scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The stroke outcome optimization project: Acute ischemic strokes from a comprehensive stroke center

John Absher[1,2,3] ✉, Sarah Goncher[1], Roger Newman-Norlund[4], Nicholas Perkins[1,2,3], Grigori Yourganov[5], Jan Vargas[1,2,3,8], Sanjeev Sivakumar[1,3,8], Naveen Parti[1,3,8], Shannon Sternberg[3,8], Alex Teghipco[4], Makayla Gibson[4], Sarah Wilson[6], Leonardo Bonilha[7] & Chris Rorden[4]

Stroke is a leading cause of disability, and Magnetic Resonance Imaging (MRI) is routinely acquired for acute stroke management. Publicly sharing these datasets can aid in the development of machine learning algorithms, particularly for lesion identification, brain health quantification, and prognosis. These algorithms thrive on large amounts of information, but require diverse datasets to avoid overfitting to specific populations or acquisitions. While there are many large public MRI datasets, few of these include acute stroke. We describe clinical MRI using diffusion-weighted, fluid-attenuated and T1-weighted modalities for 1715 individuals admitted in the upstate of South Carolina, of whom 1461 have acute ischemic stroke. Demographic and impairment data are provided for 1106 of the stroke survivors from this cohort. Our validation demonstrates that machine learning can leverage the imaging data to predict stroke severity as measured by the NIH Stroke Scale/Score (NIHSS). We share not only the raw data, but also the scripts for replicating our findings. These tools can aid in education, and provide a benchmark for validating improved methods.

Background & Summary

Stroke is a leading cause of long-term disability in the United States. Despite a decrease in stroke incidence per year of life among the elderly, this decline is counteracted by extended life expectancies and significant upswings in occurrences among younger adults. When paired with effective acute interventions that enhance survival, the overall result is a growing number of individuals living with stroke-related impairments[1]. Our overarching goal is to provide a large, public and diverse dataset that combines the medical imaging that is typical of acute stroke management along with demographic and impairment measures. These datasets can aid developers of tools to map brain injury, determine residual brain health, and create reliable diagnostic and prognostic measures.

Many "risk scores" have been developed to estimate the impact of acute ischemic stroke (AIS) and intracranial hemorrhage (ICH) in individual patients[2–5] and 94 individual de novo Clinically Predictive Models (CPMs) of stroke outcome were found in the Tufts PACE Clinical Prediction Model Registry (as of July 26, 2023)[6]. Stroke risk and stroke outcomes are influenced by many factors such as sex, racial/ethnic group, and socioeconomic class[6,7]. Validated CPMs for stroke are needed for predicting various outcomes, including overall functional recovery and the development of vascular dementia[6,8].

Prisma Health-Upstate has been collecting Get with the Guidelines (GWTG-stroke) data on all acute strokes seen at Greenville Memorial Hospital since 2009. GWTG data includes basic demographic information (age, race, sex), zip code, stroke etiology, vitals and blood work at admission and discharge, medical history and current medications, time to and type of thrombolytic therapy administered, complications, in-patient treatment

[1]University of South Carolina School of Medicine, Greenville, SC, 29605, USA. [2]Clemson University School of Health Research, CUSHR, Clemson, SC, 29634, USA. [3]Departments of Medicine, Neurosurgery, and Radiology, Prisma Health, Greenville, SC, 29601, USA. [4]Department of Psychology, University of South Carolina, Columbia, SC, 29203, USA. [5]Partnership for an Advanced Computing Environment, Georgia Institute of Technology, Atlanta, GA, 30332, USA. [6]Linguistics Program, University of South Carolina, Columbia, SC, 29203, USA. [7]Department of Neurology, University of South Carolina, Columbia, SC, 29208, USA. [8]These authors contributed equally: Jan Vargas, Sanjeev Sivakumar, Naveen Parti, Shannon Sternberg. ✉e-mail: absher@mailbox.sc.edu

and interventions, lifestyle interventions, stroke interventions, and neuroimaging data[9]. Because the GWTG data also includes two measures of post-stroke function, the NIH stroke scale (NIHSS) and modified Rankin Scale (mRS), we used these as the primary indicators of stroke severity for this dataset. However, we would note that each of these scales has its own limitations, and the best metrics of stroke impact should ideally be composed of data from multiple assessments, and include instructions for central adjudication, consistent rater training and correct application of novel statistical techniques[10]. More than 15,000 subjects are included in our local GWTG dataset. GWTG is used to promote the quality of hospital stroke care by tracking key processes and demographics known to relate to favorable outcomes, such as door-to-needle time or whether an AIS patient received thrombolytic[11]. Approximately 1100–1300 unique subjects are entered into our comprehensive stroke centers (CSC) GWTG database each year. GWTG data are important for understanding AIS outcomes, because they capture the systems of care that are influential in such outcomes. Collection and distribution of the data described in the SOOP repository is approved under protocol Pro00078716 of the Prisma Health Committee A (initial approval 10/29/2018, status = ongoing). Notably, the informed consent requirement for the current retrospective data analysis was waived by this Institutional Review Board. SOOP participants treated by study investigators may be recontacted and consented in person or remotely with the help of a legally authorized representative if needed, making future, prospective, longitudinal investigations a possibility. There are many distinct AIS outcomes. For example, weakness, numbness, visual loss, and cognitive dysfunction often result from stroke. Interactions among these factors may vary among subjects with an isolated, first-time AIS, compared to those individuals with recurrent or multifocal AIS. Also, stroke related cognitive impairments occur in up to a third of stroke patients[12,13]. Relatively few studies have attempted to predict specific outcomes such as motor function, aphasia, neglect, and depression[14–16]. Regardless of the measure of interest, a comprehensive approach to outcome prediction requires consideration of both known and unknown predictor variables and confounds–the quality of AIS care processes, patient characteristics, structural and functional consequences of acute and pre-existing brain damage, rehabilitation strategies, resilience factors, and other influences[17–21]. For example, age at stroke[22,23] and exercise are correlated with recovery, and factors like age, lesion volume, and residual brain health synergistically predict outcome. Large data repositories may capture information on a broad range of known and unknown outcome predictors and confounders to promote public health research[22,23]. Consequently, there is growing interest in data sharing consortia for AIS[24,25]. Such large data sharing collaboratives have been valuable adjuncts to understanding many diseases and disorders.

As an example, members of our team have experience applying machine learning to predict stroke sequelae and recovery trajectories in chronic and acute stroke. While these approaches have demonstrated potential, their effectiveness is often tempered by limitations such as small sample sizes, which can lead to overfitting or underfitting depending on the complexity of the algorithm used. Moreover, each algorithm carries inherent strengths and weaknesses that must be carefully considered to optimize model performance and ensure generalizability across diverse patient populations. The results to date demonstrate AIS outcome classification that is statistically significant, but insufficiently rigorous to change the standard of care for individuals.

While the majority of our group's research has focused on diagnostic aspects of acute and chronic stroke, such as the relationship between lesion size, lesion location and chronic impairment, we acknowledge the critical importance of prognostic studies in predicting long-term outcomes on the basis of data available in the acute stage. Indeed, early prediction of likely recovery trajectories and chronic outcome is of paramount importance to stroke survivors, as well as their caregivers, as it can provide both a roadmap for future recovery as well as a set of expectations/limits in which to frame treatment outcomes. Accurate prediction of long-term outcomes likely requires the creation of comprehensive models that considers multiple factors, including lesion size and location and overall brain health along with various data such as demographic, medical, health and lifestyle factors. Additionally, it may be that the relative prognostic and diagnostic value of each of these factors differs in the acute and chronic stages of stroke recovery. Ultimately, gaining a better understanding of these interactions may be informative to clinicians and rehabilitation scientists striving to understand and manage stroke across time.Therefore, we combine GWTG quality data, clinical data from the electronic health record, and magnetic resonance imaging (MRI) morphometry data to examine a large population of AIS subjects from a large comprehensive stroke center (CSC). This paper introduces our efforts to develop a reproducible, sharable AIS CPM. An AIS data sharing consortium using these or similar methods could vastly improve outcome predictions in acute stroke[26].

The term prognosis implies a type of clinical prediction model (CPM) that clinicians and families value immensely. AIS prognosis has both clinical and research significance. Stroke patients and their families are eager to know how they will recover, and how likely it is that long-term consequences like vascular dementia may develop. Likely, therapy teams rely on prognostic information to tailor their approach, deciding on the intensity and type of rehabilitation efforts. These can range from strategies aimed at restoring lost functions (rehabilitation) to those designed to help patients adapt to impairments through alternative techniques (compensatory strategies). Both forms are integral parts of the rehabilitation process, and the decision to emphasize one over the other, or how to effectively combine them, is heavily informed by an individual's predicted recovery trajectory. Prognosis guides expectations and suggests the approaches we recommend clinically for every individual with AIS. Perfect prognostic information would enable research teams to detect meaningful effects of treatment interventions with smaller sample sizes, thus accelerating and economizing clinical trials. We focus in this report on our initial efforts to develop a CPM for prediction of stroke impairment using a large AIS population from a single CSC.

Publicly sharing clinical data can empower discoveries by scientists who do not have access to medical data. Further, datasets can also be aggregated to improve performance and overcome problems with local overfitting. Open datasets can also aid in education as well as providing shared benchmarks for validating and comparing competing solutions. Our SOOP dataset is similar to two other recent shared datasets. While we emphasize the differences, we note the potential for using these large, curated datasets synergistically. The ISLES 2022 dataset

includes MR images and lesion maps from 400 stroke survivors. All individuals are from Europe, with consequences on training diversity[27]. The images are already completely brain extracted, which might limit methods that attempt to model image intensity homogeneity biases, as well as developing robust methods that can cope with diverse features such as wide diploic spaces and post bregmatic dips. The data also lacks demographic details beyond age, which limits the utility to developing automated lesion identification. Liu *et al.*[28] provide acute imaging and demographic data from 2888 individuals from the state of Maryland in the USA, capturing a more diverse population. This dataset also includes rich demographic and outcome measures. However, a limitation of this dataset is that the distribution the dataset is released as a restricted-use collection under a Data Use Agreement (DUA) which requires collaboration with a data review board and restrictions on data handling (e.g. data must be contained on an external drive where the computer is disconnected from the internet during all analyses). These restrictions limit the ability to use this dataset in many educational settings. In contrast to these existing works, our Stroke Outcome Optimization Project (SOOP) provides truly open imaging data from stroke survivors as well as similar data from individuals where stroke was excluded. Also, we provide both demographic measures as well as popular acute measures of stroke impairment and quality metrics.

While BIDS-capable pipelines exist for data from neurologically healthy adults, the presence of stroke can disrupt spatial normalization of imaging data[29]. Beyond providing the normalized acute stroke MRIs and associated clinical data, we also describe, validate and share a full processing pipeline that imports clinical data stored in the emerging BIDS-format for data sharing and generates impairment predictions. Adding data processed through this pipeline from additional stroke datasets may expand the range of impairments and outcomes that may be predicted.

Methods

Ethical statement. This retrospective evaluation of GWTG data, electronic health records extracts, and imaging data was approved by the local ethics boards. The dataset is considered exempt based on the retrospective nature of the study and the rigorous patient de-identification. Approval for re-contact and prospective examination of survivors from this cohort and their care partners has also been obtained. IRB approval for collection of the data contained in this repository was obtained from Prisma Health Committee A, Greenville SC (Pro00078716, 10-29-2018).

Cohort. Our study sample included individuals captured within the GWTG database at Prisma Health-Upstate from the start of 2019 through the end of 2020, representing all identified acute stroke encounters over the entire two-year period. All participants included in this study were exempt from informed consent prior to participation, in accordance with approval received from the Institutional Review Board. Exclusion criteria were then applied. Individuals with subarachnoid, subdural, or intracerebral hemorrhage were excluded. Individuals lacking brain MRI were excluded. Individuals with stroke mimics, transient ischemic attacks, or other confounding structural or functional brain disorders (e.g., brain tumor, refractory epilepsy) were also excluded. The final sample included all eligible individuals with AIS, deemed unlikely to have significant major comorbidities to common clinical sequelae of stroke that could adversely impact outcome ($n = 1415$). Out of 1415 total participants, 305 had large-artery atherosclerosis, 343 had cardioembolism (e.g. atrial fibrillation/flutter, prosthetic heart valve, recent mI), 107 had small-vessel disease (e.g. subcortical, brain stem or lacunar infarct < 1.5 cm), 80 had a stroke of other determined etiology. In our study, 526 cases were reported as cryptogenic strokes, indicating that despite thorough diagnostic evaluations, no definitive cause could be identified. Additionally, 54 participants were classified as unable to determine (UTD), reflecting instances where insufficient documentation or inconclusive evidence prevented any stroke etiology classification. We also included 254 individuals where stroke was initially suspected (requiring the same imaging as for stroke) but later excluded as a probable diagnosis. Behavioral and demographic data were available and are provided for 1106 stroke survivors, with details including gender, age, race, body mass index, NIH stroke scale, mortal status, and acute Modified Rankin Scale. Speech and language pathology findings for the Western Aphasia Battery are provided for each of the subjects for whom this information is available. Of these, medical records listed 784 as white, 257 as black or African American, with 538 women and 568 men, age (after being limited to individuals aged 89 or less due to privacy concerns associated with distribution of age data above this value[30]) ranges from 16 to 89 years with a median of 65, mean of 64.8 and a standard deviation of 14. NIHSS scores ranged from 0 to 30 with a median of 5 and a mean of 8 (standard deviation of 7.90).

Magnetic resonance imaging data. MRI scans for each person were completed within 30 days following their admission to the hospital. Most were obtained within 48 hours days of the acute stroke. For each individual we selected the T1-weighted, T2-weighted, Fluid Attenuated Inversion Recovery (FLAIR), and diffusion sequence that provided the best brain coverage and signal to noise ratio. The diffusion sequence included a TRACE image as well as an image with a contrast similar to an apparent diffusion coefficient (ADC). However, scan settings varied greatly between individuals, with sequence details stored in the text-based BIDS-format 'sidecar' provided with each Neuroimaging Informatics Technology Initiative (NIFTI) format image. In particular the T1w-modality varied tremendously, both in terms of coverage, resolution and contrast. We note that the clinically useful Gadolinium-enhanced T1w scans differ considerably from typical unenhanced sequences popular with basic science.

MRI images were converted from Digital Imaging and Communications in Medicine (DICOM) format to NIfTI format using dcm2niix[31]. We extended the 'spm_deface' script included with the latest version of Statistical Parametric Mapping software (SPM12)[32] to remove identifiable features from the face and neck. While some teams distribute images after complete brain extraction, we intentionally share images that include the scalp. Our rationale is that these regions can aid in modeling image inhomogeneity and our diverse dataset can help

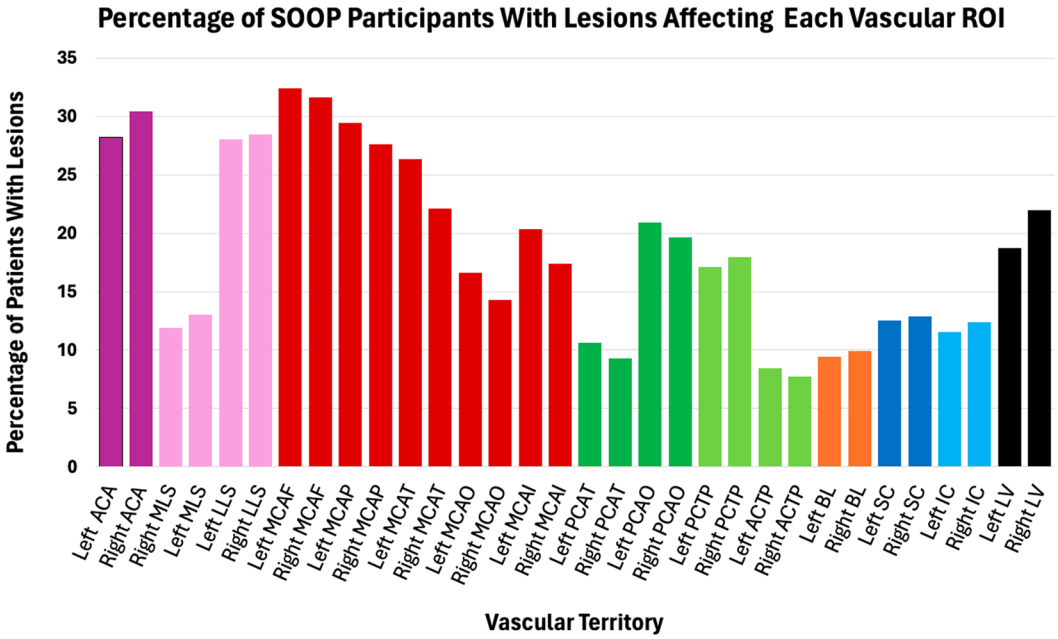

Fig. 1 Percentage of patients experiencing damage to 32 distinct vascular territories described by the digital arterial territory atlas created by Faria and colleagues[33]. The largest percentage of patients experienced MCA injury, but a significant number also experienced ACA and MLS/LLS injuries. The wide variety of lesion location, as well as the bilateral distribution (see Fig. 3) makes SOOP particularly useful to researchers and clinicians interested in recovery of functions that are primarily lateralized (i.e. language) or considered bilateral (i.e. motor).

others develop brain extraction tools trained on a diverse dataset with features such as post-bregmatic depression and wide diploic spaces.

We also conducted stroke lesion mapping to identify and demarcate the extent of the injury. Specifically, three trained neuroscientists (RNN, MG, SW) manually traced lesion boundaries on each axial slice of participants' T2w structural image. The percentage of patients with strokes including specific vascular regions, as defined by Faria's digital arterial territory atlas[33], can be found in Fig. 1.

While there is no universally accepted method for demarcation of acute stroke lesions, our process adheres to several established guidelines. All raters used MRIcroGL12[34] software to manually inspect and trace lesions on ADC diffusion weighted images (DWI) in which acute lesions appeared as hypointense. Three raters trained in the use of MRIcroGL12 in our lab, and experienced with the process of creating lesion masks performed the lesion demarcations (authors RN, MG, and SW). The first step in lesion demarcation was to scroll through the entire ADC image and locate area(s) that, with absolute certainty, contained acutely lesioned tissue. From there, the trained rater demarcated the region in all spatially adjacent (in the superior inferior direction) slices that appeared to contain contiguous lesioned tissue. Lesions data were then exported as binary NIFTI formatted files in subject native DWI space. In these files, a value of '1' denotes lesioned voxels and '0' denotes non-lesioned voxels. Notably, this newly created NIFTI file was aligned with and had the same dimensions as the DWI image on which it was drawn. Lesion masks were produced in native (subject specific) space, and the resulting lesion masks were also normalized to standard anatomical (MNI) space and associated neuroanatomical atlases. A similar process was used to identify participants that additionally showed evidence of chronic stroke lesions, which showed up as hypointense on the same ADC images (acute and chronic stroke lesion files are stored separately in the SOOP OpenNeuro database). Video recordings were made of all lesion demarcation, using Quicktime's 'New Screen Recording' function, and are available upon reasonable request to the corresponding author.

Importantly, recent evidence suggests that not all lesioned tissue exhibits uniform damage characteristics. Specifically, work by Krishnamurthy and colleagues demonstrated that T2w/T1w MRI signal ratios can be used to identify pericavitational areas with varying degrees of tissue integrity, termed Tissue Integrity Gradation via T2w T1w Ratio (TIGR)[35]. This method reveals a gradient of damage within lesions, rather than a binary map like the one generated by our manual lesion demarcation approach, and this may provide for more sensitive lesion-symptom mapping. We acknowledge these developments and propose incorporating such advanced methodologies in future protocols to enhance the precision of lesion identification and characterization[36].

Data Records

The anonymized images for the Stroke Outcome Optimization Project (SOOP) are available from OpenNeuro (https://openneuro.org/datasets/ds004889). The imaging, demographic and behavioral measures are organized using the brain imaging data structure (BIDS)[35] and shared publicly on the OpenNeuro web site[37]. This curated structure provides human readable filenames with a clear file hierarchy for storing data. A benefit of this system

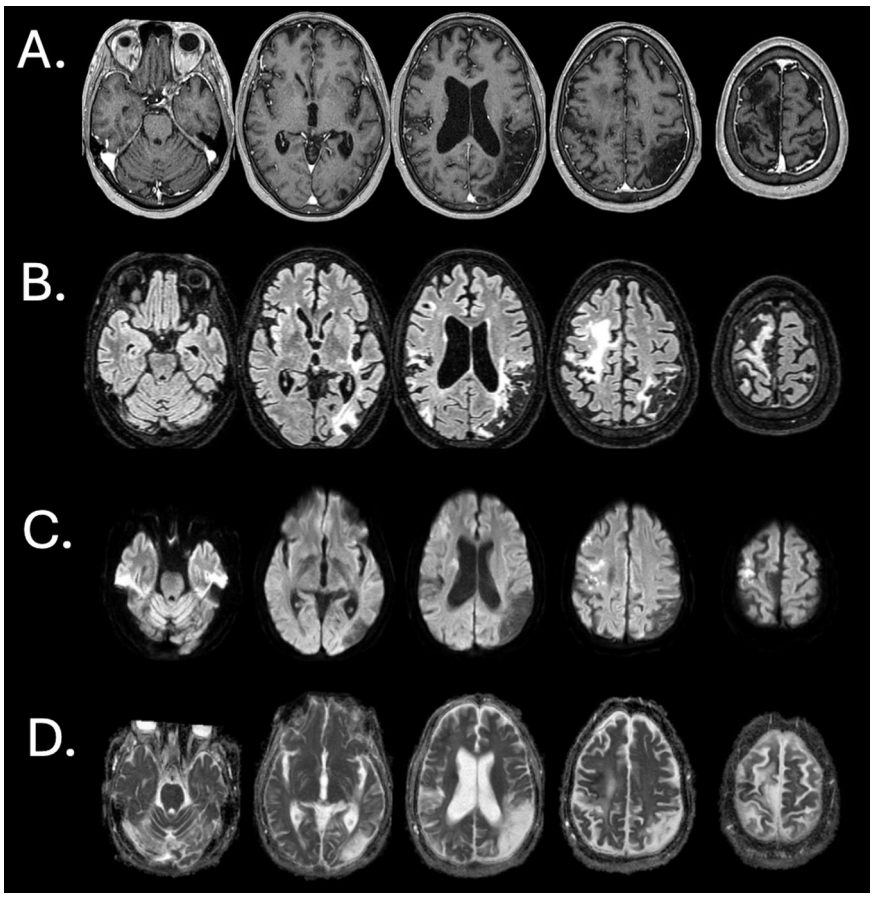

Fig. 2 Example data for one individual (participant 342). For each individual we provide a scan with T1-weighting (**A**), a T2-weighted fluid-attenuated inversion recovery (FLAIR) (**B**), as well as two images from an echo-planar imaging diffusion sequence. With regards to the diffusion sequences shown, this study chose to use very short DWI sequences referred to Apparent Diffusion Coefficient (ADC) (**C**) and TRACE (**D**) scans, as opposed to longer (10–20 minute) DWI sequences used to calculate tractography. These shorter DWI scans allowed for detection of abnormal diffusion using a very short acquisition time, which is apt for clinical settings.

is that it allows automated tools to process and aggregate datasets. Beyond the raw imaging data, the data is provided in text formats that allow inspection. Specifically, the demographic and impairment measures are stored in the tab-separated value text format spreadsheet 'participants.tsv', which includes a labeled header row to describe the variables, and each subsequent row provides the values for a single participant with the first column providing participant identification (e.g. 'sub-11'). The text file 'participants.json' provides in-depth descriptors for each of these labels. The defaced imaging data is stored in a separate folder for each individual (e.g. 'sub-11'). Each participant's folder contains two subfolders: the 'anat' folder stores the anatomical scans (here the T1 and FLAIR modalities) and the 'dwi' folder stores the diffusion data (here the TRACE and ADC images). The MR images are stored in NIfTI format, with each including a text-format JSON file that provides sequence details. The root directory also contains a folder named 'derivatives' which includes the folder 'lesion_masks' containing one folder (e.g. 'sub-11') for each patient where a stroke was observed. These folders provide the lesion maps, drawn on the individual's TRACE image. For individuals who had pre-existing injuries a total of three lesion maps are provided ('-lesionChronic_mask', '-lesionAcute_mask', '-lesion_mask') while those with only recent injury exclusively include the latter two images. The normalized FLAIR images for the SOOP participants are available from the Open Science Framework (OSF)[38].

Technical Validation

Our focus on predicting NIH stroke scale, which is a popular but notably non-comprehensive measure of AIS severity[10], may facilitate education, data sharing and collaboration. We emphasize that the provided dataset can be used by others to improve clinical tools, including automatic lesion mapping, spatial normalization, brain integrity measurements, and predicting outcomes. To demonstrate the richness of the data, we have provided simple scripts that illustrate current best practices to allow easy replication, education and a basic validation benchmark for comparing future tools and methods. Specifically, these scripts use the lesion maps drawn on the TRACE image to predict impairment on the NIH stroke scale.

Briefly, the first task is to warp the lesion masks drawn in the native space of each individual's TRACE image to a common template image. Here we leverage the Clinical Toolbox for SPM[39] to first coregister the

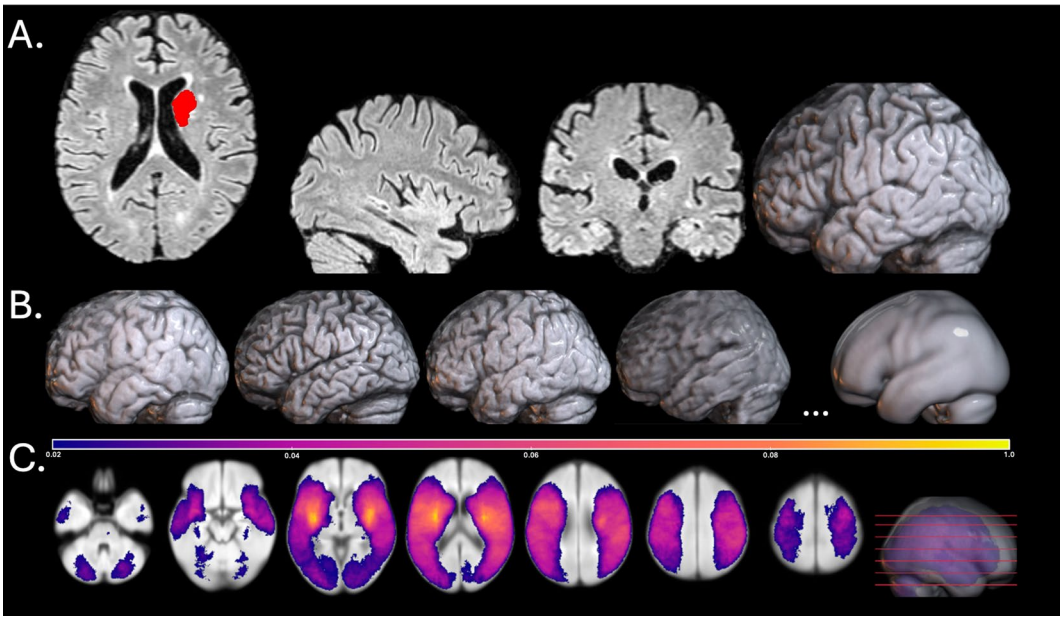

Fig. 3 Top Panel: Bitmap image generated by our validation scripts for a single participant (#23). This image shows axial, sagittal and coronal slices, as well as a rendered image, in standard MNI space. The lesion, located in the right caudate nucleus, is depicted in red. White matter hyperintensities (periventricular) are visible on the anterior boundary of the left and right ventricles. Users can inspect this bitmap can as part of the quality assurance process. In particular, the unified segmentation and normalization method we use develops a virtuous cycle between the spatial warping and the tissue segmentation that drives the brain extraction. Therefore, an accurate volume rendering (right panel) is consistent with a successful spatial warping to standard space. Middle Panel: Despite the variable differences in quality and resolution of each individual FLAIR scan (renderings for four representative individuals shown on the left side), all are normalized into a standard space, as seen by the rendering of the mean normalized FLAIR scan from all individuals (right side). Bottom Panel: Lesion incidence map (N = 1461) for the SOOP dataset. Hotter colors show regions with higher injury incidence.

low-resolution TRACE image to the high resolution FLAIR image (this warps the lesion to FLAIR space) and subsequently conducts unified segmentation and normalization[40] to warp the individuals FLAIR image to a common template (so that the lesion maps from all individuals are in a standard space), as shown in Figs. 2 and 3. We then calculate the proportion of injury for each region in a vascular atlas[33], resulting in a tab separated value where each row provides information from a single subject and each column lists the proportion injury for each territory in the atlas. A script removes columns that are damaged in fewer than a specified proportion of the population. We chose 5% of the participants, following conventions to improve statistical power and spatial biases[41]. Therefore, Fig. 3 appears to omit anterior cerebral artery strokes, which occur at a frequency lower than 5%[42,43]. Note that these spreadsheets match the layout of the 'participants.tsv', allowing us to concatenate the lesion information, demographics and outcome measures for our subsequent analyses.

Finally, we provide a script (deep_learn.py) that computes a simple leave-one-out prediction of the NIH stroke scale based on the imaging measures as well as participant age (see Code Availability section). Users can simply download the entire SOOP project from our GitHub repository (https://github.com/neurolabusc/StrokeOutcomeOptimizationProjectDemo) and run the python file, deep_learn.py, to generate the graph in this manuscript (more detailed instructions for running *deep_learn.py* using Python are included on the GitHub page for this project). Note that the goal of machine learning is to use features synergistically to provide the best prediction. Participant age is known to predict initial stroke score[44], though in our sample it does not prove a reliable predictor on its own. The analysis utilizes a TensorFlow-based implementation in Python, employing a sequential neural network architecture. The model consists of three layers: an initial dense layer with 64 nodes and a rectified linear unit (ReLU) activation function, followed by a second dense layer with 32 nodes and ReLU activation, and finally, an output layer with a single node. We also provide an identical analysis using support vector machines, which can sometimes be more robust for relatively small datasets. We wish to emphasize that our data is amenable to more sophisticated analyses, but our goal is to provide a simple solution using off-the-shelf solutions. Both models significantly predict stroke scale, with the Neural Network correlation $r = 0.543$, p-value < 0.00001, and the SVR $r = 0.550$, p-value < 0.00001. Results for both models are shown in Fig. 4.

Usage Notes

The Stroke Outcome Optimization Project is publicly shared on OpenNeuro using the community developed BIDS structure to enable usage with any BIDS-compatible pipeline. We hope that this will encourage the development, validation and education for novel tools that are capable of handling multiple types of data that influence stroke outcome. The Technical Validation section describes a simple set of analyses using current best

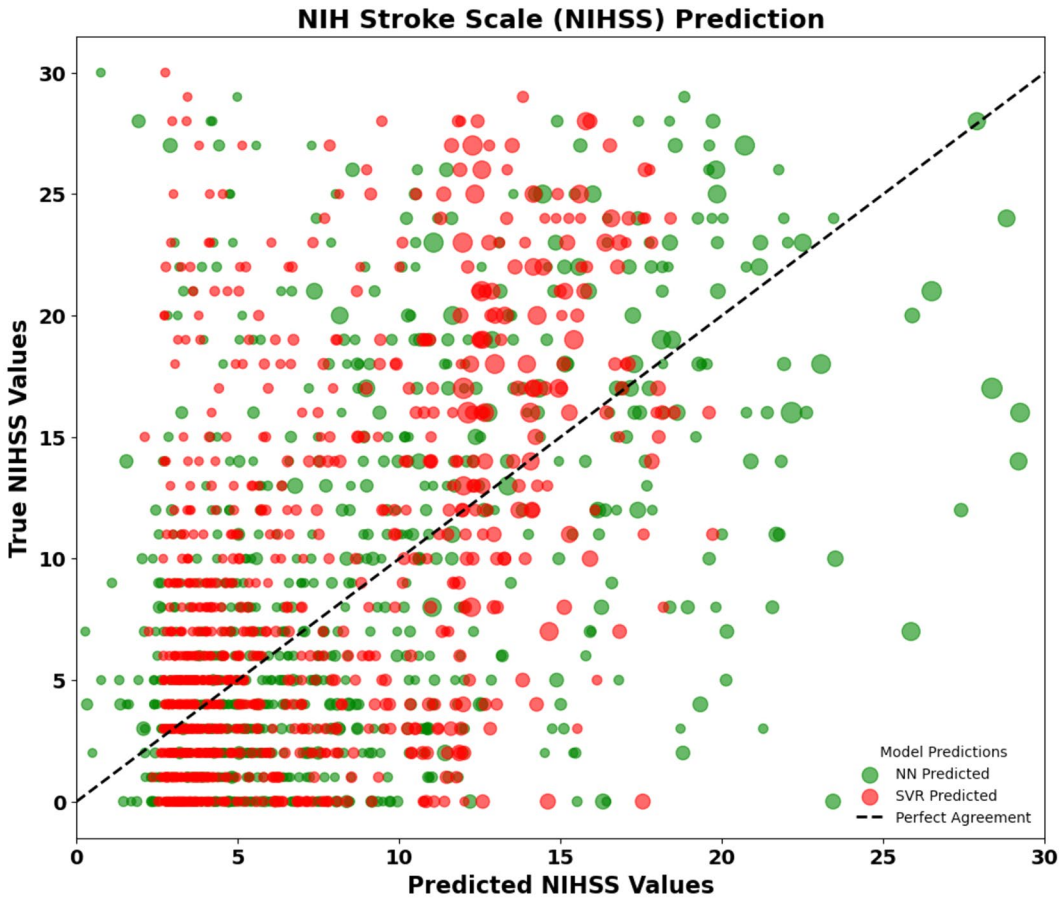

Fig. 4 We created an easy-to-modify script that attempts to predict NIH Stroke Scale (NIHSS) scores based on participant age and lesion load to each brain region described in the vascular territory brain atlas created by Faria and colleagues[33]. Our script deep_learn.py, which is contained in our open-source GitHub repository: https://github.com/neurolabusc/StrokeOutcomeOptimizationProjectDemo), can be run in a Python environment or using Jupyter Notebooks, to predict NIHSS scores using two different algorithms: support vector regression (SVR - red) and neural network (NN - green). This GitHub page contains more detailed instructions on dependencies and how to run this script. Comparison of the performance of these algorithms shows that NN outperforms SVR for this classification task. Other researchers can easily modify this script to run it on subsets of our data (e.g. males vs. females, large vs. small lesions determined by a median split, etc) or compare the performance of other types of machine learning or AI models. *Each circle represents a unique participant. Lesion sizes were converted to z-scores and are represented by the size of each dot. Data points with predicted NIHSS Values $>=30$ (N = 2) or $<=0$ (N = 8) were excluded from the graph for visualization.

practices. The Matlab and Python scripts for reproducing these results are available from GitHub. By design, these scripts focus on simplicity for clarity and training. These scripts provide a basic validation benchmark so others can evaluate the performance of more sophisticated solutions.

Anonymizing and curating large datasets for public sharing requires substantial investment of resources. Our initial release focuses specifically on ischemic stroke. We recognize that our exclusion criteria impact the generalizability of machine learning predictions by omitting structural and clinical comorbidities, and hemorrhagic strokes, that are represented in our entire stroke cohort; by omitting infrequent stroke subtypes such as anterior cerebral artery strokes and hemorrhages, generalizability to an entire stroke population is limited. However, our future goal is to remove exclusion criteria systematically. We provide code and methodology that may be used for data collection across comprehensive stroke centers. Larger datasets will be required to model the impact of uncommon or rare influences on stroke outcome, and we plan to systematically incorporate such comorbidities into our evolving models. Educational and occupational background, race and ethnicity, tobacco, alcohol and drug use, treatment timing and success, and many other factors impact long term stroke outcome. The current work is our initial effort to develop a CPM for stroke using electronic health records (EHR) and MRI data that are routinely acquired during acute ischemic stroke management. We hope to stimulate machine learning methods that will enable a comprehensive accounting of many factors that may influence aphasia outcomes in particular, and eventually other stroke outcomes. We also plan on releasing additional behavioral data to researchers. These additional data will eventually include GWTG data including, comprehensive medical history and current medication information, medications, comorbidities, as well as estimated SES (based on zip-code)

and possibly other MRI-derivatives (such as quantity and location of white matter hyperintensities, perivascular space and microbleeds) These details will be generalized to protect identities.

Code availability

We refined dcm2niix for converting the source DICOM MRI scans to BIDS format, with improvements incorporated in this open source software (https://github.com/rordenlab/dcm2niix). Our defacing method is available from GitHub (https://github.com/neurolabusc/mydeface). We provide minimal Matlab and Python scripts to organize, process, and analyze these data using machine learning. These scripts are all stored in a self-contained archive at GitHub (https://github.com/neurolabusc/StrokeOutcomeOptimizationProjectDemo), allowing others to replicate and extend the findings we describe in the Technical Validation section.

Received: 19 January 2024; Accepted: 22 July 2024;
Published online: 02 August 2024

References

1. Writing Group Members. *et al*. Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. *Circulation* **133**, e38–360 (2016).
2. Weimar, C. *et al*. Prediction of recurrent stroke and vascular death in patients with transient ischemic attack or nondisabling stroke: a prospective comparison of validated prognostic scores. *Stroke* **41**, 487–493 (2010).
3. O'Brien, E. C. *et al*. Quality of Care and Ischemic Stroke Risk After Hospitalization for Transient Ischemic Attack: Findings From Get With The Guidelines-Stroke. *Circ. Cardiovasc. Qual. Outcomes* **8**, S117–24 (2015).
4. Smith, E. E. *et al*. Risk score for in-hospital ischemic stroke mortality derived and validated within the Get With the Guidelines-Stroke Program. *Circulation* **122**, 1496–1504 (2010).
5. Menon, B. K. *et al*. Risk score for intracranial hemorrhage in patients with acute ischemic stroke treated with intravenous tissue-type plasminogen activator. *Stroke* **43**, 2293–2299 (2012).
6. Wessler, B. S. *et al*. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. *Diagn Progn Res* **1**, 20 (2017).
7. Benjamin, E. J. *et al*. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association. *Circulation* **137**, e67–e492 (2018).
8. Wessler, B. S. *et al*. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. *Circ. Cardiovasc. Qual. Outcomes* **8**, 368–375 (2015).
9. Smaha, L. A. The American Heart Association get with the guidelines program. *Am. Heart J.* **148**, S46–S48 (2004).
10. Taylor-Rowan, M., Wilson, A., Dawson, J. & Quinn, T. J. Functional Assessment for Acute Stroke Trials: Properties, Analysis, and Application. *Front. Neurol.* **9**, 191 (2018).
11. Song, S. *et al*. Association of Get With The Guidelines-Stroke Program Participation and Clinical Outcomes for Medicare Beneficiaries With Ischemic Stroke. *Stroke* **47**, 1294–1302 (2016).
12. Mijajlović, M. D. *et al*. Post-stroke dementia - a comprehensive review. *BMC Med.* **15**, 11 (2017).
13. Pendlebury, S. T. *et al*. Methodological factors in determining rates of dementia in transient ischemic attack and stroke: (I) impact of baseline selection bias. *Stroke* **46**, 641–646 (2015).
14. Grefkes, C. & Fink, G. R. Connectivity-based approaches in stroke and recovery of function. *Lancet Neurol.* **13**, 206–216 (2014).
15. Rehme, A. K. *et al*. Individual prediction of chronic motor outcome in the acute post-stroke stage: Behavioral parameters versus functional imaging. *Hum. Brain Mapp.* **36**, 4553–4565 (2015).
16. Volz, L. J. *et al*. Time-dependent functional role of the contralesional motor cortex after stroke. *Neuroimage Clin* **16**, 165–174 (2017).
17. Yoshimoto, T. *et al*. Impact of Previous Stroke on Clinical Outcome in Elderly Patients With Nonvalvular Atrial Fibrillation: ANAFIE Registry. *Stroke* **53**, 2549–2558 (2022).
18. Gallanagh, S., Quinn, T. J., Alexander, J. & Walters, M. R. Physical activity in the prevention and treatment of stroke. *ISRN Neurol.* **2011**, 953818 (2011).
19. Liew, S.-L. *et al*. Association of Brain Age, Lesion Volume, and Functional Outcome in Patients With Stroke. *Neurology* **100**, e2103–e2113 (2023).
20. Garcia, D. A., Arredondo, R. & Morris, M. A review of rehabilitation strategies for stroke recovery. P*roceedings of the* (2012).
21. Yan, H.-Y. & Lin, H.-R. Resilience in Stroke Patients: A Concept Analysis. *Healthcare (Basel)* **10**, (2022).
22. Manolio, T. A. *et al*. New models for large prospective studies: is there a better way? *Am. J. Epidemiol.* **175**, 859–866 (2012).
23. Sudlow, C. *et al*. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
24. Lees, K. R., Khatri, P. & STAIR IX Collaborators. Stroke Treatment Academic Industry Roundtable Recommendations for Individual Data Pooling Analyses in Stroke. *Stroke* **47**, 2154–2159 (2016).
25. Liebeskind, D. S. *et al*. Imaging in StrokeNet: Realizing the Potential of Big Data. *Stroke* **46**, 2000–2006 (2015).
26. Krakauer, J. W. & Marshall, R. S. The proportional recovery rule for stroke revisited. *Annals of neurology* **78**, 845–847 (2015).
27. Kopal, J., Uddin, L. Q. & Bzdok, D. The end game: respecting major sources of population diversity. *Nat. Methods* **20**, 1122–1128 (2023).
28. Liu, C.-F. *et al*. A large public dataset of annotated clinical MRIs and metadata of patients with acute stroke. *Sci Data* **10**, 548 (2023).
29. Brett, M., Leff, A. P., Rorden, C. & Ashburner, J. Spatial normalization of brain images with focal lesions using cost function masking. *Neuroimage* **14**, 486–500 (2001).
30. Office for Civil Rights (OCR). Guidance regarding methods for DE-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) Privacy Rule. H*HS.gov* https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (2012).
31. Li, X., Morgan, P. S., Ashburner, J., Smith, J. & Rorden, C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* **264**, 47–56 (2016).
32. Friston, K. J., Ashburner, J. T., Nichols, T. E. & Penny, W. D. *Statistical Parametric Mapping the Analysis of Funtional Brain Images*. (Elsevier/Academic Press, Amsterdam, 2007).
33. Liu, C.-F. *et al*. Digital 3D Brain MRI Arterial Territories Atlas. *Sci Data* **10**, 74 (2023).
34. Rorden, C. & Brett, M. Stereotaxic display of brain lesions. *Behav. Neurol.* **12**, 191–200 (2000).
35. Gorgolewski, K. J. *et al*. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).
36. Krishnamurthy, L. C. *et al*. Not All Lesioned Tissue Is Equal: Identifying Pericavitational Areas in Chronic Stroke With Tissue Integrity Gradation via T2w T1w Ratio. *Front. Neurosci.* **15**, 665707 (2021).
37. Rorden, C., Absher, J. & Newman-Norlund, R. Stroke Outcome Optimization Project (SOOP). *OpenNeuro* https://doi.org/10.18112/openneuro.ds004889.v1.1.2 (2024).

38. Rorden, C., Absher, J. R., Gibson, M., Teghipco, A. & Newman-Norlund, R. Stroke Outcome Optimization Project (SOOP). *OSF* https://doi.org/10.17605/OSF.IO/YQKTJ (2024).
39. Rorden, C., Bonilha, L., Fridriksson, J., Bender, B. & Karnath, H.-O. Age-specific CT and MRI templates for spatial normalization. *Neuroimage* **61**, 957–965 (2012).
40. Ashburner, J. & Friston, K. J. Unified segmentation. *Neuroimage* **26**, 839–851 (2005).
41. Sperber, C. & Karnath, H. Impact of correction factors in human brain lesion-behavior inference. *Hum. Brain Mapp.* **38**, 1692 (2017).
42. Arboix, A. *et al*. Infarction in the territory of the anterior cerebral artery: clinical study of 51 patients. *BMC Neurol.* **9**, 30 (2009).
43. Thirugnanachandran, T. *et al*. Anterior Cerebral Artery Stroke: Role of Collateral Systems on Infarct Topography. *Stroke* **52**, 2930–2938 (2021).
44. Simmons, C. A., Poupore, N. & Nathaniel, T. I. Age Stratification and Stroke Severity in the Telestroke Network. *J. Clin. Med. Res.* **12**, (2023).

Acknowledgements

Administrative support, space, and resources were provided by Clemson University and Prisma Health. Technical support was provided by Clemson University, Prisma Health, and the University of South Carolina. Philanthropic support was provided by Furman University, including stipends for several students: Elizabeth Nethercoat, Michael Garovich, Natalie Dunn, Molly Oroho, Cade Azzariti, Hailey Turk, and Davis Dear. Patrick Burton received stipend support through a Health Sciences Center Seed Grant. We also acknowledge Jenna Durham, Sarah Hierholzer, Leigh Ann Spell, Wes Wimpey, and Alex Ewing for their contributions. This work was supported by the National Institute of Health (P50DC014664, RF1MH133701). We would like to acknowledge the participants, students, faculty, and staff who have supported the Center for the Study of Aphasia Recovery.

Author contributions

J.A. designed the research study, supervised data acquisition, wrote the first draft of the manuscript, and edited the final versions for publication. C.R. developed the manuscript for the journal format and developed all scripts and validation testing. L.B. developed the machine learning scripts described in the manuscript. J.A., C.R., R.N., S.K. and A.T. performed the MRI processing and analysis. S.G. and J.A. evaluated each subject's inclusion/exclusion criteria, and performed speech language pathology (SLP) ratings. Nicholas Perkins wrote the code responsible for EPIC and MRI data abstraction. C.R., J.V.M. and N.P. collaborated on MRI data collection, conversion to NIfTI format, de-identification, and coding. G.Y., J.A., S.K., A.T., R.N. and C.R. developed and maintained the neuroimaging software, and collaborated on data analysis. M.G. and S.W. assisted R.N. with manual lesion delineation. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.A.