab.html

72. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: Results of the emerge consortium. Sci Transl Med. 2011; 3(79):79re1. https://doi.org/10.1126/scitranslmed.3001807 PMID: 21508311

73. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011; 4:13. https://doi.org/10.1186/1755-8794-4-13 PMID: 21269473

74. Kraaijeveld CL, van Gijn J, Schouten HJ, Staal A. Interobserver agreement for the diagnosis of transient ischemic attacks. Stroke. 1984; 15(4):723–5. PMID: 6464066

75. National heart lung and blood institute. How is a stroke diagnosed? 2016 [Accessed 10 Jul, 2016]; https://www.nhlbi.nih.gov/health/health-topics/topics/stroke/diagnosis

76. Ni Y, Alwell K, Moomaw CJ, Adeoye O, Flaherty ML, Ferioli S, et al. Towards automated incidence rate reporting: Leveraging machine learning technologies to assist stroke adjudication in a large-scale epidemiological study. Stroke. 2017; 48:A135.

77. Smolensky P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial Intelligence. 1990; 46(1–2):159–216.

78. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, et al. Extracting a stroke phenotype risk factor from veteran health administration clinical reports: An information content analysis. J Biomed Semantics. 2016; 7:26. https://doi.org/10.1186/s13326-016-0065-1 PMID: 27175226

79. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015; 350:h1885. https://doi.org/10.1136/bmj.h1885 PMID: 25911572