**Big data approaches to phenotyping stroke risk and severity using automated multi-center stroke lesion segmentation**

Ona Wu, Ph.D1; Stefan Winzeck, MS1,2; Anne-Katrin Giese, MD3; Brandon L. Hancock, BS1; Mark R. Etherton, MD PhD3; Mark J. R. J. Bouts, PhD1; Kathleen Donahue, BS3; Markus D. Schirmer, PhD3; Robert E. Irie, PhD1; Steven J. T. Mocking, MS1; Elissa C. McIntosh, MA1; Raquel Bezerra, MD1; Konstantinos Kamnitsas MS4; Petrea Frid, MD5; Johan Wasselius, MD, PhD5,6; John W. Cole, MD7; Hui Xu, MD, PhD7; Lukas Holmegaard, MD8; Jordi Jiménez-Conde, MD, PhD9; Robin Lemmens, MD, PhD10; Eric Lorentzen, Ph. Lic.11 ; Patrick F. McArdle, PhD12; James F. Meschia, MD13; Jaume Roquer, MD, PhD9; Tatjana Rundek, MD, PhD14; Ralph L. Sacco MD, MS14; Reinhold Schmidt, MD15; Pankaj Sharma, MD, PhD16; Agnieszka Slowik, MD, PhD17; Tara Stanne, PhD11; Vincent Thijs, MD, PhD18; Achala Vagal, MD19; Daniel Woo, MD, MS20; Stephen Bevan, PhD21; Steven J. Kittner, MD, MPH22; Braxton D. Mitchell, PhD, MPH22; Jonathan Rosand, MD, MSc3,23; Bradford B. Worrall, MD, MSc24; Christina Jern, MD, PhD8; Arne G. Lindgren, MD, PhD5,25; Jane Maguire, PhD26; Natalia S. Rost, MD MPH3; on behalf of the MRI-GENIE and GISCOME Investigators.

**Affiliations:**

1. Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital (MGH), 149 13th Street, Charlestown, MA 02129

2. Division of Anaesthesia, Department of Medicine, University of Cambridge, UK

3. JP Kistler Stroke Research Center, Department of Neurology, MGH

4. Department of Computing, Imperial College London, London, UK

5. Department of Clinical Sciences Lund, Neurology, Lund University, Lund, Sweden;

6. Department of Radiology, Neuroradiology, Skåne University Hospital, Malmö, Sweden.

7. Department of Neurology, University of Maryland School of Medicine and Veterans Affairs Maryland Health Care System, Baltimore, MD, USA.

8. Institute of Neuroscience and Physiology, the Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden.

9. Department of Neurology, Neurovascular Research Group (NEUVAS), IMIM-Hospital del Mar (Institut Hospital del Mar d’Investigacions Mèdiques), Universitat Autonoma de Barcelona, Barcelona, Spain.

10. KU Leuven - University of Leuven, Department of Neurosciences, Experimental Neurology; VIB – Center for Brain & Disease Research; University Hospitals Leuven, Department of Neurology, B-3000 Leuven, Belgium.

11. Institute of Biomedicine, the Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden.

12. Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA.

13. Department of Neurology, Mayo Clinic, Jacksonville, FL, USA.

14. Department of Neurology, Miller School of Medicine, University of Miami, Miami, FL, USA.

15. Department of Neurology, Clinical Division of Neurogeriatrics, Medical University Graz, Graz, Austria.

16. Institute of Cardiovascular Research, Royal Holloway University of London (ICR2UL), Egham, UK, Ashford and St Peter’s Hospital, UK.

17. Department of Neurology, Jagiellonian University Medical College, Krakow, Poland.

18. Stroke Division, Florey Institute of Neuroscience and Mental Health, Heidelberg, Australia and Department of Neurology, Austin Health, Heidelberg, Australia.

19. Department of Radiology, University of Cincinnati College of Medicine, Cincinnati, OH, USA

20. Department of Neurology and Rehabilitation Medicine, University of Cincinnati College of Medicine, Cincinnati, OH, USA

21. School of Life Science, University of Lincoln, Lincoln, UK

22. Geriatric Research and Education Clinical Center, Veterans Administration Medical Center, Baltimore, MD, USA.

23. Henry and Allison McCance Center for Brain Health.

24. Departments of Neurology and Public Health Sciences, University of Virginia, Charlottesville, VA, USA.

25. Department of Neurology and Rehabilitation Medicine, Neurology, Skåne University Hospital, Lund, Sweden;

26. University of Technology Sydney, Sydney, Australia

**Corresponding Author:**

Ona Wu, PhD

Athinoula A Martinos Center for Biomedical Imaging

149 13th Street, CNY 2301

Charlestown, MA 02129

E-mail: ona@nmr.mgh.harvard.edu

Telephone: (617) 643-3873

**Cover Title:** Automated segmentation of multicenter DWI datasets

**Total number of tables and figures:** Tables 1; Figures 4

**Key Words:**  computer based model; diffusion-weighted imaging; risk factors; stroke, ischemic; statistical model;

**Subject Terms:** Cerebrovascular Disease/Stroke, Ischemic Stroke, Magnetic Resonance Imaging (MRI), Risk Factors

**Word Count:** 5911

**ABSTRACT**

**Background and Purpose**: To evaluate deep learning algorithms’ segmentation of acute ischemic lesions on clinical diffusion-weighted (DWI) datasets and explore the potential role of this tool for phenotyping stroke risk, severity and etiology.

**Methods**: Ischemic stroke data sets from the MRI-GENetics Interface Exploration (MRI-GENIE) repository consisting of 12 international genetic research centers were retrospectively analyzed using an automated deep learning segmentation algorithm consisting of an ensemble of 3D convolutional neural networks (CNNs). Three ensembles were trained using data from: (1) a subset of subjects from MRI-GENIE, (2) an independent cohort, and (3) mixture of (1) and (2). The algorithms’ performances were compared against manual outlines for a subset of the MRI-GENIE dataset. Univariable and multivariable logistic regression with respect to demographics, stroke subtypes and vascular risk factors were performed to identify phenotypes associated with large acute DWI volumes and greater stroke severity.

**Results**: The mixture of ensemble performed best. Subset analysis comparing automated and manual lesion volumes in 650 patients found excellent correlation (ρ=0.93, p<0.0001). DWI datasets from 2770 patients from MRI-GENIE were segmented (median [IQR] lesion volumes 3.7 [0.9-16.1] cm3). Patients with small artery occlusion (SAO) stroke subtype had smaller lesion volumes than other subtypes (p<0.0001) and different topology. Atrial fibrillation, and non-SAO subtypes were associated with larger acute lesion volumes in univariable and multivariable analysis (p<0.05). Multivariable analysis indicated that larger lesion volumes (p<0.0001), and atrial fibrillation were predictors (p<0.0001) of greater admission stroke severity.

**Conclusions**: Automated accurate clinical DWI lesion segmentation using deep learning algorithms trained with multi-center and diverse data is feasible. Both lesion volume and topography can provide insight into stroke etiology with sufficient sample size from “big” clinical imaging phenotype datasets.

**INTRODUCTION**

Pooling of multi-center datasets has resulted in recent progress towards understanding the genetic pathways underlying ischemic stroke1 and have identified novel loci associated with ischemic stroke.2, 3 Brain imaging, more precise in characterizing the sequelae of ischemic brain injury than clinical exams, could provide further insight. Analysis of big data repositories required for genome wide association studies and vascular risk factors research will necessitate high-throughput, automated and scalable techniques to measure acute ischemic lesions on diffusion-weighted MRI (DWI) precisely and on a multi-center basis. Importantly, for genetic discovery studies which require the processing of thousands of disparate studies, these algorithms will have to perform accurately on MRI scans that have been obtained for routine clinical purposes, across multiple vendors, not using standardized acquisition protocols.

It has been previously shown that an automated method consisting of an ensemble of 3D convolutional neural networks (CNN) trained on multiparametric DWI data can improve segmentation of acute ischemic lesions compared to single CNNs or CNNs whose segmentations are based on a single diffusion parametric map (Winzeck et al, unpublished data, 2019). However, this study’s results were based on training and evaluation data from a single center and one MRI vendor. Performance might be further improved using multi-center training data since a known short-coming of deep learning algorithms trained at one center is that they may be over-trained and fail to perform accurately with multi-center datasets. Another question is whether diversity of training data can further improve segmentation results. Here, we investigate the robustness of a single center trained neural network for the task of segmenting DWI lesions from the MRI-GENetics Interface Exploration (MRI-GENIE) study, a large multi-center stroke genetics imaging repository, and compare its performance against an algorithm trained with a data subset from MRI-GENIE. In addition, we investigate whether a more diverse training data set, consisting of data from the single center cohort and MRI-GENIE, can lead to improved performance.

Once we have optimized our segmentation algorithm, we applied it to the entire MRI-GENIE cohort to compare the relationships of the predicted volumes against known cerebrovascular risk factors associated with large DWI lesion volumes to test consistency of our automated results with those previously demonstrated using manual approaches. We will also investigate difference in stroke lesion topography as a function of stroke subtype. We will not explore genetic pathways associated with large DWI volumes since that is beyond the scope of this study.

**METHODS**

***Subjects***

All analyses were retrospectively performed under local institutional review board approval.

*Single-Center Cohort*

The single-center data was an independent dataset consisting of 267 acute ischemic stroke patients who presented at a single academic medical center admitted between 1996–2012 with MRI obtained within 24h since the patient was last known to be well (LKW). All patients had not been treated with revascularization treatment prior to MRI. The training cohort were drawn from repositories for which manual outlines were available which had been drawn for other studies.4-6

*MRI-GENIE Cohort*

The MRI-GENIE cohort consisted of 3,301 acute ischemic stroke data sets from 12 international centers from the Stroke Genetics Network.1 All patients had signed informed consent and were included if the following were available: clinical MRI, demographics, vascular risk factors, genome-wide genotyping and stroke subtype measured with the causative classification of stroke (CCS) system.7 Revascularization interventions were not reported. Six centers also had NIH Stroke Scale (NIHSS) scores collected from patients within 11 days of admission as part of the Genetics of Ischaemic Stroke Functional Outcome study.8

***MRI***

*Single-Center cohort*

All data had been obtained for clinical purposes from 1.5T scanners (General Electric Medical Systems). For the majority of cases, the diffusion-weighting (b-value) was 1000 s/mm2. (Please see <http://stroke.ahajournals.org> for details.)

*MRI-GENIE cohort*

Data for the evaluation cohort were acquired on 1 T, 1.5 T and 3T MRI equipment from 6 vendors (please see <http://stroke.ahajournals.org> for details). All datasets were reviewed for availability of raw diffusion data or at least 2/3 of the following diffusion parametric maps: isotropic trace DWI (iDWI), apparent diffusion coefficient (ADC) and b0 maps. Data were excluded which had insufficient image quality for humans to identify the infarct.

*Diffusion Parametric Image calculation*

For studies with the raw diffusion data available, the individual high b-value volumes were corrected for eddy current distortions prior to calculation of iDWI (geometric mean of the high b-value acquisitions), and ADC maps (slope of the linear regression fit of the log of the DWI and b0, b-value=0 s/mm² images).9 If only 2 of the post-processed ADC, iDWI or b0 images were present, the third was calculated from the other two: iDWI=b0 \* exp (–b-value**\***ADC). If the b-value was not encoded in the images, 1000 s/mm2 was assumed.

***Automated segmentation with CNN***

ADC, iDWI and b0 datasets were up-sampled to 1 mm3 isotropic voxels. A brain mask was created by multiplying the skull-stripped up-sampled b0 image10 with an ADC mask (upsampled ADC map greater than 0). The brain mask was used to normalize ADC, iDWI and b0 values by subtracting the mean intensity values (limited to 1-99% of range to exclude outliers) and dividing by the standard deviation. (Please see <http://stroke.ahajournals.org>, Figure S1).

3D CNNs were trained on a workstation with an NVIDIA Tesla K40 GPU using the DeepMedic (v0.7.0) framework with two pathways.11 Details of the DeepMedic framework can be found in the publication11 and in the Supplemental Methods (please see <http://stroke.ahajournals.org>). By design, the DeepMedic training process involves random sampling of data to avoid over-fitting and retain computational efficiency. Despite training a CNN with the same hyper-parameters and data, this sampling strategy inherently results in variations in the lesions segmentation. To compensate for sampling-induced noise, six independent CNNs were trained and the voxel-wise average of their posterior maps for the lesion class were computed, resampled back to the original image resolution and then thresholded at 50% and masked with the brain mask created at the normalization step (please see <http://stroke.ahajournals.org>, Figure S1).

We trained three ensembles: single-center, MRI-GENIE, and mixed. For the single-center ensemble, we trained 6 CNNs using only data from the single center. We repeated this training of 6 independent CNNs with the same number of subjects from MRI-GENIE cohort to create the MRI-GENIE ensemble. We then selected 3 of the CNNs trained on the independent cohort and 3 of the CNNs trained on the MRI-GENIE cohort to create the Mixed ensemble. In this manner, we controlled for performance benefits from an imbalance of number of subjects involved in training across ensembles.

***Segmentation Performance Evaluation***

Two readers (Readers 1 and 2) blinded to the pipeline results created manual outlines for a subset of MRI-GENIE patients (Manual Cohort A). These outlines had been drawn as part of another study involving patients from 8/12 centers and were randomly selected.12 25 datasets were randomly selected from Manual Cohort A for training Reader 2 (please see <http://stroke.ahajournals.org>, Supplemental Methods). Another 25 datasets were randomly selected to compare inter-rater manual agreement and were outlined independently by Readers 1 and 2. A second set of 25 datasets (Manual Cohort B) were randomly selected from a ninth center and outlined by a de novo third reader (Reader 3). All Readers had access to the iDWI, ADC and b0 maps but did not systematically refer to ADC and b0 images and created outlines without using thresholds.

Manual Cohort B was used to determine which of the 3 ensemble models performed best on unseen data from a “new” site. Using outlines from Manual Cohort A for choosing the best ensemble model would bias results towards the ensemble model trained with MRI-GENIE data since the evaluation cohort would have been drawn by Readers 1 and 2. The ensemble model with the best performance metrics (described below) segmenting Manual Cohort B was applied to the entire MRI-GENIE cohort and used for subsequent analyses.

Automated segmentation results were compared to manual outlines volumetrically (Spearman’s non-parametric correlation). Voxel-wise performance was assessed by Dice (measure of overlap between automated and manual lesions), sensitivity, and precision (positive predictive value) scores. Dice score, precision and sensitivity were computed as follows: Dice = 2TP/(2\*TP+FP+FN); Precision = TP/(TP+FP); Sensitivity = TP/(TP+FN) for which TP=true positives, FP=false positive and FN=false negatives. All metrics range from 0-100%, for which higher values indicate better performances. All voxel-wise metrics were evaluated at 1 mm resolution to reduce confounds from different MRI acquisition resolutions. Performance between the best performing ensemble and its individual CNNs were compared (paired two-sided Wilcoxon signed rank tests). Comparisons were limited to patients for whom DWI lesions were detected by the manual readers and whose data were not used for training.

***Statistical analysis***

Analyses were conducted using JMP Pro 14.0 (*SAS Institute*, Cary, NC) unless otherwise noted. Univariable comparisons were done with Wilcoxon two-sample rank sum test for continuous variables, and two-sided Fisher’s Exact Test for categorical variables. Results are described in terms of mean±standard deviation, or median [interquartile range]. For all analyses, p-values < 0.05 were considered statistically significant.

Because of the pooling of multi-center data, heterogeneity of clinical and imaging phenotypes across centers was measured using the I2 metric,13 calculated with the meta package in R (<http://meta-analysis-with-r.org>). Moderate heterogeneity was defined as I2 between 0.5 and 0.75, and high heterogeneity as I2 > 0.75. Mixed modeling (fitted with restricted maximum likelihood method) was used with center treated as a random effect to investigate the association between segmentation performance and lesion volume, time-to-MRI and MRI acquisition parameters (field-of-view, b-value, field strength and vendor).

One-way analysis of variance followed by post-hoc Wilcoxon rank tests was used to compare lesion volumes as a function of center and CCS scores. Multivariable linear regression for predictors of large lesion volumes was performed by including variables that were significant (p<0.05) on a univariable basis. Genetic research center was included as a random effect in all models. For univariable and multivariable analyses, stroke subtype was dichotomized to SAO or non-SAO.

To investigate the topological differences between stroke subtypes, b0 data were registered to the ICBM T2 atlas14, 15 (which had been registered to MNI 152 1 mm atlas from FSL16) using non-linear registration techniques.16, 17 The automatically segmented lesions were spatially registered to MNI 152 1 mm atlas using the derived transformation parameters. Lesion incidence maps were calculated for all MRI-GENIE DWI data. Frequency maps as a function of stroke subtype were generated by dividing the incidence map by the number of subjects and compared using cross-correlation (fslcc16). Results were displayed using MRIcroGL ([http://www.mricro.com](http://neuroling.arizona.edu/resources.html)) in radiologic convention, with MNI coordinates provided in mm. Videos were made using

**RESULTS**

***Subjects***

The demographics for the 267 patients in the single-center training cohort are shown in Supplemental Table S1 (please see <http://stroke.ahajournals.org>). For the independent multi-center dataset, of the 3301 patients available in MRI-GENIE, 531 were excluded for reasons described above (please see <http://stroke.ahajournals.org> for Supplemental Figure S2). Distribution of the remaining 2770 patients as a proportion of the MRI-GENIE and Stroke Genetics Network cohorts by center are shown in Supplemental Figure S3 (please see <http://stroke.ahajournals.org>). The mean±SD age was 63.4±14.7 years old (N=2765). Median [IQR] NIHSS score was 3 [2-7] (N=1312), obtained 1 [0-1] days from admission. Time-to-MRI was 1 [0-4] (N=2464) days from admission. Moderate to high heterogeneity across centers was observed across parameters (please see <http://stroke.ahajournals.org>, Supplement). Demographics breakdown by center can be found in Supplemental Tables S2 and S3 (please see <http://stroke.ahajournals.org>). Demographics for the MRI-GENIE training cohort are provided in Supplemental Table S4 (please see <http://stroke.ahajournals.org>).

***Segmentation Performance Evaluation***

*Ensemble versus ensemble models*

Two cases in Manual Cohort B for which no lesion was detected by the Reader 3 were excluded and analysis was limited to the remaining 23 patients. The distribution of Manual Cohort B volumes was 2.0 [0.9-9.1] cm3. The mixed ensemble model had a Dice score (0.86 [0.79-0.89]) that was significantly superior to that of the single-center cohort ensemble (0.79 [0.66-0.86], p=0.0002) and MRI-GENIE ensemble (0.82 [0.64-0.88], p=0.007) (see Figure 1). The single-center ensemble was comparable to the MRI-GENIE ensemble (p=0.24). The mixed ensemble precision (0.86 [0.7-0.91]) was greater than the single-center ensemble (0.72 [0.54-0.91], p=0.002) but not the MRI-GENIE ensemble (0.85 [0.47-0.91], p=0.09). There was no significant difference in precision between the single-center and MRI-GENIE ensembles (p=0.66). There were no statistically significant differences in terms of sensitivity between the ensembles (mixed: 0.90 [0.84-0.95]; MRI-GENIE: 0.88 [0.81-0.96]; single-center: 0.89 [0.84-0.93]). For the remainder of the analysis, we thus used the mixed ensemble model.

*Ensemble versus individual CNN models*

Manual Cohort A consisted of 666 patients. 16 patients were excluded since no DWI lesion was identified by the readers. 267 patients used for training 3 of the CNNs in the mixed ensemble were excluded for performance metrics analysis to avoid biasing the results. Median [IQR] manual lesion volumes for the remaining 383 subjects were 2.9 [0.8-19.2] cm3. The Ensemble model outperformed all individual CNN models (please see <http://stroke.ahajournals.org>, Figure S4) in terms of Dice score (0.77 [0.57–0.88]), and precision 0.83 [0.57-0.94]) but not sensitivity (0.82 [0.66-0.91]) for which it underperformed one individual CNN model, performed equivalently to 2 others and outperformed 3. There was one patient for which the ensemble model did not detect any lesion (measured volume was 0.2 cm3), and thus precision could not be calculated for the one patient and excluded in the analysis. For the remainder of the analyses, we therefore focus on the ensemble results.

*Effects of MRI acquisition on performance*

The median automated lesion volumes for the 383 subjects were 3.9 [0.9-18.3] cm3, with median difference from measured volumes of -0.02 [-0.9-0.8] cm3, and correlated significantly with respect to manual outlines (ρ=0.93, p<0.0001). Mixed modeling with center as a random effect showed that Dice score (p<0.0001), precision (p<0.0001), and sensitivity (p<0.0001) were significantly greater with larger lesion volumes. No differences in performance (p>0.05) were found as a function of time-to-MRI (though very late scans could have poor segmentation results, please see <http://stroke.ahajournals.org>, Figure S5), field-of-view, field strength and vendor. However, there were significant differences for high b-value for Dice (b-value=1500 s/mm2 (N=52): 0.52 [0.30-0.76] vs b-value=1000 s/mm2 (N=331): 0.80 [0.63-0.88], p=0.009) and precision (b-value=1500 s/mm2 (N=52): 0.43 [0.21-0.80] vs b-value=1000 s/mm2 (N=330): 0.86 [0.65-0.94], p=0.02), but not sensitivity (p=0.64).

*Inter-rater analysis: Human versus human versus machine*

For inter-rater analysis, no statistically significant differences were found between the performance metrics of Reader 2 and Ensemble method for the 25 independent cases in terms of (a) median differences with Reader 1 (Reader 2: 0.4 [-1.7-4.4]; Ensemble: -1.6 [-3.4-0.9] cm3; p=0.10) (b) Dice score (Reader 2: median 0.86 [IQR 0.81-0.95]; Ensemble: 0.88 [0.78-0.95]; p=0.93), (c) precision (Reader 2: 0.93 [0.83-0.97], Ensemble: 0.94 [0.81-0.96], p=0.95), and (d) sensitivity (Reader 2: 0.86 [0.79-0.93]; Ensemble: 0.90 [0.83-0.94], p=0.11). Figure 2 shows an example of discordant results between Reader 1 and 2 in a subject with acute, subacute and chronic infarcts. The Ensemble model segmented a lesion that encompassed regions intermediate between the two readers.

***Risk Factors for Large Acute Ischemic Lesion Volumes***

The segmentation results using the Ensemble model was 3.7 [0.9-16.6] cm3 across all 2770 patients. Automated lesion volumes between patients with and without non-zero manual outlines were not significantly different (p=0.69). The remainder of the analyses therefore used results from all 2770 patients.

DWI lesion volumes (lnDWIv) differed significantly (p<0.0001) by center and by CCS (Figure 3). Lesion volumes differed significantly in terms of self-identified race (white), AF, and SAO strokes (see Table 1). Lesion volumes were significantly correlated with NIHSS (N=1312, ρ=0.38, p<0.0001), but not age (ρ=-0.03, p=0.08). Using univariable regression, AF (p<0.0001), self-identified white race (p=0.02), and non-SAO subtype (p<0.0001) were found to be independent predictors of larger lesion volumes. Multivariable analysis showed only non-SAO stroke subtype (p<0.0001), and AF (p=0.0009) remained significant predictors of larger lesion volumes (please see <http://stroke.ahajournals.org>, Table S5).

For centers with NIHSS available, non-SAO subtype (p<0.0001), AF (p=0.008), hypertension (p=0.048), age (p=0.02) and NIHSS (p<0.0001) were independent univariate predictors of larger lesion volumes (please see <http://stroke.ahajournals.org>, Table S6). Multivariable analysis resulted in only non-SAO stroke subtypes (p<0.0001), younger age (p=0.007) and larger NIHSS scores (p<0.0001) being significant predictors of larger DWI volumes. Since NIHSS is itself influenced by many vascular risk factors, univariable analysis was performed to identify significant predictors (p<0.05) of higher NIHSS in patients with DWI volumes available (non-SAO, AF, and acute lesion volume). Lesion volume (p<0.0001), and AF (p<0.0001) remained independent predictors of greater NIHSS in multivariable analysis (please see <http://stroke.ahajournals.org>, Table S7).

***Stroke Lesion Topology***

MRI data from 7 patients could not be non-linearly co-registered due to image artifacts. Stroke lesion incidence maps for remaining 2763 are shown in Figure 4. Figure 5 shows the frequency maps for each stroke subtype (please see <http://stroke.ahajournals.org>, Supplemental Videos). SAO had the greatest difference compared to non-SAO (R=0.39) followed by Other (R=0.89) with the others showing high similarity (LAA vs non-LAA, R=0.95; cardioembolic vs non-cardioembolic, R=0.94; undetermined vs non-undetermined, R=0.94).

**DISCUSSION**

We have demonstrated accurate fully automated segmentation of ischemic lesions of multi-center clinically acquired DWI data is feasible. These data sets were obtained as part of routine clinical practice involving multiple field strengths, sequences, vendors, and acquisition protocols that spanned over several years, resulting in very heterogeneous populations. We showed that an ensemble algorithm trained with data from a single center can perform comparably well with a model trained with multi-center data. However, we showed that even better results can be obtained when using an ensemble algorithm trained on diverse training data sets. The performance of all models were similar to another study which used CNN approaches to evaluate 361 patients from multiple centers, achieving mean Dice score of 0.67.18 Our study is distinct from the other study18 in that we performed a thorough comparison of the effects of various training datasets on the performance of machine learning algorithms. In addition, the other study developed a CNN that was trained (N=380) and tested (N=361) on data from the same centers, naturally achieving high performance metrics. One of the key questions we sought to address was whether models trained on one center’s DWI data can be transferred to segment DWI data from another center accurately. If the answer was no, for genetic discovery studies that require big datasets, a new algorithm would need to be trained whenever a new center’s dataset was incorporated into a repository. Fortunately, our results suggest that a robust segmentation network developed using a mixture of ensembles trained on diverse datasets can provide accurate segmentation.

We also explored factors inherent to multi-center datasets that can affect the performance of machine learning algorithms. Interestingly, variations in DWI acquisition across vendors did not affect the algorithm performance, with the exception of high b-value. The results of poor performance on data acquired with very high b-values (1500 s/mm2) is not surprising since diffusion data acquired with b-values>1000 s/mm2 are known to exhibit non-mono-exponential behavior leading to non-pathological differences between normal gray and white matter.19 We found that smaller lesion volumes had poorer Dice scores, consistent with previous results using the ensemble approach (Winzeck et al, unpublished data, 2019) and other methods.18 Even for human readers, the intra-rater and inter-rater volume measurements fluctuate by 2.7 cm3.20 The performance of our method may therefore be underestimated since the MRI-GENIE cohort consisted of primarily smaller lesion volumes. Segmentation accuracy did not depend on the timing of the MRI scan with respect to stroke presentation. We mitigated the impact of timing by training with data acquired over a range of times with our mixed ensemble model.

The high throughput afforded us by our automated method allowed us to test hypotheses that could be limited by sample size. Our study demonstrated the potential benefits from using large-scale data sets to identify vascular risk factors for more severe strokes, and hence less favorable outcomes. This in turn can inform genetic discovery studies. Although our finding that SAO stroke subtypes involve smaller DWI lesion volumes is well known, and stroke subtype classification typically uses imaging, it is reassuring that our automated algorithms generated consistent results. Without having manual lesion volumes available for all 2770 subjects, this finding along with results showing positive correlation of NIHSS with automated DWI segmentations reinforces our confidence in the accuracy of our results. Our study also showed that AF is an independent risk factor for larger stroke lesion volumes in line with other studies.21, 22 Since AF classification is independent of imaging findings, these results could be considered an independent validation of our method. In addition, given our large dataset, we were able to create incidence maps as a function of stroke subtype consisting of at least 200 subjects for each subtype. However, 200 subjects may still be insufficient to test hypotheses, emphasizing the need for even larger imaging repositories and automated pipelines for extracting phenotypic information.

Major strengths of our study are the very large number of patients (N=2770) in our study and their diversity. However, the pooling of multi-center datasets for which data were collected retrospectively can also be considered a limitation due its heterogeneity. For the majority of patients, the b-value of the DWI acquisition was not available and thus a b-value=1000 s/mm2 was assumed, affecting ADC calculation. However, our pipeline normalized input images prior to segmentation and thus mitigated the effects of incorrect b-values. Furthermore, our findings of associations of NIHSS, non-SAO stroke subtype, age and AF with larger lesion volumes are consistent with previous independent studies. Thus, the heterogeneity of the MRI-GENIE cohort could also be considered a major strength of the study since it allows us to test the generalizability of our approach for disparate datasets.

In conclusion, we have demonstrated the feasibility of automated acute lesion segmentation of multicenter, big stroke data repositories that can be used to facilitate the genetic discovery of acute clinical MRI phenotypes. Although we did not present genetic analysis here, we have laid the foundations for future investigations of single nucleotide polymorphisms associated with large DWI lesions. High-throughput studies investigating the association between imaging phenotypes with genetics, stroke severity and long-term functional outcomes in large multi-center data sets will become realistic using automated tools powered by artificial intelligence.

**Acknowledgments:** We acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

**Sources of Funding:**

NIH: R01NS030678 (AV, DLW), R01NS039987 (BBW), R01NS042733 (BBW, JFM), R01NS059775 (OW), R01NS063925 (OW), R01NS082285 (NSR, OW), R01NS086905 (NSR, OW), R01NS100417 (AV), P50NS051343 (OW), U01NS069208 (BBW, JR, OW, SKK), U10NS077311 (AV), R37NS29993 (RLS, TR), K23NS064052 (NSR), EB015325 (MGH), 1S10RR023401 (MGH), 1S10RR019307 (MGH), 1S10RR023043 (MGH)

Evelyn F. McKnight Brain Institute: (RLS, TR)

Swedish Stroke Association: (AL, JW)

Swedish Heart and Lung Foundation: (AL, CJ)

Lund University: (AL)

Region Skåne: (AL)

Skåne University Hospital: (AL)

Freemasons Lodge of Instruction Eos in Lund: (AL)

Foundation of Färs & Frosta—one of Sparbanken Skåne’s ownership Foundations: (AL)

Swedish Research Council: (CJ)

Spanish Ministry of Science and Innovation, Instituto de Salud Carlos III Funding for Research in Health: PI051737 (JJC), PI10/02064 (JJC), PI12/01238 (JJC), PI15/00451 (JJC)

Fondos FEDER/EDRF Red de Investigación Cardiovascular: RD12/0042/0020 (JJC)

Fundació la Marató TV3: 76/C/2011 (JJC)

Recercaixa’13: JJ086116 (JJC)

Fonds voor Wetenschappelijk Onderzoek: 1841918N (RL)

Crafoord Foundation (JW)

Henry Smith Charity, Qatar National Research Fund, Stroke Association (UK), Dept of Health (UK), British Council, UK-India Education Research Initiative (UKIERI): BRAINS

**Disclosures:** None

**REFERENCES**

1. Meschia JF, Arnett DK, Ay H, Brown RD, Jr., Benavente OR, Cole JW, et al. Stroke genetics network (SiGN) study: Design and rationale for a genome-wide association study of ischemic stroke subtypes. *Stroke*. 2013;44:2694-2702

2. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet*. 2018;50:524-537

3. Network NSG, International Stroke Genetics C. Loci associated with ischaemic stroke and its subtypes (SiGN): A genome-wide association study. *Lancet Neurol*. 2016;15:174-184

4. Wu O, McIntosh E, Bezerra R, Diwan I, Garg P, Mocking S, et al. Prediction of lesion expansion in patients using acute MRI. *Stroke*. 2012;43:A3319

5. Wu O, Schwamm LH, Garg P, Pervez MA, Yoo AJ, Chautinet A, et al. Using MRI as the witness: Multimodal MRI-based determination of acute stroke onset. *Stroke*. 2010;41:E273-E273

6. Wu O, Koroshetz WJ, Østergaard L, Buonanno FS, Copen WA, Gonzalez RG, et al. Predicting tissue outcome in acute human cerebral ischemia using combined diffusion- and perfusion-weighted MR imaging. *Stroke*. 2001;32:933-942

7. Giese AK, Schirmer MD, Donahue KL, Cloonan L, Irie R, Winzeck S, et al. Design and rationale for examining neuroimaging genetics in ischemic stroke: The MRI-GENIE study. *Neurol Genet*. 2017;3:e180

8. Maguire JM, Bevan S, Stanne TM, Lorenzen E, Fernandez-Cadenas I, Hankey GJ, et al. Giscome – genetics of ischaemic stroke functional outcome network: A protocol for an international multicentre genetic association study. *European Stroke Journal*. 2017;2:229-237

9. Sorensen AG, Wu O, Copen WA, Davis TL, Gonzalez RG, Koroshetz WJ, et al. Human acute cerebral ischemia: Detection of changes in water diffusion anisotropy by using MR imaging. *Radiology*. 1999;212:785-792

10. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp*. 2002;17:143-155

11. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3d CNN with fully connected crf for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61-78

12. Giese AK, Xu H, Schirmer MD, Donahue KL, Dalca AV, Gaynor BJ, et al. Genetics of acute ischemic lesion volume: The MRI-genetics interface exploration (MRI-GENIE) study. *Stroke*. 2018;49 (Suppl 1):AWMP56

13. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327:557-560

14. UCLA Brain Mapping Center. ICBM T2 atlas. 2018

15. Mazziotta JC, Toga AW, Evans A, Fox P, Lancaster J. A probabilistic atlas of the human brain: Theory and rationale for its development. The international consortium for brain mapping (ICBM). *Neuroimage*. 1995;2:89-101

16. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage*. 2012;62:782-790

17. Andersson JLR, Jenkinson M, S. S. Non-linear registration, aka spatial normalisation. *Technical Report FMRIB Technical Report TR07JA2*. 2007

18. Chen L, Bentley P, Rueckert D. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *Neuroimage Clin*. 2017;15:633-643

19. Yoshiura T, Wu O, Zaheer A, Reese TG, Sorensen AG. Highly diffusion-sensitized MRI of brain: Dissociation of gray and white matter. *Magn Reson Med*. 2001;45:734-740

20. Luby M, Bykowski JL, Schellinger PD, Merino JG, Warach S. Intra- and interrater reliability of ischemic lesion volume measurements on diffusion-weighted, mean transit time and fluid-attenuated inversion recovery MRI. *Stroke*. 2006;37:2951-2956

21. Sakamoto Y, Koga M, Toyoda K, Osaki M, Okata T, Nagatsuka K, et al. Low DWI-ASPECTS is associated with atrial fibrillation in acute stroke with the middle cerebral artery trunk occlusion. *J Neurol Sci*. 2012;323:99-103

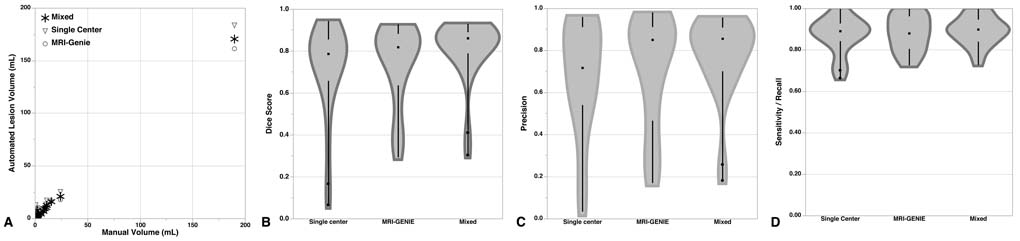
22. Henninger N, Goddeau RP, Jr., Karmarkar A, Helenius J, McManus DD. Atrial fibrillation is associated with a worse 90-day outcome than other cardioembolic stroke subtypes. *Stroke*. 2016;47:1486-1492

**TABLES**

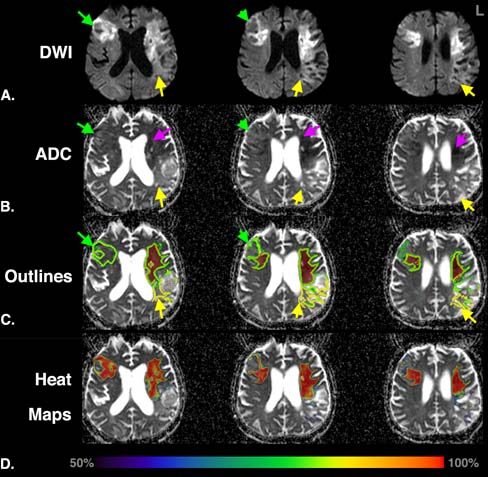
**Table 1:** Differences in automatic DWI lesion volumes (cm3) as a function of demographics and vascular risk factors. The first column (All) shows the number of patients (%) positive for the phenotype out of 2770 patients unless otherwise noted.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Characteristic | All | Yes | No | P-value |
| Sex, male | 1701 (61.4) | 3.5 [0.9-16.5] | 4.0 [0.9-16.9] | 0.99 |
| Ethnicity, Hispanic (N=2723) | 178 (6.5) | 4.9 [1.0-21.6] | 3.7 [0.9-16.6] | 0.17 |
| Race, white (N=2560) | 2287 (89.3) | 4.1 [1.0-17.9] | 2.5 [0.8-13.4] | **0.02** |
| **Medical History** |  |  |  |  |
| Hypertension (N=2749) | 1813 (66.0) | 3.4 [0.9-15.9] | 4.5 [0.9-18.0] | 0.08 |
| Diabetes Mellitus (N=2741) | 643 (23.5) | 3.2 [0.8-13.4] | 4.0 [0.9-17.4] | 0.10 |
| Atrial Fibrillation (N=2736) | 410 (15.0) | 7.3 [1.4-31.4] | 3.3 [0.8-15.1] | **<0.0001** |
| Coronary Artery Disease (N=2719) | 491 (18.1) | 4.0 [1.0-16.5] | 3.6 [0.9-16.7] | 0.42 |
| Current or former smoker | 1448 (52.3) | 3.4 [0.9-17.0] | 3.9 [0.9-16.1] | 0.93 |
| **Causative Classification of Stroke** |  |  |  |  |
| Large artery atherosclerosis | 599 (21.6) | 7.1 [1.8-25.8] | 2.9 [0.8-14.5] | **<0.0001** |
| Cardioembolic | 426 (15.4) | 8.6 [2.1-32.7] | 3.1 [0.8-14.7] | **<0.0001** |
| Small Artery Occlusion | 432 (15.6) | 0.8 [0.4-1.7] | 5.4 [1.321.3] | **<0.0001** |
| Undetermined | 1112 (40.1) | 3.5 [0.9-15.8] | 3.8 [0.9-17.8] | **0.16** |
| Other | 201 (7.3) | 7.5 [1.6-29.9] | 3.4 [0.9-16.0] | **<0.0001** |

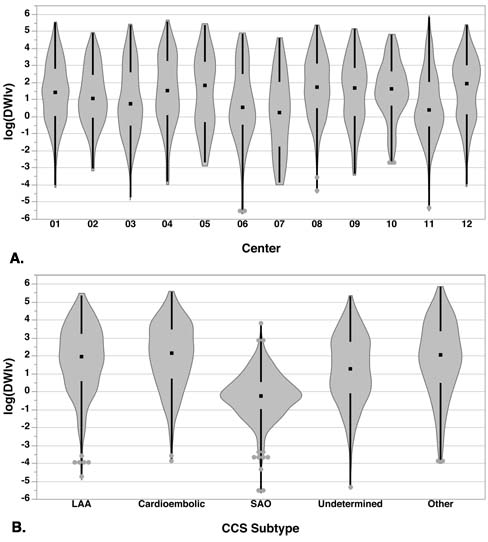
**FIGURE LEGENDS**

****

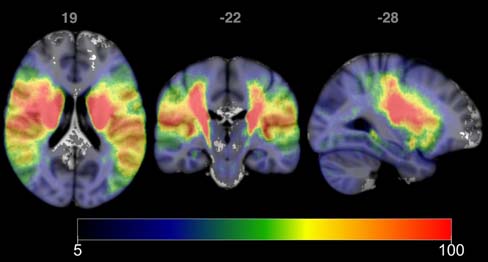
**Figure 1:** Evaluation of 3 ensemble segmentations compared to Manual Cohort B outlines. Shown are (A) correlations (Single-Center Ensemble: ρ=0.74, p<0.0001; MRI-GENIE Ensemble: ρ=0.88, p<0.0001; Mixed Ensemble: ρ=0.91, p<0.0001) and violin plots (middle dot are median values, spaces between black lines are IQR) of (B) Dice score (C) precision and (D) sensitivity. See text for details.

****

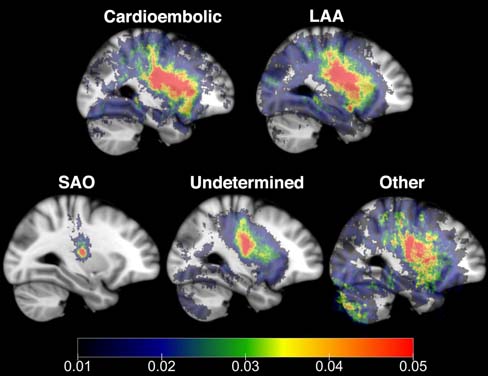
**Figure 2:** Example of discordance between Readers 1 and 2 and automated segmentations for a 76-year old man, with Undetermined stroke etiology, imaged 20 days from stroke onset. Shown are (A) iDWI and (B) ADC maps, (C) lesion segmentations (Ensemble: green outline, Reader 1: red region, Reader 2: yellow outline) and (D) ensemble probability (a. k. a. heat) maps for tissue infarction. Regions only outlined by Reader 2 (yellow arrows) have elevated ADC and iDWI. Regions where the Ensemble agreed with Reader 2 (green arrows) have pseudo-normal ADC and elevated DWI. Regions for which both readers and the ensemble model agree have reduced ADC (purple arrows) and highest probability. Lesion segmentation for this subject was complicated by the mixture of acute, subacute and chronic regions of ischemic injury that led to discordant results from similarly trained human readers. This showcases the benefit of automated algorithms that generate consistent reproducible results. Images are in radiological orientation. ADC, Apparent Diffusion Coefficient; iDWI, isotropic trace DWI.

****

**Figure 3:** Distribution of log of the DWI lesion volumes by (A) center and (B) causative classification of stroke (CCS) subtype. There was a significant difference between centers in terms of the DWI volumes (p<0.0001). Centers 6, 7 and 11 had the smallest median lesion volumes (< 2 cm3), while Centers 5 and 12 had the largest median volumes (>6 cm3). Small artery occlusion had the smallest median volume compared to all the other subtypes (p<0.0001) followed by Undetermined (p<0.0001).

****

**Figure 4:** Incidence maps for 2763 patients. Each voxel shows number of patients with lesions within each voxel using range 5 to 100. Maximum incidence was 295. Images are shown in radiological format.

****

**Figure 5:** Frequency maps (incidence map/number of subjects) showing topological distribution of infarcts across different stroke subtypes – cardioembolic (N=425), LAA (N=597), SAO (N=431), Other (N=200), and Undetermined (N=1110). SAO frequency map had the lowest correlation compared with the other subtypes, followed by Other. Images are shown in radiological format.