Title Page

Title: Enhanced physician performance when using an artificial intelligence model to detect ischemic stroke on computed tomography

Authors:
James M Hillis, MBBS DPhil*[1,2,3]
Bernardo C Bizzo, MD PhD*[1,3,4]
Romane Gauriau, PhD[1]
Christopher P Bridge, DPhil[1,3,5]
John K Chin, MD[1]
Buthaina Hakamy, MBBS[1]
Sarah Mercaldo, PhD[1,3,4]
John Conklin, MD[3,4]
Sayon Dutta, MD MPH[3,6]
William A Mehan, MD MBA[1,3,4]
Robert W Regenhardt, MD PhD[2,3]
Ajay Singh, MD[3,4]
Aneesh B Singhal, MBBS MD[2,3]
Jonathan D Sonis, MD MHCM[3,6]
Marc D Succi, MD[3,4]
Tianhao Zhang, PhD[7]
Bin Xing, PhD[7]
John F Kalafut, PhD[7]
Keith J Dreyer, DO PhD[1,3,4]
Michael H Lev, MD[3,4]
R Gilberto González, MD PhD[3,4,5]

*These authors contributed equally to the work.

Author Affiliations:
[1]Data Science Office, Mass General Brigham, Boston, MA, USA
[2]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
[3]Harvard Medical School, Boston, MA, USA
[4]Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
[5]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital
[6]Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, USA
[7]GE Healthcare

Corresponding author:
Dr. James Hillis
Digital CRO, Data Science Office, Mass General Brigham
Suite 1303, Floor 13, 100 Cambridge St, Boston, MA 02114
james.hillis@mgh.harvard.edu
617-726-2000

45

46 Date of revision: January 16 2023

Abstract (word count: 55)

Acute ischemic stroke can be subtle to detect on non-contrast computed tomography imaging. We show that a novel artificial intelligence model significantly improves the performance of physicians, including ED physicians, neurologists and radiologists, in identifying and quantifying the volume of acute ischemic stroke lesions. This model may lead to improved clinical decision-making for stroke patients.

Article (word count: 1,118)

The early imaging features of acute ischemic stroke on non-contrast computed tomography (CT) can be subtle. We previously reported the development of a deep learning model that detects ischemic core on CT studies and was superior to three experienced neuroradiologists.[1] As part of its training, the model utilized segmentations obtained from the region of acute infarct on paired magnetic resonance imaging (MRI) studies that were then registered onto the CT studies. This design had recognized that MRI better detects early acute ischemic stroke but CT is a cheaper, quicker and more widely available imaging modality.

In further evaluating this model, we wished to see how its use impacted physicians who were interpreting non-contrast CT studies. We therefore designed a multi-reader multi-case study whereby physicians interpreted 180 CT cases both with and without the use of the model (Supplementary Figure 1). There were 8 physicians (2 emergency physicians, 2 emergency radiologists, 2 neurologists, 2 neuroradiologists) who each interpreted 90 cases with the model output and 90 cases without the model output (Figure 1A). After a four-week washout period, they interpreted the cases in the opposite manner (i.e., the cases that they had previously interpreted without the model output were now interpreted with the model output).

The model output included the binary classification of whether ischemic core ≥5mL was present. If there was ischemic core ≥5mL, the output also included the volume of the ischemic core and the segmented region of the ischemic core. The physicians were asked up to three multiple choice questions: the presence of any ischemic core (positive, negative), their level of

certainty for the presence or absence (scale of 1-7; see Supplementary Table 1 for details), and, if they stated the presence of ischemic core, their volume estimate (0-20mL, 20-50mL, >50mL; upper bound of range considered inclusive). Their responses were compared to the ground truth interpretations demonstrated by diffusion imaging on a paired MRI case.

The physicians performed significantly better at binary detection when using the model compared with not using the model. Their area under the receiver operating characteristic curve (AUC) improved from 0.696 to 0.836 (difference 0.141; 95% CI: 0.081-0.200; p<0.001; Figure 1B and Table 1). Their sensitivity improved from 53.9% to 66.9% (difference 13.1%; 95% CI: 6.5-19.6%; p<0.001). Their specificity improved from 78.9% to 89.2% (difference 10.3%; 95% CI: 6.6-14.0%; p<0.001). Their interpretation time also improved from 62.30 seconds to 43.92 seconds (difference 18.38 seconds; 95% CI: 15.59-21.18 seconds; p<0.001).

These improvements were maintained across most specialties as part of a subgroup analysis (Table 1). The physicians with the greatest improvement were emergency physicians and emergency radiologists, which is consistent with them having the least experience in interpreting brain imaging and therefore having the greatest opportunity to benefit. The physicians with the least improvement were neuroradiologists, who have the most experience.

The improvement in sensitivity was most pronounced for larger infarct volumes (Table 1). The physicians' sensitivity improved from 54.2% without the model output to 84.6% with the model output for 20-50mL infarcts, and 61.7% to 84.2% for >50mL infarcts. These large infarcts are

more likely to involve a large vessel occlusion, which means the patients could benefit from treatment with endovascular thrombectomy.[2-8] The increased ability to detect these infarcts could prompt physicians to obtain CT angiography to detect a large vessel occlusion and cue the physicians interpreting the CT angiography to the likely vessel involved. This increased ability to detect may also lead to improved triage and sooner evaluation for endovascular thrombectomy. A decreased time to thrombectomy has previously been shown to improve outcomes.[9]

There was, however, a decline in sensitivity for infarct volumes 0-5mL (Table 1). This decreased performance was expected given that the model only outputs whether it has detected ischemic core ≥5mL (i.e., so the model should classify these cases as negative). The reason for the model using this volume threshold is to avoid false positive interpretations from small regions of noise on a CT study; the clinical concern is for such interpretations to increase MRI utilization to evaluate for infarct more definitively. While we acknowledge the subsequent decreased sensitivity amongst physicians for these infarct volumes, it is important to recognize that this study occurred outside of the clinical environment where other factors, especially the acute onset of neurologic symptoms, could alert physicians to the occurrence of ischemic stroke.

The physicians also performed better at volume quantification when using the model (Figure 1C). They correctly identified the volume range 73.8% of the time for >50mL infarcts with the model output compared to 27.1% without the model output. They correctly identified the

volume range 58.8% of the time for 20-50mL infarcts with the model output compared to 30.4% without the model output.

A key consideration moving forward is how the model might impact clinical workflow. Currently the clinical paradigm is that an ischemic stroke should be assumed when a patient presents with stroke-like symptoms and a non-contrast CT does not reveal an abnormality; the key reason for obtaining the CT is to exclude intracranial hemorrhage. This study demonstrates the model improves physicians' detection of ischemic core. Further research should be conducted to determine whether the CT could be used more than it currently is for confirmation of ischemic core. For instance, confirmation of ischemic core on non-contrast CT could be particularly helpful when a patient's symptoms are ambiguous and ischemic stroke has not been considered as a likely differential diagnosis. The presence of ischemic core could help trigger and triage the next management steps such as CT angiography, administration of thrombolytic medication and consideration of endovascular thrombectomy.[10]

The model most helped physicians with less experience in interpreting brain imaging and may be most beneficial in rural areas with fewer subspecialty physicians. The model may separately be most helpful at hospitals that can perform non-contrast CT but do not have emergent access to advanced imaging modalities like CT perfusion or MRI. It may therefore assist in reducing the urban-rural inequities in acute stroke care.[11]

A limitation of this study was that the cases were taken from the test set from model development. While these cases were sequestered and not exposed to the model during development, they are likely to be the most similar to the training cases and therefore provide the best model performance. The assessment of generalizability of the model will benefit from evaluation on a more diverse dataset prior to clinical use. We note that the physicians were not aware of the standalone model performance on these cases during this study.

Overall, this study demonstrates how the use of an artificial intelligence model enhances physicians' identification and volume quantification of ischemic core on non-contrast CT. It suggests that the model could provide benefit in the acute stroke clinical environment.

Methods

Study design

This retrospective multi-reader multi-case study was conducted using radiology cases from hospitals within the Mass General Brigham network. It was approved by the Mass General Brigham Institutional Review Board with waiver of informed consent per the Common Rule. It was conducted in accordance with relevant guidelines and regulations including the Health Insurance Portability and Accountability Act (HIPAA). This report followed the Standards for Reporting Diagnostic Accuracy (STARD 2015) reporting guideline.

Case selection and model inference

The 180 cases were selected from the test set that had been used at the time of model development.[1] As described previously, they had accompanying MRI studies that were used to establish the ground truth interpretations (within 3 hours of the CT for positive cases and 5 days for negative cases). They were selected using stratified randomization from the entire primary test set such that there were 30 cases with ischemic core 0-20mL, 30 cases with ischemic core 20-50mL, 30 cases with ischemic core >50mL and 90 cases without any ischemic core. This randomization ensured there were cases on which the model performed accurately and inaccurately (see Supplementary Table 2 for standalone model performance on these cases).

For the studies without the model output, the physicians were provided with only the axial 5mm series. For the studies with the model output, the physicians were provided with the axial

5mm series and an identical series with the model output incorporated into the imaging pixel data (Figure 1A). The physicians were able to visualize these two series simultaneously in adjacent window panes. The outputs included the binary classification of whether ischemic core ≥5mL was present, and, if it was present, the volume of the ischemic core and the segmented region of the ischemic core. The volume threshold of 5mL is proposed for future clinical use given it optimizes specificity by avoiding false positive interpretations through incorrect interpretation of small regions of noise on CT.

Physicians

The physicians were chosen to ensure representation of likely future clinical users including emergency physicians, emergency radiologists, neurologists and neuroradiologists. They were all board-certified for their relevant specialty. They were trained on the annotation tasks and completed twelve training cases. They had not been involved with model development and were not aware of prior model performance results. They were informed that specificity was prioritized over sensitivity as part of the model design given that future clinical users should similarly be aware of this fact.

Reader study process

The multi-reader multi-case study design involved the physicians interpreting all radiology studies twice: both with and without the model output (Supplementary Figure 1). The interpretations were performed as part of two sessions that were separated by at least four weeks as recommended by the US Food and Drug Administration[12]; each study was interpreted

once in each session. Within each session, half of the studies were interpreted with the model output and half of the studies without the model output; the studies were interpreted in the opposite manner for the other session. The studies were split into two batches to facilitate these interpretations. The order of studies within each session was randomized and differed for each physician. The interpretation of the first batch of studies with or without model output for each physician was also randomized.

The physicians were assisted in working through the cases by an internal web-based annotation system that required them to interpret the cases in the defined order. This annotation system incorporated the FDA-cleared eUnity image visualization software (Version 6 or higher). The physicians were firstly asked about the presence of any ischemic core (positive, negative). They were then asked their level of certainty for this ischemic core on a scale of 1-7 (see Supplementary Table 1 for options). If they stated that ischemic core was present, they were also asked to estimate the volume (0-20mL, 20-50mL, >50mL; the upper bound of these ranges was considered inclusive). The annotation system also recorded the start and stop times for the interpretation of each case. When submitting the annotations for each case, the physicians could opt to move to the next case or exit; they did not need to interpret all cases in a single sitting.

Statistical techniques

This study was a pilot multi-reader multi-case study for this model. The predefined primary assessment was comparison of the AUC with and without the model outputs. The predefined

secondary assessments were comparison of sensitivity, specificity and interpretation time with and without the model outputs. The specialty subgroup analysis, infarct volume subgroup analysis and volume quantification analysis were calculated as exploratory assessments. Given the pilot nature of this study, there was not an estimated effect size and powering was not performed. There were no missing data with the exception of the excluded interpretation times as described below.

The selection of cases involved multiple randomization procedures including for the cases selected in each batch, the interpretation order for each physician, and whether the first batch for a physician was with or without model output. This randomization used the Latin squares methodology.

The analysis of the AUC was based on the 7-point scale assessments from each physician and the ground truth interpretations based on the MRI. The overall comparison between the two methods (with and without model outputs) was derived from the difference between the mean AUCs[13] via analysis of variance (ANOVA), taking into account the variability components: methods, physicians, cases and their interactions. The analysis was implemented by using the OR-DBM MRMC software package version 2.5 or higher.[14,15] The ANOVA model treated both physicians and cases as random samples, to be able to draw inferences for the whole population of physicians and cases. Two-sided p-values were based on 95% confidence intervals of the difference in AUCs.

The analysis of sensitivity and specificity used the binary interpretations from each physician and the ground truth interpretations based on the MRI. The comparison between the two methods (with and without model outputs) was based on the generalized estimating equations (GEE) model[16] by using SAS procedure PROC GENMOD, taking into account repeated observations from the MRMC design. Based on SAS GENMOD, case was treated as the repeated subject, accounting for correlations of physician and method within case, by using an exchangeable covariance structure. Two-sided p-values were based on 95% confidence intervals of the difference in sensitivity and specificity.

The analysis of interpretation time was performed in SAS using a repeated measures ANOVA. Two-sided p-values were based on 95% confidence intervals of the difference in interpretation time. Thirteen interpretation times (from 2880 interpretations) were excluded; the paired study for the same physician (i.e., the equivalent case with or without model output) was also excluded to ensure balance of such exclusions. These exclusions occurred for two reasons. Firstly, the physicians could notify study management if an interpretation time should be excluded (e.g., they were interrupted). Secondly, the annotation system initially recorded the start time for each case incorrectly; to ensure consistency across the entire cohort, the interpretation time was rederived by calculating the difference in start time with the subsequent case or the difference in stop time with the previous case (the minimum of these times was taken); the interpretation times were excluded if this derived number did not appear consistent with an expected duration (e.g., it appeared a physician only interpreted one case in a sitting so the difference in times with the previous and subsequent cases would be incorrect).

The analysis of volume outputs used the volume estimates from each physician and the ground truth interpretations based on the MRI. The frequencies of how the volume estimates matched with ranges of ground truth volumes of the two methods (with and without model outputs) were calculated using Microsoft Excel.

Acknowledgements

The authors thank the broader Mass General Brigham Data Science Office and GE Healthcare teams for their assistance with this project.

Funding

This study was funded by GE Healthcare. JMH, BCB, RG, CPB, JKC, BH, SM, JC, SD, WAH, RWR, AS, ABS, JDS, MDS, KJD, MHL, RGG were employees of Mass General Brigham and/or Massachusetts General Hospital at the time of this study, which had received institutional funding from GE Healthcare for the study. TZ, BX, JFK were employees of GE Healthcare at the time of this study.

Competing interests

RWR reports the following competing interests: Rapid Medical – Clinical Trial DSMB; Microvention – Site PI; Penumbra – Site PI; National Institute of Neurological Disorders and Stroke – Research Grant; Society of Vascular and Interventional Neurology – Research Grant; Heitman Foundation – Research Grant.

Data availability

The data used were obtained from hospitals within the Mass General Brigham network. Data use was approved by relevant Institutional Review Board. The data are not publicly available and restrictions apply to their use.

Author contributions

JMH, BCB, JKC, TZ, JFK, KJD, MHL and RGG conceptualized the study. JMH, BCB, RG, CPB, JKC and BH conducted the study including producing the model output. JC, SD, WAH, RWR, AS, ABS, JDS and MDS were physician readers. SM, TZ and BX performed the randomization and statistical analysis. JMH drafted the manuscript text and figures. All authors reviewed the manuscript.

References

1 Bizzo, B. C. *et al. Head CT Deep Learning Model for Early Stroke Identification Outperforms Human Experts* (Research Square, 2021).

2 Berkhemer, O. A. *et al.* A randomized trial of intraarterial treatment for acute ischemic stroke. *N Engl J Med* **372**, 11-20, doi:10.1056/NEJMoa1411587 (2015).

3 Goyal, M. *et al.* Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med* **372**, 1019-1030, doi:10.1056/NEJMoa1414905 (2015).

4 Saver, J. L. *et al.* Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *N Engl J Med* **372**, 2285-2295, doi:10.1056/NEJMoa1415061 (2015).

5 Campbell, B. C. *et al.* Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med* **372**, 1009-1018, doi:10.1056/NEJMoa1414792 (2015).

6 Jovin, T. G. *et al.* Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N Engl J Med* **372**, 2296-2306, doi:10.1056/NEJMoa1503780 (2015).

7 Nogueira, R. G. *et al.* Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct. *N Engl J Med* **378**, 11-21, doi:10.1056/NEJMoa1706442 (2018).

8 Albers, G. W. *et al.* Thrombectomy for Stroke at 6 to 16 Hours with Selection by Perfusion Imaging. *N Engl J Med* **378**, 708-718, doi:10.1056/NEJMoa1713973 (2018).

9 Jahan, R. *et al.* Association Between Time to Treatment With Endovascular Reperfusion Therapy and Outcomes in Patients With Acute Ischemic Stroke Treated in Clinical Practice. *JAMA* **322**, 252-263, doi:10.1001/jama.2019.8286 (2019).

10 Powers, W. J. *et al.* Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke* **50**, e344-e418, doi:10.1161/STR.0000000000000211 (2019).

11 Hammond, G., Luke, A. A., Elson, L., Towfighi, A. & Joynt Maddox, K. E. Urban-Rural Inequities in Acute Stroke Care and In-Hospital Mortality. *Stroke* **51**, 2131-2138, doi:10.1161/STROKEAHA.120.029318 (2020).

12 US Food and Drug Administration. *Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in Premarket Notification (510(k)) Submissions*, <https://www.fda.gov/media/77642/download> (2022).

13 Hillis, S. L., Obuchowski, N. A., Schartz, K. M. & Berbaum, K. S. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. *Stat Med* **24**, 1579-1607, doi:10.1002/sim.2024 (2005).

14 Hillis, S. L., Berbaum, K. S. & Metz, C. E. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol* **15**, 647-661, doi:10.1016/j.acra.2007.12.015 (2008).

15 Hillis, S. L. & Schartz, K. M. Multireader sample size program for diagnostic studies: demonstration and methodology. *J Med Imaging (Bellingham)* **5**, 045503, doi:10.1117/1.JMI.5.4.045503 (2018).

16 *Generalized Estimating Equations*,
<https://support.sas.com/rnd/app/stat/topics/gee/gee.pdf> (

Figure Legends

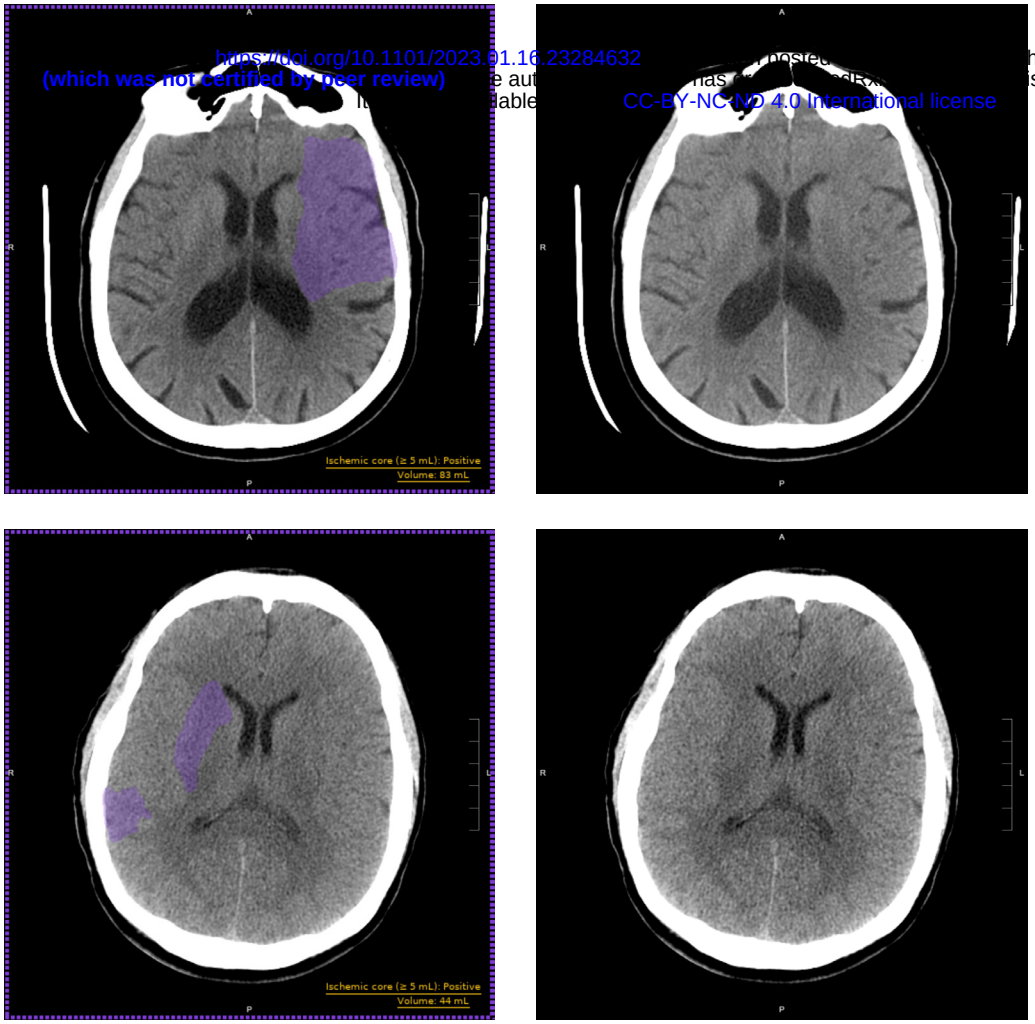
Figure 1: A. Example of two cases with and without model output showing text incorporating binary classification for ischemic core ≥5mL and volume, and purple segmented region. The first case, with infarct volume 77mL on MRI, was detected by 1 out of the 8 physicians without model output and 6 with model output. The second case, with infarct volume 45mL on MRI, was detected by 2 without model output and 8 with model output. **B.** Receiver operating characteristic curves for both with model output and without model output. **C.** Confusion matrices comparing physician volume with ground truth volume; a correct volume estimate occurs on the diagonal; a volume of 0mL reflects a negative case.

Tables

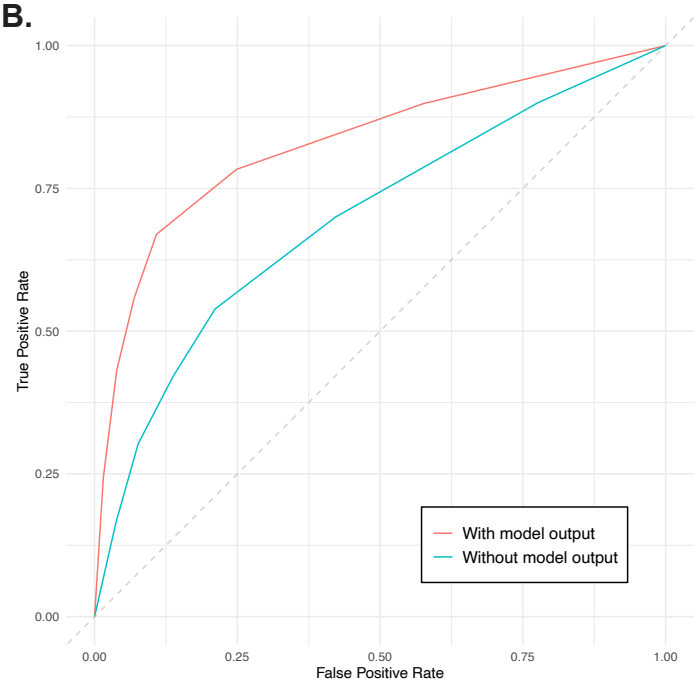
Table 1: Results summary for physician performance comparing interpretations performed with or without the model outputs.

	$Metric_{with}$	$Metric_{without}$	Difference (95% CI)	p value
AUC				
Overall	0.836	0.696	0.141 (0.081 to 0.200)	<0.001
ED Physician	0.820	0.632	0.188 (0.119 to 0.256)	<0.001
ED Radiologist	0.805	0.617	0.188 (0.098 to 0.279)	0.003
Neurologist	0.854	0.738	0.116 (0.081 to 0.151)	0.001
Neuroradiologist	0.866	0.796	0.070 (-0.337 to 0.477)	0.328
Sensitivity				
Overall	66.9%	53.9%	13.1% (6.5 to 19.6%)	<0.001
ED Physician	64.4%	43.9%	20.6% (10.4% to 30.8%)	<0.001
ED Radiologist	74.4%	48.9%	25.6% (16.0% to 35.1%)	<0.001
Neurologist	70.6%	69.4%	1.1% (-8.0% to 10.3%)	0.814
Neuroradiologist	58.3%	53.3%	5.0% (-1.0% to 11.4%)	0.123
Specificity				
Overall	89.2%	78.9%	10.3% (6.6 to 14.0%)	<0.001
ED Physician	91.1%	78.3%	12.8% (6.1% to 19.4%)	<0.001
ED Radiologist	76.7%	72.8%	3.9% (-4.7% to 12.5%)	0.376
Neurologist	91.1%	69.4%	21.7% (15.5% to 27.8%)	<0.001
Neuroradiologist	97.8%	95.0%	2.8% (-0.4% to 6.0%)	0.090
Interpretation time				
Overall	43.92s	62.30s	-18.38s (-21.18s to -15.59s)	<0.001
ED Physician	40.74s	61.91s	-21.17s (-25.95s to -16.38s)	<0.001
ED Radiologist	42.01s	70.58s	-28.57s (-37.03s to -20.10s)	<0.001
Neurologist	40.59s	60.79s	-20.20s (-24.32s to -16.07s)	<0.001
Neuroradiologist	52.33s	55.94s	-3.60s (-7.12s to -0.09s)	0.044
Sensitivity for different infarct volumes (based on MRI ground truth)				
0-5mL	16.0%	34.7%	-18.8% (-28.2% to -9.3%)	<0.001
5-20mL	56.3%	62.5%	-6.3% (-21.1% to 8.6%)	0.410
20-50mL	84.6%	54.2%	30.4% (23.6% to 37.3%)	<0.001
>50mL	84.2%	61.7%	22.5% (14.1% to 30.9%)	<0.001

A.



B.



C.

With model output

			Physician volume			
		N	0mL	0-20mL	20-50mL	>50mL
Ground truth (MRI) volume	0mL	720	89.2%	6.4%	3.8%	0.7%
	0-20mL	240	67.9%	20.4%	9.6%	2.1%
	20-50mL	240	15.4%	16.7%	58.8%	9.2%
	>50mL	240	15.8%	2.5%	7.9%	73.8%

Without model output

			Physician volume			
		N	0mL	0-20mL	20-50mL	>50mL
Ground truth (MRI) volume	0mL	720	78.9%	10.7%	6.9%	3.5%
	0-20mL	240	54.2%	26.7%	16.3%	2.9%
	20-50mL	240	45.8%	17.1%	30.4%	6.7%
	>50mL	240	38.3%	16.7%	17.9%	27.1%