Natural Language Processing of Radiology Reports to Detect Complications of Ischemic Stroke

Matthew I. Miller, BA[1], Agni Orfanoudaki, PhD[2], Michael Cronin, BS[1], Hanife Saglam, MD[3], Ivy So Yeon Kim, MPH[4], Oluwafemi Balogun, MD, MPH[4,5], Maria Tzalidi, MD[6], Kyriakos Vasilopoulos, MD[6], Georgia Fanaropoulou, MD[6], Nina M. Fanaropoulou, BSc[7], Jack Kalin, BA[1], Meghan Hutch, BS[8,9], Brenton R. Prescott, MS[4], Benjamin Brush, MD[10], Emelia J. Benjamin, MD, ScM[1,5], Min Shin, PhD[11], Asim Mian, MD[12], David M. Greer, MD[1,4], Stelios M. Smirnakis, MD, PhD[9,13,14], Charlene J. Ong, MD, MPHS [1,4,9,10,13]

[1]Boston University School of Medicine, Boston, MA, USA

[2]Saïd Business School, University of Oxford, Oxford, UK

[3] Department of Neurology, West Virginia University School of Medicine, Morgantown, USA

[4]Department of Neurology, Boston Medical Center, Boston, MA, USA

[5]Boston University School of Public Health, Boston, MA, USA

[6]School of Medicine, University of Crete, Heraklion Greece

[7] School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece

[8]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA

[9]Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA

[10]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

[11]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

[12]Department of Radiology, Boston Medical Center, Boston, MA, USA

[13]Harvard Medical School, Boston, MA, USA

[14]Jamaica Plain Veterans Administration Hospital, Boston, MA, USA

Word Count: 3614 **Figures and Tables:** 7

Correspondence: Charlene J. Ong, MD, MPHS, cjong@bu.edu, 617-638-8647

Department of Neurology, Boston University School of Medicine, 85 E. Concord St., Suite 1116, Boston, MA, 02118

Keywords: Natural Language Processing, Stroke, Critical Care, Diagnostic Imaging

Details:

We assert that the following conditions for publication have been met:

1. The present manuscript complies with all instructions to authors set forth by *Neurocritical Care.*
2. All authors listed in this work have fulfilled authorship requirements and approved the present manuscript.
3. This manuscript has not been published in any other journal and is not currently under consideration for publication elsewhere. This work adheres to ethical guidelines and was approved by the Institutional Review Boards of Massachusetts General and Brigham & Women's Hospitals, as well as Boston Medical Center.
4. Author contributions are as follows: MIM, AO, CJO, and SMS conceived and designed the overall study. MIM, AO, and BB wrote original computer code in Python to train and test all models. Individual radiology reports were reviewed and labeled by MIM, CJO, HS, OB, MT, KV, GF, NMF, and JK. MIM, AO, CJO, BRP, MH, and ISYK organized and administered collected data for modeling. AM is a practicing neuroradiologist who reviewed selected reports to benchmark interrater reliability. MS provided technical assistance with model training and development. MIM, MC, ISYK, and CJO wrote the manuscript. MIM, AO, MC, HS, ISYK, OB, MT, KV, GF, NMF, JK, MH, BRP, BB, EJB, MS, AM, DMG, SMS, and CJO helped with reviews and revision. DMG and EJB provided overall study direction and critical review.
5. The authors do not have any relevant conflicts of interest that may impact this work.
6. This study does not require submission of STROBE, CONSORT, PRISMA, STARD, SQUIRE, ARRIVE, or TRIPOD checklists.
7. We acknowledge the following funding sources for contributing authors:

- CJO receives support from NIH/NINDS K23NS116033 and the Peter Paul Young Career Development Foundation at Boston University.
- EJB is supported in part by 2R01 HL092577; 1R01 HL141434 01A1; 1R01AG066010; 1R01AG066914; American Heart Association, AHA_18SFRN34110082.
- DMG receives support from R01 NS102574.
- SS receives support from RO1 EY024019 and VA Merit award I01RX002981

Abstract:

Background: Abstraction of critical data from unstructured radiologic reports using Natural Language Processing (NLP) is a powerful tool to automate the detection of important clinical features and enhance research efforts. We present a set of NLP approaches to identify critical findings in acute ischemic stroke patients from radiology reports of computed tomography (CT) and magnetic resonance imaging (MRI).

Methods: We trained machine learning classifiers to identify categorical outcomes of edema, midline shift (MLS), hemorrhagic transformation, and parenchymal hematoma, as well as rule-based systems (RBS) to identify intraventricular hemorrhage (IVH) and continuous MLS measurements within CT/MRI reports. Using a derivation cohort of 2289 reports from 550 individuals with acute middle cerebral artery (MCA) territory ischemic strokes, we externally validated our models on reports from a separate institution as well as from patients with ischemic strokes in any vascular territory.

Results: In all datasets, a deep neural network with pretrained biomedical word embeddings (BioClinicalBERT) achieved the highest discrimination performance for binary prediction of edema (Area Under Precision Recall Curve [AUPRC] > 0.94), MLS (AUPRC > 0.98), hemorrhagic conversion (AUPRC > 0.89), and parenchymal hematoma (AUPRC > 0.76). BioClinicalBERT outperformed Lasso regression ($p < 0.001$) for all outcomes except parenchymal hematoma ($p = 0.755$). Tailored RBS for IVH and continuous MLS outperformed BioClinicalBERT ($p < 0.001$) and linear regression respectively ($p < 0.001$).

Conclusions: Our study demonstrates robust performance and external validity of a core NLP toolkit for identifying both categorical and continuous outcomes of ischemic stroke from unstructured radiographic text data. Medically-tailored NLP methods have multiple important “Big Data” applications including

scalable electronic phenotyping, augmentation of clinical risk prediction models, and facilitation of automatic alert systems in the hospital setting.

Introduction:

Natural Language Processing (NLP) is a powerful tool to abstract and categorize critical data from unstructured radiologic reports for research and clinical purposes [1]. Assembling large cohorts for clinical and epidemiologic studies has historically required intensive manual chart review, limiting both sample size and generalizability [2-5]. Such efforts are particularly laborious in the case of critical outcomes after ischemic stroke such as life-threatening edema and hemorrhagic transformation, for which dedicated disease classification codes may be inaccurate or nonexistent. Direct derivation of relevant data from radiology reports ("electronic phenotyping") [6] represents an emerging strategy to improve the quality of retrospective datasets and develop clinical systems for triaging patients with critical findings [7].

NLP, a subfield of artificial intelligence dedicated to the study of human language [8], has emerged as a promising new approach to retrieve clinically-rich data contained in medical records and diagnostic reports [2, 3, 9, 10]. NLP makes use of a variety of approaches to analyze language and classify text according to pre-defined outcomes of interest. Methods that identify desired text-based information include specifically tailored rule-based systems (RBS) [11] as well as more general combinations of featurization methods (the conversion of text to numerical values) with supervised learning methodologies [12]. In the field of neurology, recent NLP studies have focused on identifying and subtyping ischemic stroke [13-15] and cerebral aneurysms [16], predicting neurologic outcomes after ischemia [17], and triaging patients with transient ischemic attack [18]. However, few studies have focused on feasibly-implementable methods to classify critical and acute stroke sequelae [19].

Previously, we tested a variety of featurization and machine learning (ML) classification methods to detect the presence, acuity, and location of ischemic stroke [4]. In the present work, we sought to expand our electronic phenotyping of neuroradiology reports by applying various NLP methodologies to detect critical complications of stroke relevant to neurocritical care contexts. Specifically, we characterized the following intracranial abnormalities using both Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) reports: 1) edema 2) midline shift (MLS) 3) hemorrhagic conversion 4) parenchymal hematoma (within hemorrhagic conversion), and 5) intraventricular hemorrhage (IVH). We hypothesized that recently-developed ML models pretrained on biomedical text would demonstrate improved classification of binary outcomes compared with traditional ML algorithms [20]. Moreover, we also hypothesized that customized RBSs would perform better at detecting rare and continuous outcomes of stroke than more general ML-based approaches. Our work expands progress in large-scale neurocritical care research by demonstrating the application of NLP methods for electronic phenotyping of relevant emergent radiographic complications of ischemic stroke.

Methods:

Study Population and Data Sources

To derive and assess the performance and generalizability of our NLP models, we accrued and labeled three distinct datasets of head CT, CT angiography, MRI, and MR angiography radiology reports from patients 18 years or older with an ICD-9 (433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11, 434.91) or ICD-10 (I63) diagnosis of ischemic stroke during admission. The three datasets are described as follows:

1. **MGB MCA Ischemic Stroke:** This dataset consisted of 2289 reports from 550 individuals presenting to Massachusetts General or Brigham and Women's (MGB) hospitals with acute ischemic stroke occurring in the middle cerebral artery (MCA) territory between 2006-2018. Reports were

identified by our previous NLP algorithm [4] from an initial Research Patient Data Registry (RPDR) query of 60,654 radiology reports, and the presence of MCA stroke was confirmed by expert reviewers (BB, CJO, HS).

2. **BMC MCA Ischemic Stroke:** This dataset comprised 377 randomly-selected reports from 154 patients out of a total of 4818 reports from 2069 patients with NLP-identified acute MCA ischemic stroke from Boston Medical Center (BMC) between 2005-2018. BMC, New England's largest safety net hospital, features a population with notable socioeconomic and demographic differences relative to the MGB system despite geographic proximity. Acute MCA ischemic stroke and outcomes of interest were confirmed and labeled by trained experts (JK, MM, OB).

3. **MGB Any Territory Ischemic Stroke:** This dataset comprised 520 reports from 214 patients with ICD-9/10 identified ischemic stroke in any vascular territory (not only MCA territory) from MGB hospitals between 2003-2018. Trained experts (HS, KV, MT) labeled all outcomes. No reports overlapped with the MGB MCA Ischemic Stroke dataset.

The MGB Human Research Committee and Boston University Medical Center Internal Review Boards (2017P002564, H-38431) approved this study.

Radiographic Report Labeling:

We sought to identify a core set of outcomes covering major inflammatory and hemorrhagic sequelae of acute ischemic stroke. Therefore, all reports included in this study were manually-labeled for the presence of edema, MLS, hemorrhagic conversion, parenchymal hematoma (as inferred according to ECASS-II criteria [21]), and IVH. Full labeling criteria are included in Supplementary Methods. Notably, we further subdivided MLS detection into two separate prediction tasks: a *categorical* task, in which we labeled reports for the binary presence of any MLS, and a *continuous* task in which we recorded the precise degree of MLS in millimeters. Research staff received training from three board-

certified neurologists (BB, CJO, SS) who performed regular audits of data collection to ensure accuracy and consistency. Deidentified reports were randomly distributed among study team members, and datasets were managed using a Research Electronic Data Capture (REDCap) electronic database [22]. Additionally, a board-certified neuroradiologist (AM) independently and in a blinded fashion assessed the accuracy of all labels in a random sample of 75 reports (25 per dataset) to provide expert-level benchmarking of study labels.

Text Preprocessing:

To reduce heterogeneities that impact classification, we processed raw radiology reports across all datasets in the following ways:

1. We converted all text to lowercase and removed excess white space.
2. We removed all text within the report header, deidentified the names of providers and patients, and removed any other instances of non-narrative text (e.g., characters representing divisions between report sections).
3. We "tokenized" terms to quantify common word patterns describing a single entity by condensing commonly-used groups of words to their corresponding "n-grams" (e.g. "subdural hematoma" → "subduralhematoma", "altered mental status" → "altermentalstatus"). The full token list is provided in the Supplementary Methods.

Featurization Methods:

Following preprocessing, we used three methods to convert ("featurize") unstructured text reports into numerical features capable of being used as direct input for statistical learning [8].

The first and simplest of these approaches, known as "Bag of Words" (BOW), counts each unique word or n-gram within a radiology report as a unique feature known as a "token." The value of each token is determined by the number of times each given word occurs in a unique report.

Term Frequency-Inverse Document Frequency (TF-IDF) is similar to BOW, but penalizes the value of each token according to the total number of reports in which that token occurs so that common words and n-grams like “and” and “the” are assigned lower values.

Word embedding strategies are more sophisticated NLP featurization methods that assign values to tokens based on their relationship to other words [23]. We used an embedding method called Bidirectional Encoder Representations from Transformers (BERT) model for our binary classification tasks [23]. Specifically, we used a domain-specific version of BERT embeddings (“BioClinicalBERT”) pretrained on 18 billion words worth of PubMed articles [24] and approximately 880 million words worth of critical care notes from the MIMIC-III v1.4 dataset [25]. These embeddings, originally developed by Alsentzer and colleagues [26], acted as a starting point for training of our stroke-specific classifiers.

Classifier Training:

For each binary task, we trained Lasso regression [27], K-nearest neighbor [28], random forest [29], and multilayer perceptron [30] algorithms using the BOW and TF-IDF featurization techniques. We also further developed a BioClinicalBERT transformer network using pretrained BERT embeddings, as described above. This type of network is a specialized type of sequential deep learning model which could be further “fine-tuned” for detection of post-stroke outcomes. For the task of continuous MLS estimation, we employed linear regression with both BOW and TF-IDF. Further details regarding the merits and hyperparameter tuning of each algorithm are included in our Supplementary Methods. We employed a standard desktop central processing unit (CPU) for training of all models with the exception of our BioClinicalBERT transformer, in which we used Graphics Processing Units (GPU) for training and testing.

We trained our models on 80% of the original MGB MCA Ischemic stroke cohort and tested it on a withheld set of 20% for internal validation. For our derivation cohort, we used bootstrapping to randomly split the data five times into training and testing sets. Both external validation cohorts were tested in full across all five splits of the data. For the parenchymal hematoma outcome, training and testing were conducted only on reports with a ground truth label of hemorrhagic conversion. Given that there were multiple tasks with differing prevalence and clinical relevance, we used 0.5 as the initial decision threshold for all classifiers.

We performed hyperparameter tuning with a grid search and 10-fold cross-validation during each round of bootstrapping. BioClinicalBERT models were trained for 10 epochs with a maximum sequence length of 512 and a learning rate of $4x10^{-5}$, which corresponds closely with hyperparameters reported in previous studies [26]. All models were developed and tested in Python 2.7 using the scikit-learn [31] package for non-BERT classifiers and the "simpletransformers" library for the BioClinicalBERT transformer. The detailed codebase is available within our Github repository (https://github.com/matthewmiller22/STROKE_NLP_2022).

Rule-Based System (RBS):

We hypothesized that the previously-described approaches would be ill-suited for rare and continuous outcomes. Therefore, we employed alternative RBSs to identify IVH (a relatively infrequent complication of ischemic stroke [32]) and continuous MLS measurements. For IVH, we replaced phrases with periods that led to misclassification, including other hemorrhage subtypes (e.g., "subdural hemorrhage,") phrases associated with ventricular effacement (e.g., "compression [of the ventricles],") and terms associated with neurosurgical procedures (e.g., "ventriculoperitoneal [shunt]"). We identified regular expression rules capturing both negative (e.g., "no ventricular blood") and positive (e.g., "layering of blood") phraseology for IVH. To identify continuous MLS, we extended our preprocessing

pipeline by removing superfluous numerical characters, dates, and times. We assigned a value of 0 to cases in which there was an absence of "shift," "deviation," or "herniation," as well as those in which there was explicit mention of "no midline shift" or "without midline shift." We then split individual text reports by sentence and abstracted numeric values in proximity to words such as "shift," "deviation," or "herniation." Regular expressions were constructed with the Regex library in Python 2.7. Further details of the design and implementation of each RBS are included in Supplementary Data.

Evaluation and Statistical Analysis

To assess interrater reliability between labelers and a board-certified neuroradiologist, we calculated Cohen's kappa statistic. We report discrimination using Area Under Precision Recall Curve (AUPRC), sensitivity, specificity, and precision (positive predictive value). Similar to other work in machine learning for neurocritical care [33], we chose AUPRC to be our main performance metric given that raw accuracy and area under receiver operating characteristic curve (AUROC) may overestimate performance in imbalanced datasets [34]. We do, however, demonstrate ROC curves for all classifiers in Supplementary Figure S1, and supply full performance metrics including accuracy and AUROC within Supplementary Tables S1-S5. Calibration curves for classifier probabilities are also provided in Supplementary Figure S2.

To compare the performance of our two best performing classifiers for each binary task, we used McNemar's test [35], a statistical procedure used in previous NLP studies to assess the equivalence of two models tested on the same data. We also used McNemar's test for comparison of our IVH RBS to the best-performing ML classifier. For each clinical outcome, we report the average χ^2 value across all five data partitions with 95% confidence intervals and p-values. Given that McNemar's test assumes independent observations, we performed this procedure using a dedicated dataset containing only one randomly-select report patient. To assess the performance of our RBS for continuous MLS, we

compared the absolute errors of RBS MLS estimates to linear regression estimates using a one-sided student's t-test. To account for our six total hypotheses, we used a Bonferroni correction [36] and set the significance level for all hypotheses to α=0.008. All statistical tests were performed using the scipy.stats library within Python 2.7.

A graphical overview of our methods, including cohort selection and model development is illustrated within Figure 1.

Results:

In our MGB MCA Ischemic Stroke dataset, we labeled 2289 CT and MRI radiology reports for all stroke outcomes (Fig 1). For external validation, we labeled 377 reports in the BMC MCA Ischemic Stroke set and 520 reports from the All Territory MGB dataset. The baseline characteristics of the overall cohort, including demographics and clinical outcomes, are described in Tables 1-2. Across all three datasets, we calculated the Cohen's κ values for interrater reliability in categorical outcomes, finding strong consistency for labeling of edema (κ=0.70-1.00), MLS (κ=0.74-1.00), hemorrhagic conversion (κ=0.82-1.0), and IVH (κ=1.00) and moderate reliability for parenchymal hematoma (κ=0.46-0.623) [37]. Full Cohen's κ results may be found within Supplementary Table S6.

For our binary outcomes, we compared performance of Lasso regression, K-nearest neighbor, random forest, multilayer perceptron, and BioClinicalBERT algorithms. In our internal test set of MGB MCA Ischemic strokes, BioClinicalBERT consistently outperformed other classifiers trained with BOW and TF-IDF featurizations. Specifically, BioClinicalBERT achieved high performance in detecting edema (AUPRC 0.98, sensitivity 0.99, specificity: 0.92), MLS (AUPRC 1.00, sensitivity 0.99, specificity 0.99), and hemorrhagic conversion (AUPRC 0.95, sensitivity 0.87, specificity 0.97). BioClinicalBERT detected parenchymal hematoma slightly less well (AUPRC 0.80, sensitivity 0.79, specificity 0.78). Our

RBS for IVH detection yielded respective sensitivities and specificities of 0.94/0.99 in the MGB MCA Ischemic Stroke dataset.

In both external validation datasets BioClinicalBERT discriminated well between reports with edema (AUPRC > 0.94, sensitivity > 0.91, specificity > 0.87), MLS (AUPRC > 0.98, sensitivity > 0.95, specificity > 0.99), hemorrhagic conversion (AUPRC > 0.89, sensitivity > 0.72, specificity > 0.94), and parenchymal hematoma (AUPRC > 0.76, sensitivity > 0.79, specificity > 0.70). Precision recall curves for all datasets, outcomes, and classifiers are included in Figure 2.

We compared BioClinicalBERT to the next-best performing classifier (Lasso regression) using McNemar testing in the MGB MCA Ischemic Stroke derivation cohort. We found statistically significant improvements in the detection of edema (χ^2:12.07, $p < 0.001$), midline shift (χ^2:29.43, $p < 0.001$), and hemorrhagic conversion (χ^2:18.64, $p < 0.001$) using BioClinicalBERT. Performance metrics for these two models are compared side-by-side within Table 3, and further details on McNemar testing are available in Supplementary Tables S7-S11.

Additionally, we observed strong external performance from the RBS for IVH and continuous MLS. The IVH RBS achieved sensitivities > 90% and specificities > 99% across all datasets. McNemar's testing demonstrated this system to outperform BioClinicalBERT (average χ^2: 427.0, $p < 0.001$, Supplementary Table S12). For continuous MLS estimation, the mean absolute error of our RBS was less than 0.2 mm for all datasets (Table 4, Figure 3). In comparing absolute errors of this system to standard regression models, one-sided t-testing revealed that the RBS provided substantial improvements in absolute error over linear regressions trained with both BOW ($t=21.14$, $p < 0.001$) and TF-IDF ($t=22.06$, $p < 0.001$). (Supplementary Tables S13-S14).

Lastly, among traditional machine learning classifiers, we observed the fastest CPU runtimes in Lasso regression and K-nearest neighbor models, which each took < 60 seconds per round of

bootstrapping. Runtimes for BioClinicalBERT on a GPU similarly took < 60 seconds per epoch. Full runtime statistics may be found in Supplementary Tables S15-S16.

Discussion:

NLP is a promising approach to classify important unstructured data for research and clinical purposes [6, 38, 39]. In the present study, we analyzed a variety of NLP methods tailored to specific tasks in neurocritical care. Overall, we observed that a BioClinicalBERT transformer network pretrained on large critical care corpora outperformed other binary classifiers trained with both BOW and TF-IDF featurizations for detection of edema, MLS, hemorrhagic conversion, and parenchymal hematoma after acute ischemic stroke. Specifically, our fine-tuned BioClinicalBERT network discriminated well between reports with and without edema (AUPRC > 0.94), MLS (AUPRC > 0.98), and hemorrhagic conversion (AUPRC > 0.89) across three separate datasets that differed by hospital and ischemic stroke location. We found that performance was not quite as good for parenchymal hematoma (AUPRC > 0.76), which we speculate may be related to variations in how hemorrhage is described and the difficulty in distinguishing between confluent petechiae versus parenchymal hematoma from report alone. Notably, BioClinicalBERT ran efficiently on GPU hardware with runtimes of < 60 seconds per epoch across all outcomes. We observed similar computational efficiency when training our second-highest performing model (Lasso regression) on a standard desktop CPU. Finally, for IVH and continuous MLS detection, we found that customized RBS outperformed more generalized machine learning methods.

The field of neurocritical care is increasingly recognizing the potential for clinical and research applications that leverage the ability of NLP to organize information from unstructured radiologic datasets. In our own work, we previously reported improved detection of ischemic stroke, stroke location, and acuity using a modality-agnostic approach with word embeddings adapted from neurologic texts [4], which notably demonstrated significant mismatches between patient outcomes and officially-

documented diagnostic codes. Our current study expands the scope of these efforts by classifying sequelae of stroke and identifying which features are better suited to machine learning approaches versus expert-tailored RBS. Importantly, given variations in radiologist documentation [13], we have also shown that our current toolkit achieved robust external performance across differing hospital systems and patient populations.

While our results demonstrate successful electronic phenotyping of neuroradiology reports, the full scope of applications for such NLP tools remains to be fully explored. From a research perspective, we believe that our methods can help expedite, scale, and control quality in clinical investigative endeavors by automating or semi-automating the identification of important outcomes in large retrospective datasets. Past work has highlighted the potential for inaccuracy when relying on administrative records such as ICD codes for classifying ground truth hospital outcomes [40, 41]. This mismatch is of direct consequence to stroke research, with Mandava and colleagues notably arguing that the heterogeneous nature of stroke subtypes contributes to mistaken statistical inferences on clinical trial data [42]. Consequently, further development of NLP methodologies such as ours could enable appropriate stratification of retrospective patient information according to critical stroke sequelae, thereby furthering efforts already underway in the field to automatically characterize ischemic stroke phenotypes [14, 15]. Additionally, NLP tools could potentially facilitate timely identification of patients for prospective studies. Improved characterization of stroke from radiologic reports may be used to screen for inclusion or exclusion criteria in clinical trials, which may in turn optimize administrative efficiency [43]. Applying NLP to identify or flag potentially-eligible patients for research studies could decrease labor and costs associated with prospective investigation and lower the barrier for these research endeavors [44]. While automated text mining for trial recruitment has been shown in other disciplines such as oncology [45, 46], our work may help to bring similar NLP strategies into prospective trials for stroke.

We also envision a range of clinical applications for this work in the future. Notably, we feel that NLP tools such as these may be used to safeguard the timely communication of critical radiologic results in hospital settings. While standardized systems for reporting are available in many tertiary care centers, delays in communication between radiologists and bedside providers remain a potential source of error in patient care [47], especially during periods of understaffing including nights, weekends, or holidays. Moreover, given that radiologists in high-income nations are increasingly participating in global health delivery via "teleradiology," our work also has implications for low-resource settings where critical results reporting may be underutilized [48]. Beyond improvements in communication, we also suggest that automated flagging of stroke sequelae could facilitate automated quality improvement initiatives. Escobar and colleagues demonstrated better 30-day outcomes after implementing an automated model identifying concerning data trends that triggered a series of response-protocols [49]. Conceivably, our tools could be similarly used to develop quality improvement initiatives including electronic health record-based warnings such as flagging anticoagulation use in patients with documented hemorrhage on radiology reports or suggested step-wise protocols for level of care escalation in response to cerebral edema. Finally, additional work has shown how medical NLP systems could be potentially integrated with evidence-based guidelines to manage specific conditions identified by imaging reports [50]. Automatic retrieval of such information may allow for immediate review of evidence based guidelines for providers' convenience, rather than relying on the clinician to initiate database searches de novo.

Although the potential applications of NLP may be intriguing, we acknowledge that the necessary technological infrastructure is lacking for immediate clinical translation of our findings. While other studies have suggested pathways toward implementation of NLP systems in neurocritical care settings, including improved prehospital stroke diagnosis [7], prediction of stroke outcome from radiology

reports [17], and automated phenotyping of stroke subclass [15] for treatment recommendations, only a fraction of medical NLP applications reach the level of in-hospital deployment [3]. It is important to recognize that fundamental proofs of principle in machine learning often fall short of effectively providing clinicians with the kind of clinical support systems that they often suggest to be possible [51]. With this perspective, we must state that our present NLP models, although well-performing overall, still require the development of user-friendly interfaces, external validation, and well-designed implementation trials before they can be regarded as ready for patient care [47]. We concede that our work is foundational, and hope that it can serve a guide toward clinician-led directives that leverage technology to serve patients and improve outcomes. At the present time, we advocate that our NLP methods are best viewed as a promising analytic strategy for others conducting "big data" research in neurocritical care rather than a ready tool for clinical use..

We also recognize several methodological limitations within our study. While our manually-labeled dataset of neurologic complications after ischemic stroke is one of the largest in literature to date, we acknowledge that we had a modest sample size and number of unique patients. We also note that our gold-standard labeling of radiology reports reflects a professional interpretation of raw imaging data and may not represent the clinical ground truth of a patient's condition. Additionally, we found lower performance of parenchymal hematoma detection relative to other outcomes, which may be in part due to our annotation strategy. Our labelers assigned parenchymal hematoma occurrence based on interpretation of ECASS-II criteria and not direct specific radiology phrases leading to more heterogeneity and lower interrater-reliability. It is also possible that we overestimated the degree of external validity of our methods. Given that our clinical datasets contain some patients with multiple reports, models may have been trained with linguistically-correlated data, which could potentially limit generalizability to other unseen datasets. Because the two hospital systems used for comparison are

located within the same city in the United States, a substantial number of staff radiologists between the two institutions have received training or experience between both hospitals during their careers. Therefore, it is possible that reporting styles in the MGB and BMC datasets may be more similar than comparators derived from other regions.

Nevertheless, our results suggest that a new generation of NLP models holds promise for characterizing and subtyping large amounts of radiologic text data in ischemic stroke. Future studies may apply our methodology and models to identify cohorts, outcomes of interest and/or covariates for epidemiologic and clinical research in neurocritical care. Further work in this area is necessary to ensure generalizability of these methods for future research and clinical applications.

Conclusion:

Automated NLP approaches detect critical complications of ischemic stroke including edema, MLS, hemorrhagic conversion, parenchymal hematoma and IVH. A pretrained BioClinicalBERT model yielded the highest overall performance in detecting common binary outcomes, and outperformed traditional machine learning classifiers trained with BOW and TF-IDF featurizations, respectively. Our fine-tuned BioClinicalBERT also demonstrated strong external validity, generalizing well from our derivation cohort of MCA territory ischemic strokes to external datasets from a separate institution and to patients presenting with a variety of locations of cerebral infarctions. Our results reinforce the utility of deep attention neural networks in clinical NLP tasks. The presented approach has multiple important clinical and research applications, including scalable electronic phenotyping, augmentation of clinical risk prediction models, and facilitation of automatic alert systems in the hospital setting. Additional research in this area will continue the development of novel strategies for electronic phenotyping in neurocritical care.

References:

1. 1. Foreman, B., *Neurocritical Care: Bench to Bedside (Eds. Claude Hemphill, Michael James) Integrating and Using Big Data in Neurocritical Care.* Neurotherapeutics, 2020. **17**(2): p. 593-605.
2. 2. Casey, A., et al., *A systematic review of natural language processing applied to radiology reports.* BMC medical informatics and decision making, 2021. **21**(1): p. 1-18.
3. 3. Pons, E., et al., *Natural language processing in radiology: a systematic review.* Radiology, 2016. **279**(2): p. 329-343.
4. 4. Ong, C.J., et al., *Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports.* PLoS One, 2020. **15**(6): p. e0234908.
5. 5. Elkins, J.S., et al., *Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review.* Computers and biomedical research, 2000. **33**(1): p. 1-10.
6. 6. Langlotz, C.P., et al., *A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop.* Radiology, 2019. **291**(3): p. 781-791.
7. 7. Mayampurath, A., et al., *Improving Prehospital Stroke Diagnosis Using Natural Language Processing of Paramedic Reports.* Stroke, 2021. **52**(8): p. 2676-2679.
8. 8. Manning, C. and H. Schutze, *Foundations of statistical natural language processing*. 1999: MIT press.
9. 9. Arsava, E.M., et al., *Assessment of the predictive validity of etiologic stroke classification.* JAMA neurology, 2017. **74**(4): p. 419-426.
10. 10. Mozayan, A., et al., *Practical Guide to Natural Language Processing for Radiology.* RadioGraphics, 2021. **41**(5): p. 1446-1453.
11. 11. Friedlin, J. and C.J. McDonald. *A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports*. in *AMIA annual symposium proceedings*. 2006. American Medical Informatics Association.
12. 12. Hinton, G., *Deep learning—a technology with the potential to transform health care.* Jama, 2018. **320**(11): p. 1101-1102.
13. 13. Li, M., et al., *Analysis of stroke detection during the COVID-19 pandemic using natural language processing of radiology reports.* American Journal of Neuroradiology, 2021. **42**(3): p. 429-434.
14. 14. Garg, R., et al., *Automating ischemic stroke subtype classification using machine learning and natural language processing.* Journal of Stroke and Cerebrovascular Diseases, 2019. **28**(7): p. 2045-2051.
15. 15. Guan, W., et al., *Automated Electronic Phenotyping of Cardioembolic Stroke.* Stroke, 2021. **52**(1): p. 181-189.
16. 16. Castro, V.M., et al., *Large-scale identification of patients with cerebral aneurysms using natural language processing.* Neurology, 2017. **88**(2): p. 164-168.
17. 17. Heo, T.S., et al., *Prediction of Stroke Outcome Using Natural Language Processing-Based Machine Learning of Radiology Report of Brain MRI.* Journal of personalized medicine, 2020. **10**(4): p. 286.

18. 18. Bacchi, S., et al., *Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations.* Stroke, 2019. **50**(3): p. 758-760.
19. 19. Zech, J., et al., *Natural language–based machine learning models for the annotation of clinical radiology reports.* Radiology, 2018. **287**(2): p. 570-580.
20. 20. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
21. 21. Larrue, V., et al., *Risk factors for severe hemorrhagic transformation in ischemic stroke patients treated with recombinant tissue plasminogen activator: a secondary analysis of the European-Australasian Acute Stroke Study (ECASS II).* Stroke, 2001. **32**(2): p. 438-441.
22. 22. Harris, P.A., et al., *Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support.* Journal of biomedical informatics, 2009. **42**(2): p. 377-381.
23. 23. Vaswani, A., et al. *Attention is all you need*. in *Advances in neural information processing systems*. 2017.
24. 24. Lee, J., et al., *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics, 2020. **36**(4): p. 1234-1240.
25. 25. Johnson, A.E., et al., *MIMIC-III, a freely accessible critical care database.* Scientific data, 2016. **3**(1): p. 1-9.
26. 26. Alsentzer, E., et al., *Publicly available clinical BERT embeddings.* arXiv preprint arXiv:1904.03323, 2019.
27. 27. Franklin, J., *The elements of statistical learning: data mining, inference and prediction.* The Mathematical Intelligencer, 2005. **27**(2): p. 83-85.
28. 28. Cover, T. and P. Hart, *Nearest neighbor pattern classification.* IEEE transactions on information theory, 1967. **13**(1): p. 21-27.
29. 29. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
30. 30. Chaudhuri, B. and U. Bhattacharya, *Efficient training and improved performance of multilayer perceptron in pattern classification.* Neurocomputing, 2000. **34**(1-4): p. 11-27.
31. 31. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
32. 32. Darby, D.G., et al., *Primary intraventricular hemorrhage: clinical and neuropsychological findings in a prospective stroke series.* Neurology, 1988. **38**(1): p. 68-68.
33. 33. Foroushani, H.M., et al., *Accelerating prediction of malignant cerebral edema after ischemic stroke with automated image analysis and explainable neural networks.* Neurocritical Care, 2021: p. 1-12.
34. 34. Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PloS one, 2015. **10**(3): p. e0118432.
35. 35. Dietterich, T.G., *Approximate statistical tests for comparing supervised classification learning algorithms.* Neural computation, 1998. **10**(7): p. 1895-1923.
36. 36. RA, A., *When to use the Bonferroni corrrection.* Ophthalmic Physial Opt, 2014. **34**: p. 502-508.
37. 37. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* biometrics, 1977: p. 159-174.
38. 38. Cai, T., et al., *Natural language processing technologies in radiology research and clinical applications.* Radiographics, 2016. **36**(1): p. 176-191.

39. 39. Sorin, V., et al., *Deep learning for natural language processing in radiology—fundamentals and a systematic review.* Journal of the American College of Radiology, 2020. **17**(5): p. 639-648.
40. 40. Jetté, N., et al., *How accurate is ICD coding for epilepsy?* Epilepsia, 2010. **51**(1): p. 62-69.
41. 41. Quan, H., 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): p. 1424-1441.
42. 42. Mandava, P., et al., *A critical review of stroke trial analytical methodology: outcome measures, study design, and correction for imbalances.* Translational Stroke Research, 2012: p. 833-861.
43. 43. Elkins, J.S., et al., *Recruiting subjects for acute stroke trials: a meta-analysis.* Stroke, 2006. **37**(1): p. 123-128.
44. 44. Woo, M., *An AI boost for clinical trials.* Nature, 2019. **573**(7775): p. S100-S100.
45. 45. Liu, R., et al., *Evaluating eligibility criteria of oncology trials using real-world data and AI.* Nature, 2021. **592**(7855): p. 629-633.
46. 46. Alexander, M., et al., *Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients.* JAMIA open, 2020. **3**(2): p. 209-215.
47. 47. Murphy, D.R., H. Singh, and L. Berlin, *Communication breakdowns and diagnostic errors: a radiology perspective.* Diagnosis, 2014. **1**(4): p. 253-261.
48. 48. Mollura, D.J., et al., *2016 RAD-AID conference on international radiology for developing countries: gaps, growth, and United Nations sustainable development goals.* Journal of the American College of Radiology, 2017. **14**(6): p. 841-847.
49. 49. Escobar, G.J., et al., *Automated identification of adults at risk for in-hospital clinical deterioration.* New England Journal of Medicine, 2020. **383**(20): p. 1951-1960.
50. 50. Do, B.H., et al., *Automatic retrieval of bone fracture knowledge using natural language processing.* Journal of digital imaging, 2013. **26**(4): p. 709-713.
51. 51. Citerio, G., *Big Data and Artificial Intelligence for Precision Medicine in the Neuro-ICU: Bla, Bla, Bla.* Neurocritical Care, 2022: p. 1-3.

Tables and Figures

Table 1. Cohort Characteristic by Individual

	MGB MCA Ischemic Stroke (N=550)	**BMC MCA Ischemic Stroke** (N=154)	**MGB Any Territory Ischemic Stroke** (N=214)
Demographics			
Age (years) – Mean (SD)	68.3 (14.2)	63.6 (17.1)	68.6 (15.0)
Female – N (%)	265 (48.2)	75 (48.7)	86 (40.2)
Clinical Outcomes			
Edema – N (%)	420 (76.3)	101 (65.6)	88 (41.1)
MLS (Binary)– N (%)	223 (40.5)	47 (30.5)	24 (11.2)
Hemorrhagic Conversion ^ – N (%)	260 (47.2)	45 (29.2)	63 (29.4)
Parenchymal Hematoma – N (%)	81 (31.2)	11 (24.4)	11 (17.5)
Petechial Hemorrhage – N (%)	164 (63.1)	34 (75.6)	47 (74.6)
Indeterminate – N (%)	15 (5.8)	0 (0.0)	5 (7.9)
Intraventricular Hemorrhage – N (%)	39 (7.1)	6 (3.9)	10 (4.7)

Table 1. Cohort Characteristics by Individual For unique patients, we report demographics, report characteristics, and clinical outcomes are illustrated in tabular form for the MGB MCA Ischemic Stroke, BMC MCA Ischemic Stroke, and MGB Any Territory Ischemic stroke datasets.
Abbreviations: N – *Number of patients,* ***MLS:*** *Midline Shift*
^ indicates hemorrhagic subtypes calculated as a percentage of patients/reports with hemorrhagic conversion. For patients with varying severities of hemorrhagic conversion, we report the most severe subtype observed.

Table 2. Cohort Characteristic by Report

	MGB MCA Ischemic Stroke (M=2289)	BMC MCA Ischemic Stroke (M=377)	MGB Any Territory Ischemic Stroke (M=520)
Imaging Modality			
CT –M (%)	1846 (80.6)	225 (59.7)	276 (53.1)
CT Angiography –M (%)	97 (4.2)	63 (16.7)	8 (1.5)
MRI/MRA –M (%)	346 (15.2)	89 (23.6)	236 (45.4)
Clinical Outcomes			
Edema – N (%)	1534 (67.0)	233 (61.8)	179 (34.4)
MLS (Binary)– N (%)	623 (27.2)	126 (33.4)	47 (9.0)
MLS millimeters – Mean (SD)*	7.2 (4.1)	6.8 (3.9)	6.1 (3.7)
Hemorrhagic Conversion ^ – N (%)	819 (35.8)	86 (22.8)	119 (22.9)
Parenchymal Hematoma – N (%)	277 (33.8)	23 (26.7)	15 (12.6)
Petechial Hemorrhage – N (%)	463 (56.5)	58 (67.4)	89 (74.8)
Indeterminate – N (%)	79 (9.6)	5 (5.8)	15 (12.6)
Intraventricular Hemorrhage – N (%)	109 (4.8)	12 (3.2)	18 (3.5)

Table 2. Cohort Characteristics by Report Given that a single patient may have multiple radiology reports included within the analysis, we provide baseline information across all individual stroke reports. For all reports, we provide diagnostic modalities and clinical outcomes for the MGB MCA Ischemic Stroke, BMC MCA Ischemic Stroke, and MGB Any Territory Ischemic stroke datasets.
Abbreviations: CT - *Computed Tomography****, N-****Number of Reports,* ***M -*** *Number of Reports,* ***MLS -****Midline Shift,* ***MRA -*** *Magnetic Resonance Angiography,* ***MRI -*** *Magnetic Resonance Imaging.*
*******Only indicates reports with MLS present, ^ indicates hemorrhagic subtypes calculated as a percentage of patients/reports with hemorrhagic conversion.*

Table 3: BioClinicalBERT vs. Lasso Regression

		BioClinicalBERT				Lasso				
		AUPRC	Sensitivity	Specificity	Precision (PPV)	AUPRC	Sensitivity	Specificity	Featurization	
MGB MCA Ischemic Stroke	Edema	0.98 (0.97, 1.00)	0.99 (0.98, 0.99)	0.92 (0.89,0.96)	0.98 (0.96,0.98)	0.98 (0.98, 0.99)	0.87 (0.84, 0.89)	0.84 (0.89, 0.98)	0.97 (0.95, 0.99)	TF-IDF
	MLS	1.00 (1.00, 1.00)	0.99 (0.98, 1.00)	0.99 (0.98, 1.00)	0.97 (0.95, 0.99)	0.95 (0.92, 0.98)	0.81 (0.77, 0.86)	0.96 (0.94, 0.98)	0.90 (0.86, 0.95)	BOW
	Hemorrhagic Conversion	0.95 (0.93, 0.97)	0.89 (0.85, 0.93)	0.97 (0.95, 0.99)	0.94 (0.91, 0.97)	0.89 (0.86, 0.91)	0.80 (0.75, 0.84)	0.89 (0.85, 0.92)	0.81 (0.75, 0.87)	BOW
	Parenchymal Hematoma	0.80 (0.77, 0.84)	0.79 (0.75, 0.83)	0.78 (0.76, 0.79)	0.62 (0.52, 0.65)	0.64 (0.58, 0.70)	0.75 (0.59, 0.90)	0.75 (0.60, 0.89)	0.70 (0.50, 0.89)	BOW
	IVH	0.66 (0.48, 0.83)	0.57 (0.41, 0.73)	0.99 (0.98, 0.99)	0.71 (0.53, 0.89)	0.76 (0.67, 0.84)	0.37 (0.13, 0.60)	0.99 (0.99, 1.00)	0.94 (0.80, 1.00)	BOW
BMC MCA Ischemic Stroke	Edema	0.99 (0.99, 1.00)	0.98 (0.97, 0.99)	0.88 (0.84, 0.92)	0.93 (0.91, 0.95)	0.98 (0.97, 0.98)	0.76 (0.74, 0.78)	0.96 (0.95, 0.97)	0.97 (0.97, 0.98)	TF-IDF
	MLS	1.00 (1.00, 1.00)	0.99 (0.97, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	0.80 (0.80, 0.81)	0.81 (0.79, 0.83)	0.84 (0.83, 0.85)	0.69 (0.68, 0.71)	TF-IDF
	Hemorrhagic Conversion	0.89 (0.85, 0.93)	0.72 (0.69, 0.76)	0.99 (0.98, 1.00)	0.96 (0.93, 0.98)	0.75 (0.74, 0.76)	0.61 (0.58, 0.63)	0.96 (0.95, 0.96)	0.82 (0.79, 0.84)	TF-IDF
	Parenchymal Hematoma	0.78 (0.75, 0.82)	0.79 (0.75, 0.84)	0.74 (0.70, 0.78)	0.55 (0.49, 0.61)	0.43 (0.41, 0.45)	0.24 (0.20, 0.29)	0.97 (0.96, 0.97)	0.53 (0.47, 0.59)	TF-IDF
	IVH	0.59 (0.23, 0.96)	0.54 (0.09, 1.00)	0.99 (0.97, 1.00)	0.59 (0.14, 1.00)	0.82 (0.76, 0.87)	0.15 (0.06, 0.24)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	TF-IDF
MGB Any Territory Ischemic Stroke	Edema	0.94 (0.91, 0.97)	0.91 (0.88, 0.94)	0.97 (0.95, 0.98)	0.93 (0.90, 0.96)	0.89 (0.89, 0.90)	0.64 (0.62, 0.66)	0.97 (0.97, 0.98)	0.93 (0.91, 0.93)	TF-IDF
	MLS	0.98 (0.95, 1.00)	0.96 (0.92, 1.00)	1.00 (1.00, 1.00)	0.99 (0.97, 1.00)	0.82 (0.80, 083)	0.89 (0.88, 0.90)	0.96 (0.95, 0.96)	0.67 (0.64, 0.71)	TF-IDF
	Hemorrhagic Conversion	0.90 (0.86, 0.93)	0.78 (0.73, 0.83)	0.94 (0.92, 0.96)	0.81 (0.75, 0.87)	0.77 (0.76, 0.79)	0.66 (0.64, 0.68)	0.92 (0.92, 0.93)	0.73 (0.71, 0.75)	TF-IDF
	Parenchymal Hematoma	0.76 (0.69, 0.84)	0.87 (0.86, 0.89)	0.71 (0.70, 0.72)	0.33 (0.31, 0.35)	0.37 (0.34, 0.40)	0.14 (0.11, 0.17)	0.94 (0.94, 0.94)	0.52 (0.48, 0.56)	TF-IDF
	IVH	0.89 (0.67, 1.00)	1.00 (1.00, 1.00)	0.97 (0.95, 1.00)	0.57 (0.40, 0.74)	0.94 (0.89, 0.99)	0.94 (0.94, 0.94)	0.99 (0.96, 1.00)	0.74 (0.36, 1.00)	TF-IDF

Table 3: BioClinicalBERT vs. Lasso Regression We demonstrate comparative performance statistics (AUPRC, sensitivity, specificity, and precision) for BioClinicalBERT and Lasso regression for binary classification tasks. Across three datasets and four outcomes (edema, hemorrhagic conversion, midline shift, and parenchymal hematoma), BioClinicalBERT demonstrated the greatest balanced performance while Lasso regression generally outperformed other machine learning classifiers. For each dataset and outcome, Lasso performance is shown according to the text featurization algorithm which yielded the highest AUPRC value. Full model results across all classifiers and featurizations may be found in Supplemental Tables S2-S6.
Abbreviations: BMC - *Boston Medical Center,* ***IVH -*** *Intraventricular Hemorrhage,* ***MCA -****Middle Cerebral Artery****,* ***MGB -*** *Massachusetts General Brigham and Women's,* ***MLS*** *- Midline Shift,* ***PPV*** *– Positive Predictive Value*

Table 4: RBS Performance

		Sensitivity	Specificity	Precision	Accuracy	MAE (mm)
MGB MCA Ischemic Stroke	MLS	0.95 (0.94, 0.96)	0.95 (0.95, 0.97)	0.92 (0.91, 0.93)	0.97 (0.96, 0.97)	0.18 (0.16, 0.20)
	IVH	0.94 (0.92, 0.95)	0.99 (0.99, 0.99)	0.85 (0.82, 0.88)	0.99 (0.98, 0.99)	--
BMC MCA Ischemic Stroke	MLS	0.96 (0.95, 0.96)	0.96 (0.95, 098)	0.95 (0.94, 0.95)	0.98 (0.97, 0.98)	0.14 (0.13, 0.16)
	IVH	0.91 (0.89, 0.93)	0.99 (0.99, 1.00)	0.83 (0.80, 0.85)	0.99 (0.99, 1.00)	--
MGB Any Territory Ischemic Stroke	MLS	0.95 (0.94, 0.96)	0.95 (0.93, 0.99)	0.95 (0.94, 0.96)	0.99 (0.98, 1.00)	0.03 (0.03, 0.04)
	IVH	0.94 (0.92, 0.96)	0.99 (0.99, 1.00)	0.85 (0.83, 0,87)	0.99 (0.99, 1.00)	--

Table 4: RBS Performance Rule-based systems were developed for the detection of intraventricular hemorrhage as well as the quantification of midline shift. In the case of our quantitative RBS for MLS, correct numerical detection was converted to true and false positives and negatives. Thus, for both outcomes, we report sensitivity, specificity, precision, and accuracy. Mean absolute error in millimeters is reported in the final column for the quantitative MLS system.
Abbreviations: BMC - *Boston Medical Center,* ***IVH -*** *Intraventricular Hemorrhage,* ***MCA -****Middle Cerebral Artery****,* ***MGB -*** *Massachusetts General Brigham and Women's,* ***MLS*** *- Midline Shift,* ***RBS -*** *Rule-Based System*

Figure 1. Model Testing and Development Framework

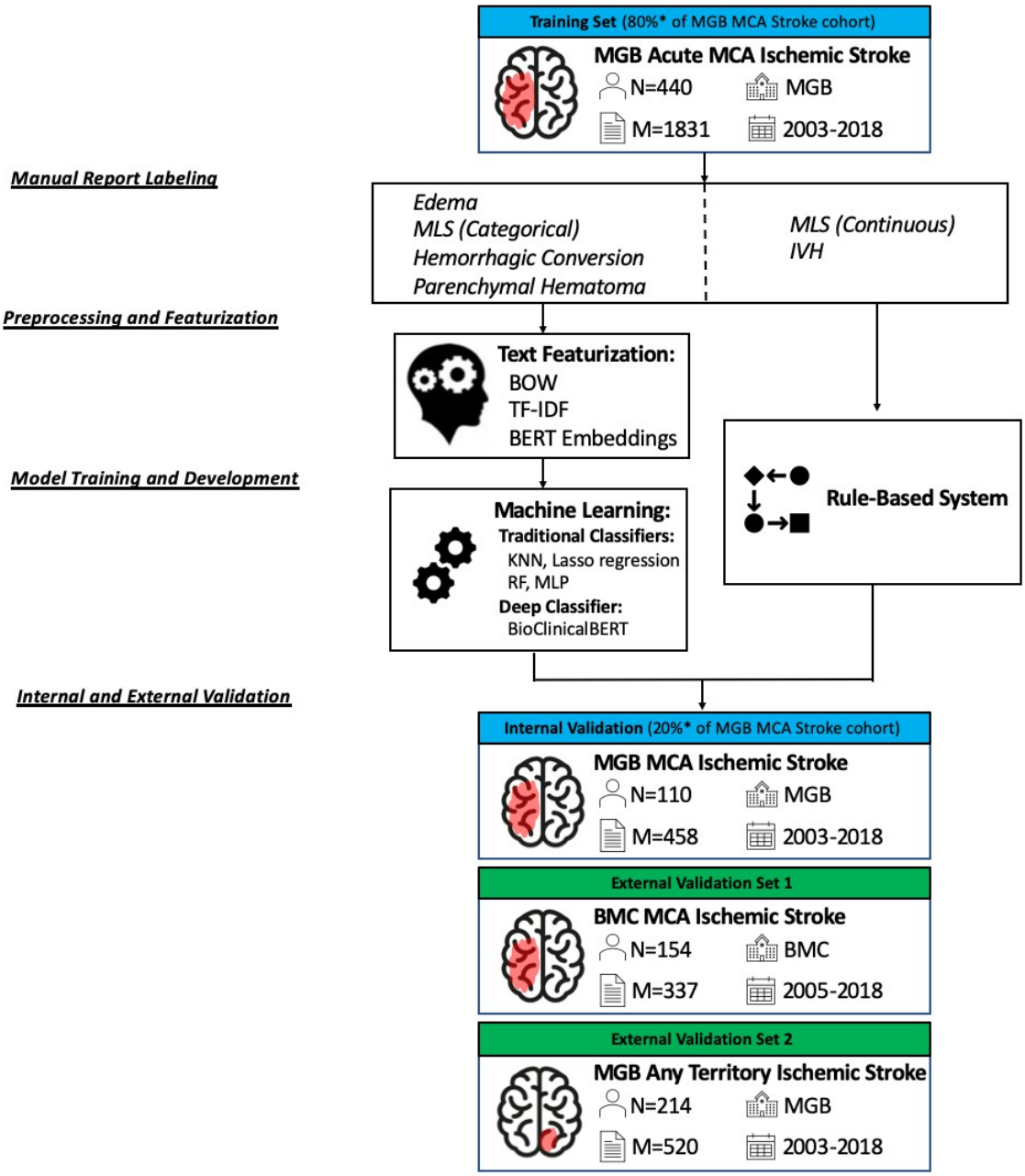


Figure 2. Precision (Positive Predictive Value) Recall Curves

Figure 3: RBS for Quantitative Midline Shift.

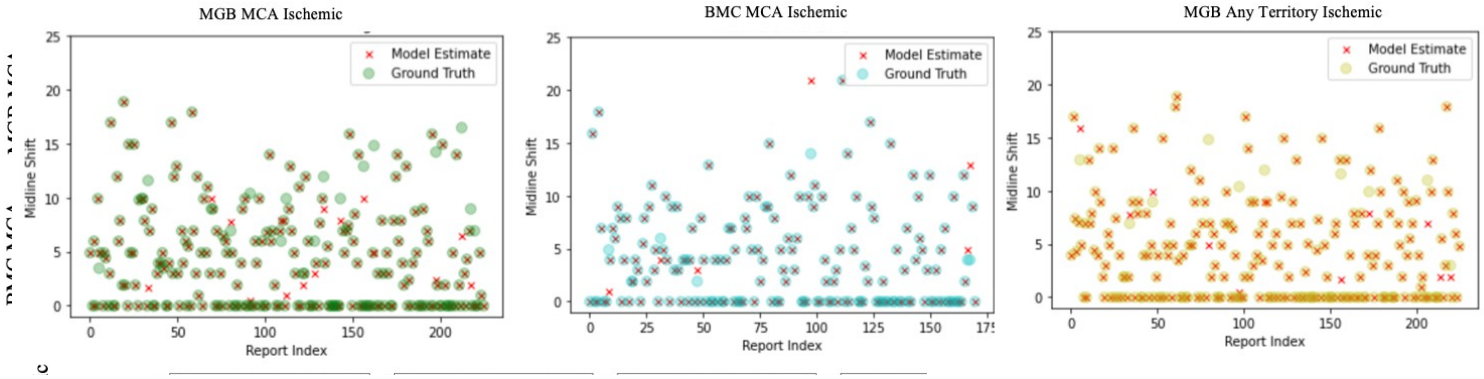


Figure Captions

Figure 1. Model Testing and Development Framework A graphic display of data labelling, preprocessing, featurization, and model training and development is illustrated.
Abbreviations: BMC - *Boston Medical Center,* ***BOW -*** *Bag of Words,* ***IVH -*** *Intraventricular Hemorrhage,* ***KNN -*** *K-Nearest Neighbor,* ***M -*** *Number of Reports,* ***MCA -*** *Middle Cerebral Artery,* ***MGB -*** *Massachusetts General Brigham and Women's,* ***MLS*** *- Midline Shift,* ***MLP*** *- Multilayer Perceptron,* ***N -*** *Number of patients,* ***RF -*** *Random Forest,* ***TF-IDF -*** *Term Frequency-Inverse Document Frequency*

Figure 2: Precision Recall Curves Precision recall (PR) curves are demonstrated for binary classification of edema, midline shift, hemorrhagic conversion, and parenchymal hematoma* per ECASS-II criteria. Results are demonstrated for BioClinicalBERT, Lasso regression, Random Forest, Multilayer Perceptron, and K-Nearest Neighbors algorithms. For all models other than BioClinicalBERT, the PR curve is shown for the text featurization algorithm (BOW or TF-IDF) that maximized the AUPRC value. PR curves, unlike ROC curves, are influenced by class imbalance as well as performance metrics (sensitivity, specificity, and precision) that are reflective of clinical heuristics. This allows notable differences in binary classification to stand out where ROC analysis might otherwise not highlight salient differences. **Note**: *Indicates parenchymal hematoma only trained using reports with ground truth-labeled hemorrhagic conversion.
Abbreviations: PR – *Precision Recall,* ***BOW –*** *Bag of Words,* ***TF-IDF –*** *Term Frequency Inverse Document Frequency,* ***ROC*** *– Receiver Operating Characteristic curve*

Figure 3: RBS for Quantitative Midline Shift Scatterplots are demonstrated with overlap of RBS-detected midline shift ("Model Estimates") and ground truths from CT and MRI reports in three separate cohorts. Points representing individual radiology reports are randomly selected from each cohort for visualization and arranged according to their order ("index") within the overall sample. Significant concordance is seen between RBS-detected MLS and ground truths. Full metrics for RBSs, including MLS, are found in Table 3.

Abbreviations: BMC - *Boston Medical Center,* ***MCA -*** *Middle Cerebral Artery****, MGB -*** *Massachusetts General Brigham and Women's,* ***MLS*** *- Midline Shift,* ***RBS -*** *Rule-Based System*