AJNR

Evaluation of an artificial intelligence model for identification of mass effect and vasogenic edema on computed tomography of the head

Isabella Newbury-Chaet, Sarah F. Mercaldo, John K. Chin, Ankita Ghatak, Madeleine A. Halle, Ashley L. MacDonald, Karen Buch, John Conklin, William A. Mehan, Stuart Pomerantz, Sandra Rincon, Keith J. Dreyer, Bernardo C. Bizzo and James M. Hillis

AJNR Am J Neuroradiol published online 28 May 2024
http://www.ajnr.org/content/early/2024/05/28/ajnr.A8358

Published May 28, 2024 as 10.3174/ajnr.A8358

ORIGINAL RESEARCH

Evaluation of an artificial intelligence model for identification of mass effect and vasogenic edema on computed tomography of the head

Isabella Newbury-Chaet, Sarah F. Mercaldo, John K. Chin, Ankita Ghatak, Madeleine A. Halle, Ashley L. MacDonald, Karen Buch, John Conklin, William A. Mehan, Stuart Pomerantz, Sandra Rincon, Keith J. Dreyer, Bernardo C. Bizzo*, James M. Hillis*

ABSTRACT

BACKGROUND AND PURPOSE: Mass effect and vasogenic edema are critical findings on CT of the head. This study compared the accuracy of an artificial intelligence model (Annalise Enterprise CTB) to consensus neuroradiologist interpretations in detecting mass effect and vasogenic edema.

MATERIALS AND METHODS: A retrospective standalone performance assessment was conducted on datasets of non-contrast CT head cases acquired between 2016 and 2022 for each finding. The cases were obtained from patients aged 18 years or older from five hospitals in the United States. The positive cases were selected consecutively based on the original clinical reports using natural language processing and manual confirmation. The negative cases were selected by taking the next negative case acquired from the same CT scanner after positive cases. Each case was interpreted independently by up to three neuroradiologists to establish consensus interpretations. Each case was then interpreted by the AI model for the presence of the relevant finding. The neuroradiologists were provided with the entire CT study. The AI model separately received thin (≤1.5mm) and/or thick (>1.5 and ≤5mm) axial series.

RESULTS: The two cohorts included 818 cases for mass effect and 310 cases for vasogenic edema. The AI model identified mass effect with sensitivity 96.6% (95% CI, 94.9-98.2) and specificity 89.8% (95% CI, 84.7-94.2) for the thin series, and 95.3% (95% CI, 93.5-96.8) and 93.1% (95% CI, 89.1-96.6) for the thick series. It identified vasogenic edema with sensitivity 90.2% (95% CI, 82.0-96.7) and specificity 93.5% (95% CI, 88.9-97.2) for the thin series, and 90.0% (95% CI, 84.0-96.0) and 95.5% (95% CI, 92.5-98.0) for the thick series. The corresponding areas under the curve were at least 0.980.

CONCLUSIONS: The assessed AI model accurately identified mass effect and vasogenic edema in this CT dataset. It could assist the clinical workflow by prioritizing interpretation of abnormal cases, which could benefit patients through earlier identification and subsequent treatment.

ABBREVIATIONS: AI = artificial intelligence; AUC = area under the curve; CADt = computer assisted triage devices; FDA = Food and Drug Administration; NPV = negative predictive value; PPV = positive predictive value; SD = standard deviation.

From the Data Science Office, Mass General Brigham, Boston, MA, USA (I.N-C., S.F.M., J.K.C., A.G., M.A.H., A.L.M, S.P., K.J.D., B.C.B., J.M.H.), Harvard Medical School, Boston, MA, USA (S.F.M., K.B., J.C., W.A.M., S.P., S.R., K.J.D., B.C.B., J.M.H.), Department of Radiology, Massachusetts General Hospital, Boston, MA, USA (S.F.M., K.B., J.C., W.A.M., S.P., S.R., K.J.D., B.C.B.), Department of Neurology, Massachusetts General Hospital, Boston, MA, USA (J.M.H.). B.C.B. and J.M.H. share senior authorship.

Disclosures: I.N-C., S.F.M., J.K.C., A.G., M.A.H., A.L.M., K.B., J.C., W.A.M., S.P., S.R., K.J.D., B.C.B., J.M.H. are employees of Mass General Brigham or Massachusetts General Hospital, which had received institutional funding from Annalise-AI for the study

Please address correspondence to James M. Hillis, MBBS DPhil, Data Science Office, Mass General Brigham, Suite 1303, Floor 13, 100 Cambridge St, Boston, MA 02114; james.hillis@mgh.harvard.edu.

SUMMARY SECTION

PREVIOUS LITERATURE: The use of artificial intelligence (AI) algorithms to triage and prioritize head CT cases with large vessel occlusion or intracranial hemorrhage is well established. Mass effect and vasogenic edema are similarly critical findings that may require emergent attention, yet there are fewer AI algorithms that identify them. This algorithm, which received US Food and Drug Administration clearance for the identification of both findings as part of computer assisted triage devices (CADt), was the first to do so for vasogenic edema.

KEY FINDINGS: This standalone model performance assessment demonstrated sensitivity and specificity of at least 89.8% for the identification of each of mass effect and vasogenic edema by an AI algorithm. This performance occurred in both thin and thick series, and a similar performance was maintained across various demographic and technical subgroups.

KNOWLEDGE ADVANCEMENT: The ability to identify these findings could assist the clinical workflow through prioritizing the interpretation of abnormal cases. The growing number of findings identified by CADt devices also broadens the pool of patients who could benefit from them.

Copyright 2024 by American Society of Neuroradiology.

INTRODUCTION

Mass effect and vasogenic edema are critical findings on CT of the head and require emergent medical attention.[1-3] Mass effect can be caused by various pathologies including tumor and hemorrhage. It manifests on CT with effacement of ventricles, basal cisterns or cerebral sulci; midline shift; and brain herniation including tonsillar herniation or uncal herniation. Vasogenic edema can similarly be caused by various pathologies and manifests as a deep white matter hypodensity extending into subcortical white matter.

Like the identification of large vessel occlusion and intracranial hemorrhage on head CT,[4, 5] the identification of mass effect and vasogenic edema by artificial intelligence (AI) algorithms could assist clinical care by triaging suspected cases for sooner interpretation and enabling sooner treatment. While there have been at least fifteen computer assisted triage devices (CADt) cleared by the US Food and Drug Administration (FDA) for intracranial hemorrhage, there are far fewer for mass effect and vasogenic edema.[6-8] This paper describes the performance of the Annalise Enterprise (CTB module) device, which is available in many non-US regulatory jurisdictions and can identify 130 different radiological findings on head CT.[9] In the US, the two findings of mass effect and vasogenic edema have received FDA clearance as separate devices including as the first CADt device to identify vasogenic edema.[10, 11]

This study was a standalone model performance assessment for the identification of mass effect and vasogenic edema: it compared the accuracy of the AI device to consensus neuroradiologist interpretations in detecting these findings. Similar to a prior study for intracranial hemorrhage,[12] the device was provided separately with thin (≤1.5mm) and/or thick (>1.5 and ≤5mm) axial series from each case so that the performance on different slice thicknesses could be calculated. The performance was also calculated for cases belonging to demographic and technical subgroups to determine the generalizability of the device.

MATERIALS AND METHODS

Study design

This retrospective standalone model performance study was conducted using radiology cases from five hospitals within the Mass General Brigham network between 2016 and 2022 using similar methods to a previously published study about intracranial hemorrhage identification.[12] The study examined the performance for the binary detection of mass effect and vasogenic edema by the AI model. It was approved by the Mass General Brigham Institutional Review Board with waiver of informed consent. It was conducted in accordance with relevant guidelines and regulations including the Health Insurance Portability and Accountability Act. This manuscript follows the Standards for Reporting Diagnostic Accuracy (STARD 2015) reporting guideline.

Case selection

The cohorts for mass effect and vasogenic edema were selected in a consecutive manner based on the original radiology reports. The cohort size for each of the positive and negative cases was based on powering calculations as described in the statistical analysis section below. The positive cases were identified through a natural language processing search engine (Nuance mPower Clinical Analytics) followed by manual report review. The negative cases were identified by taking the next negative case acquired on the same CT scanner after the positive cases to avoid temporal and technical bias. The next negative cases were taken after every Nth positive case based on the ratio of positive to negative cases to ensure the principles of consecutive selection applied.

The cohort considered all CT head cases performed at a hospital including inpatient and outpatient; there were no limitations on the original CT head clinical indication. The CT head cases were obtained from patients at least 18 years of age. The CT head cases were taken from unique patients; only the first CT head from a given patient was included. It was possible for a case to be included in both cohorts (i.e., both mass effect and vasogenic edema); there were 8 cases in both cohorts.

All cases were deidentified and underwent an image quality review by an American Board of Radiology-certified neuroradiologist. The relevant series for the model interpretations were selected at the same time as described under the series selection section below. The review was performed using the FDA-cleared eUnity image visualization software (Version 6 or higher) and an internal web-based annotation system that utilized the REDCap electronic data capture tools hosted at Mass General Brigam.[13, 14]

Series selection

The model was provided with a single selected series at the time of model inference. These series were non-contrast thin (≤1.5mm) and/or thick (>1.5 and ≤5mm) axial series for each CT head case. The series were selected such that the thin series was the thinnest available series ≤1.5mm; the thick series was randomized between the thinnest and thickest available series >1.5 and ≤5mm to ensure representation of series thicknesses across the entire range. The series were selected at the same time as the image quality review. After series selection, a DICOM metadata review was additionally performed to ensure that the slice thickness was within the appropriate range and that there was a consistent slice interval (with tolerance of ±0.2mm).

Ground truth interpretations

Ground truth interpretations were performed by up to three ABR-certified neuroradiologists. They answered whether the relevant finding was “Present” or “Absent”. The definition for mass effect was “mass effect as evidenced by effacement of ventricles, basal cisterns or

cerebral sulci, midline shift, or brain herniation (e.g., tonsillar herniation or uncal herniation)". The definition of vasogenic edema was "deep white matter hypodensity extending into subcortical white matter". The neuroradiologists also answered whether a "parenchymal abnormality including ischemia / mass / cyst / encephalomalacia" was present. They provided their interpretations independently, without access to the original radiology reports and in different worklist orders. They used the same image visualization software and annotation system as was used in the image quality review. They had access to the entire CT head case (i.e., were not restricted to the series selected for model inference). For determining consensus interpretations for the presence of mass effect or vasogenic edema, a "2+1" strategy was used: the first two neuroradiologists interpreted every case and a third neuroradiologist then interpreted cases with discrepant interpretations. A parenchymal abnormality was considered present if either of the first two neuroradiologists annotated it as present; the third neuroradiologist was not asked about its presence.

Model inference

The evaluated AI model was version 3.1.0 of the Annalise Enterprise CTB Triage Trauma device. It is the same AI model used by the Annalise Enterprise (CTB module) device, which is commercially available in some non-US markets and whose development has been previously described.[9] In brief, it consists of an ensemble of five neural networks with three heads: one for classification, one for left-right localization and one for segmentation. It can identify 130 different radiological findings and was trained on over 200,000 CT head cases, which were each labelled by at least three radiologists. These training cases came from 8 different scanner manufacturers and over 90 different scanner models. The training cases were completely independent of the cases used for this standalone model performance study.

The Annalise Enterprise CTB Triage Trauma device only provides binary classification outputs about the identification of findings, which is consistent with FDA regulations for CADt devices. The model was installed at MGB for use in this study and received only the DICOM-formatted CT head cases. It outputted a classification score between 0 and 1 for mass effect and vasogenic edema. A binary output for these findings could be derived using prespecified operating points. As part of model inference, the device contains multiple filters to look at attributes of the series to be interpreted to ensure the model does not perform inference on unsuitable images; in these situations, the device does not produce an output and is referred to as "unsuccessful model inference" within this study. While not part of the current study, internal bench testing indicated a model turn-around time of 81.6 seconds (95% CI: 80.3 to 82.9 seconds).[10, 11]

Statistical analysis

The statistical analysis was performed in R (version 4.0.2) on the full analysis set. The predefined endpoints included the areas under the receiver operating characteristic curves (AUCs) for the identification of mass effect and vasogenic edema for each of thin and thick series. The AUCs were calculated using the consensus annotations and the classification scores from the AI model. The prespecified endpoints also included the sensitivity and specificity at predetermined operating points; this paper reports the performance at operating points that have received FDA clearance. They were calculated by comparing the binary model output at each operating point with the consensus annotations (i.e., by calculating the number of true positive, false negative, true negative and false positive cases).

The positive predictive values (PPVs) and negative predictive values (NPVs) were calculated as exploratory analyses at an assumed prevalence of 0.05, 0.10, 0.15 and 0.20. The sensitivities and specificities were calculated as exploratory analyses for the subgroups of presence or absence of a parenchymal abnormality. The AUCs, sensitivities and specificities were calculated as exploratory analyses for the subgroups of sex, age, ethnicity, race and CT scanner manufacturer. These parameters were derived from clinical databases or DICOM fields for each radiology case. Any missing data were treated as "Unknown" or "Unavailable" and no data were imputed.

All CIs were calculated using bootstrapped intervals with 2,000 resamples. The sample sizes for each of the findings were powered based on preliminary model results at a balanced operating point to ensure the lower bound of the 95% CI for sensitivity was >80% and for specificity was >80%.

RESULTS

Mass effect

A cohort of 818 CT head cases were selected for the mass effect cohort (Figure 1). This cohort resulted in 650 thin series and 816 thick series for which the model could be evaluated.

Thin series

The model successfully performed inference on 632 (97.2%) thin series. This cohort for analysis included 306 (48.4%) women and 326 (51.6%) men; mean (standard deviation [SD]) age was 67.2 (± 17.5) years; there were 495 (78.3%) positive cases and 137 (21.7%) negative cases (Online Supplemental Data). The AI model identified mass effect with AUC of 0.987 (95% CI: 0.979-0.993; Figure 2A and Table 1) and, at an operating point of 0.221484, the sensitivity was 96.6% (95% CI: 94.9-98.2%) and the specificity was 89.8% (95% CI: 84.7-94.2%). At an assumed prevalence of 0.10, the PPV was 51.2% (95% CI: 40.9-66.7%) and NPV was 99.6% (95% CI: 99.4-99.8%; Table 2). The performance was maintained in the presence or absence of a parenchymal abnormality with the model achieving sensitivity and specificity of at least 80% for both subgroups (Table 3). The performance was broadly consistent across sex, age, ethnicity, race and manufacturer with all subgroups with at least 5 cases having sensitivity and specificity of at least 80% (Supplemental Table 1).

Thick series

The model successfully performed inference on 770 (94.4%) thick series. This cohort for analysis included 356 (46.2%) women and 414 (53.8%) men; mean (SD) age was 66.5 (± 17.3) years; there were 596 (77.4%) positive cases and 174 (22.6%) negative cases (Online Supplemental Data). The AI model identified mass effect with AUC of 0.983 (95% CI: 0.975-0.991; Figure 2B and Table 1) and, at an operating point of 0.160195, the sensitivity was 95.3% (95% CI: 93.5-96.8%) and the specificity was 93.1% (95% CI: 89.1-96.6%). At an assumed prevalence of 0.10, the PPV was 60.6% (95% CI: 49.4-75.0%) and NPV was 99.4% (95% CI: 99.2-99.6%; Table 2). The performance was maintained in the presence or absence of a parenchymal abnormality with the model achieving sensitivity and specificity of at least 80% for both subgroups (Table 3). The performance was broadly consistent across sex, age, ethnicity, race and manufacturer with all subgroups with at least 5 cases having sensitivity and specificity of at least 80% (Supplemental Table 2).

Vasogenic edema

A cohort of 310 CT head cases were selected for the vasogenic edema cohort. This cohort resulted in 174 thin series and 309 thick series for which the model could be evaluated (Figure 3).

Thin series

The model successfully performed inference on 169 (97.1%) thin series. This cohort for analysis included 77 (45.6%) women and 92 (54.4%) men; mean (SD) age was 65.6 (± 19.7) years; there were 61 (36.1%) positive cases and 108 (63.9%) negative cases (Online Supplemental Data). The AI model identified vasogenic edema with AUC of 0.980 (95% CI: 0.961-0.993; Figure 4A and Table 1) and, at an operating point of 0.145352, the sensitivity was 90.2% (95% CI: 82.0-96.7%) and the specificity was 93.5% (95% CI: 88.9-97.2%). At an assumed prevalence of 0.10, the PPV was 60.7% (95% CI: 46.4-82.4%) and NPV was 98.8% (95% CI: 97.9-99.7%; Table 2). The performance was maintained in the presence or absence of a parenchymal abnormality with the model achieving sensitivity and specificity of at least 80% for both subgroups (Table 3). The performance was broadly consistent across sex, age, ethnicity, race and manufacturer with all subgroups with at least 8 cases having sensitivity and specificity of at least 80% (Supplemental Table 3).

Thick series

The model successfully performed inference on 301 (97.4%) thick series. This cohort for analysis included 148 (49.2%) women and 153 (50.8%) men; mean (SD) age was 64.6 (± 19.9) years; there were 100 (33.2%) positive cases and 201 (66.8%) negative cases (Online Supplemental Data). The AI model identified vasogenic edema with AUC of 0.988 (95% CI: 0.977-0.995; Figure 4B and Table 2) and, at an operating point of 0.145352, the sensitivity was 90.0% (95% CI: 84.0-96.0%) and the specificity was 95.5% (95% CI: 92.5-98.0%). At an assumed prevalence of 0.10, the PPV was 69.1% (95% CI: 57.1-84.5%) and NPV was 98.9% (95% CI: 98.1-99.5%; Table 2). The performance was maintained in the presence or absence of a parenchymal abnormality with the model achieving sensitivity and specificity of at least 80% for both subgroups (Table 3). The performance was broadly consistent across sex, age, ethnicity, race and manufacturer with all subgroups with at least 2 cases having sensitivity and specificity of at least 80% (Supplemental Table 4).

DISCUSSION

This retrospective standalone model performance study assessed the ability of an AI device to identify mass effect and vasogenic edema on head CT. For mass effect, the AI device achieved AUC 0.987 on thin series and AUC 0.983 on thick series. For vasogenic edema, it achieved AUC 0.980 on thin series and AUC 0.988 on thick series. These AUCs, as well as the lower bounds of their 95% CIs, are in excess of the benchmark AUC of 0.95 that the FDA uses for CADt devices cleared under the QFM product code.[15] Both devices also had operating points that corresponded to sensitivity and specificity of at least 89.8%. These sensitivities and specificities, as well as the lower bounds of their 95% CIs, are in excess of the benchmark sensitivity of 80% and specificity of 80% that the FDA commonly uses for CADt devices cleared under the QAS product code (which was the product code through which these two findings were cleared).[10, 11, 16]

The proposed benefit of CADt devices is that they are "intended to aid in prioritization and triage of radiological medical images"[17] such that clinicians will be aware of abnormal studies sooner and can commence subsequent management steps. There are two other CADt devices that the FDA has approved for the identification of mass effect. The NinesAI device detects mass effect with sensitivity 96.4% and specificity 91.1%.[7] The qER device detects mass effect with sensitivity 96.39% and specificity 96.00%, and midline shift with sensitivity 97.34% and specificity 95.36%.[6] The current results are consistent, while noting that the cohorts for the assessment of each algorithm are different and therefore prevent direct comparison.

One of the ongoing challenges with CADt devices cleared by the FDA is that the regulation states "device does not mark, highlight, or direct users' attention to a specific location in the original image".[17] This assessment was therefore based on only the binary identification of mass effect or vasogenic edema, and did not incorporate a localization or segmentation analysis. As we have described previously, a localization output including a segmentation or heat-map could assist with explainability by demonstrating what the model has identified especially when a user suspects the algorithm has falsely identified a finding (i.e., a false positive case).[12, 18, 19] The growing number of head CT findings that can be identified by CADt devices paves the way for application of AI in radiology to use cases requiring broader identification of findings such as report writing.

This device demonstrated robust performance across sex, age, ethnicity, race and manufacturer subgroups. It achieved a sensitivity and specificity of at least 80% whenever there were at least 8 cases within a subgroup. This performance suggests that the device is generalizable for different patient demographics and technical parameters. The device will, however, encounter new scenarios when being used in the clinical environment and its ongoing performance should continue to be monitored. The device reassuringly also appeared to differentiate between an underlying parenchymal abnormality and mass effect or vasogenic edema, as suggested by its ability to maintain a specificity greater than 80% even when a parenchymal abnormality was present.

As we have described for similar standalone model performance assessments, a key limitation of this study is that it is retrospective and outside of the clinical workflow.[12, 18] It therefore establishes the accuracy of the model in identifying mass effect and vasogenic edema but does not assess its impact on the clinical workflow including for benefit on patient outcomes. We view this initial step as a prerequisite to ensure that the device has the potential to provide clinical benefit. Further evaluation will be required to prove such a benefit.

CONCLUSIONS

This standalone model performance assessment investigated the ability of an AI device to identify mass effect and vasogenic edema on head CT. It demonstrated performance that exceeded the FDA benchmarks for CADt devices. Its use could lead to improved care and outcomes for patients with these findings.

ACKNOWLEDGMENTS

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

This study was funded by Annalise-AI. Annalise-AI were involved in the design and conduct of the study; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. Annalise-AI were not involved in the collection, management, analysis, and interpretation of the data.

REFERENCES

1. Frey C, Hazenfield JM. Essentials of Head Trauma Imaging. Semin Ultrasound CT MR 2018;39:469-480

2. Rincon S, Gupta R, Ptak T. Imaging of head trauma. Handb Clin Neurol 2016;135:447-477

3. Fink KR, Benjert JL. Imaging of Nontraumatic Neuroradiology Emergencies. Radiol Clin North Am 2015;53:871-890, x

4. Matsoukas S, Stein LK, Fifi J. Artificial Intelligence-Assisted Software Significantly Decreases All Workflow Metrics for Large Vessel Occlusion Transfer Patients, within a Large Spoke and Hub System. Cerebrovasc Dis Extra 2023;13:41-46

5. Morey JR, Zhang X, Yaeger KA, et al. Real-World Experience with Artificial Intelligence-Based Triage in Transferred Large Vessel Occlusion Stroke Patients. Cerebrovasc Dis 2021;50:450-455

6. US Food and Drug Administration. K200921 (qER). 2020

7. US Food and Drug Administration. K193351 (NinesAI). 2020

8. U.S. Food & Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. 2023

9. Buchlak Q, Tang C, Seah J, et al. Effects of a comprehensive brain computed tomography deep-learning model on radiologist detection accuracy: a multireader, multicase study. 2022

10. US Food and Drug Administration. K231767 (Annalise Enterprise CTB Triage Trauma). 2023

11. US Food and Drug Administration. K231384 (Annalise Enterprise CTB Triage Trauma). 2023

12. Hillis JM, Bizzo BC, Newbury-Chaet I, et al. Evaluation of an artificial intelligence model for identification of intracranial hemorrhage subtypes on computed tomography of the head. Stroke: Vascular and Interventional Neurology (In press), 2024

13. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019;95:103208

14. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42:377-381

15. U.S. Food & Drug Administration. Product Classification (QFM Product Code). 2024

16. U.S. Food & Drug Administration. Product Classification (QAS Product Code). 2024

17. Department of Health and Human Services. 21CFR 892.2080 Radiological computer aided triage and notification software. 2020

18. Hillis JM, Bizzo BC, Mercaldo S, et al. Evaluation of an Artificial Intelligence Model for Detection of Pneumothorax and Tension Pneumothorax in Chest Radiographs. JAMA Netw Open 2022;5:e2247172

19. Hillis JM, Visser JJ, Cliff ERS, et al. The lucent yet opaque challenge of regulating artificial intelligence in radiology. NPJ Digit Med 2024;7:69

TABLES

Table 1: Model performance summary for identifying mass effect at operating point 0.221484 on thin series and at operating point 0.160195 on thick series, and for identifying vasogenic edema at operating point 0.145352 on thin series and at operating point 0.145352 on thick series.

	Mass effect		Vasogenic edema	
Metric	Thin series	Thick series	Thin series	Thick series
Positive N	495	596	61	100
Negative N	137	174	108	201
AUC	0.987	0.983	0.980	0.988
(95% CI)	(0.979-0.993)	(0.975-0.991)	(0.961-0.993)	(0.977-0.995)
Sensitivity	96.6	95.3	90.2	90.0
(95% CI)	(94.9-98.2)	(93.5-96.8)	(82.0-96.7)	(84.0-96.0)
Specificity	89.8	93.1	93.5	95.5
(95% CI)	(84.7-94.2)	(89.1-96.6)	(88.9-97.2)	(92.5-98.0)

Table 2: PPVs and NPVs at different levels of assumed prevalence for identifying mass effect and vasogenic edema. These measurements are based off the same operating points used to determine sensitivity and specificity in Table 2.

	Thin series		Thick series	
Assumed prevalence	PPV (95% CI)	NPV (95% CI)	PPV (95% CI)	NPV (95% CI)
Mass effect				
0.05	33.2 (24.7-48.7)	99.8 (99.7-99.9)	42.1 (31.6-58.7)	99.7 (99.6-99.8)
0.10	51.2 (40.9-66.7)	99.6 (99.4-99.8)	60.6 (49.4-75.0)	99.4 (99.2-99.6)
0.15	62.5 (52.4-76.1)	99.3 (99.0-99.6)	70.9 (60.8-82.7)	99.1 (98.8-99.4)
0.20	70.3 (60.9-81.8)	99.1 (98.6-99.5)	77.6 (68.7-87.1)	98.8 (98.3-99.2)
Vasogenic edema				
0.05	42.3 (29.1-68.9)	99.4 (99.0-99.9)	51.4 (38.7-72.1)	99.5 (99.1-99.7)
0.10	60.7 (46.4-82.4)	98.8 (97.9-99.7)	69.1 (57.1-84.5)	98.9 (98.1-99.5)
0.15	71.1 (57.9-88.1)	98.2 (96.7-99.6)	78.0 (67.9-89.6)	98.2 (97.0-99.2)
0.20	77.7 (66.1-91.3)	97.4 (95.3-99.4)	83.4 (75.0-92.5)	97.4 (95.9-98.8)

Table 3: Sensitivity and specificity for subgroups based on the presence or absence of a parenchymal abnormality. These measurements are based off the same operating points used to determine sensitivity and specificity in Table 2.

	Positive N	Negative N	Sensitivity (95% CI)	Specificity (95% CI)
Mass effect – thin series				
Parenchymal abnormality present	319	32	98.4 (96.9-99.7)	81.2 (65.6-93.8)
Parenchymal abnormality absent	176	105	93.2 (89.2-96.6)	92.4 (86.7-97.1)
Mass effect – thick series				
Parenchymal abnormality present	391	39	98.2 (96.9-99.5)	84.6 (71.8-94.9)
Parenchymal abnormality absent	205	135	89.8 (85.4-93.7)	95.6 (91.9-98.5)
Vasogenic edema – thin series				
Parenchymal abnormality present	57	28	89.5 (80.7-96.5)	82.1 (67.9-96.4)
Parenchymal abnormality absent	4	80	100.0 (100.0-100.0)	97.5 (93.8-100.0)
Vasogenic edema – thick series				
Parenchymal abnormality present	94	37	89.4 (83.0-94.7)	83.8 (70.3-94.6)
Parenchymal abnormality absent	6	164	100.0 (100.0-100.0)	98.2 (95.7-100.0)

FIGURES

Figure 1: Cohort selection diagram for mass effect

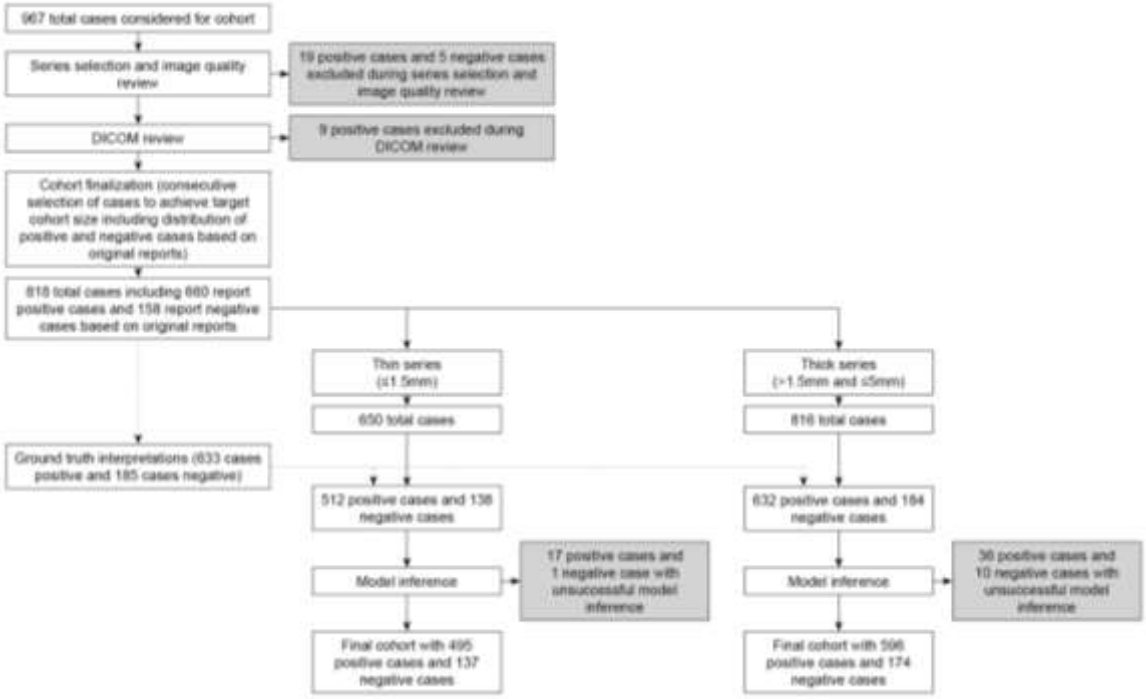


Figure 2: Performance for mass effect. A and B, Receiver operating characteristic curves for the thin series (A) and thick series (B). The shaded region reflects the bootstrapped 95% CIs. The selected point on each graph reflects the performance at the operating points described in the text. C, D and E, Example images for true positive case with parenchymal abnormality (C), true positive case without parenchymal abnormality (D) and true negative case with parenchymal abnormality (E). The model classification score output is provided for each case.

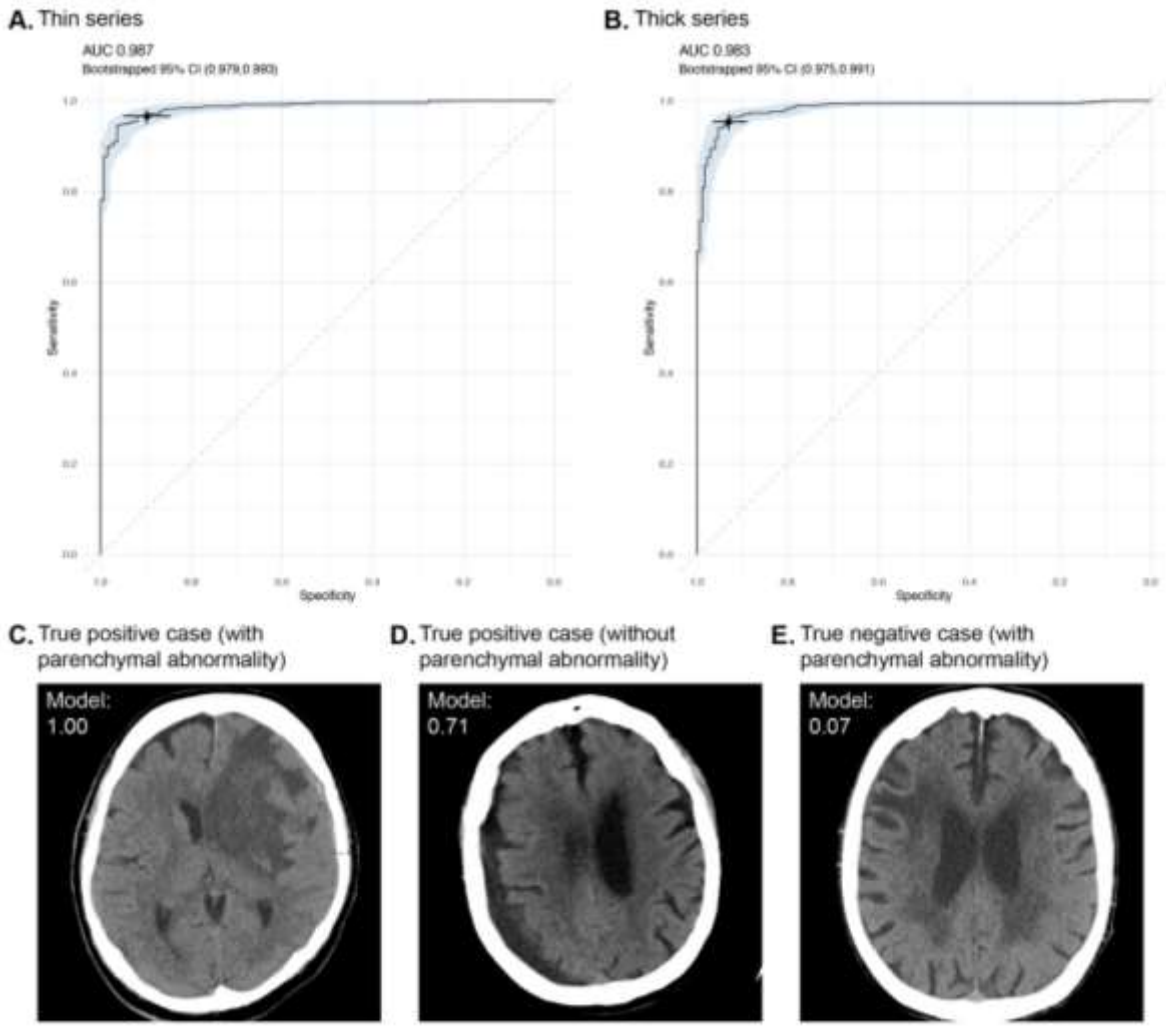


Figure 3: Cohort selection diagram for vasogenic edema

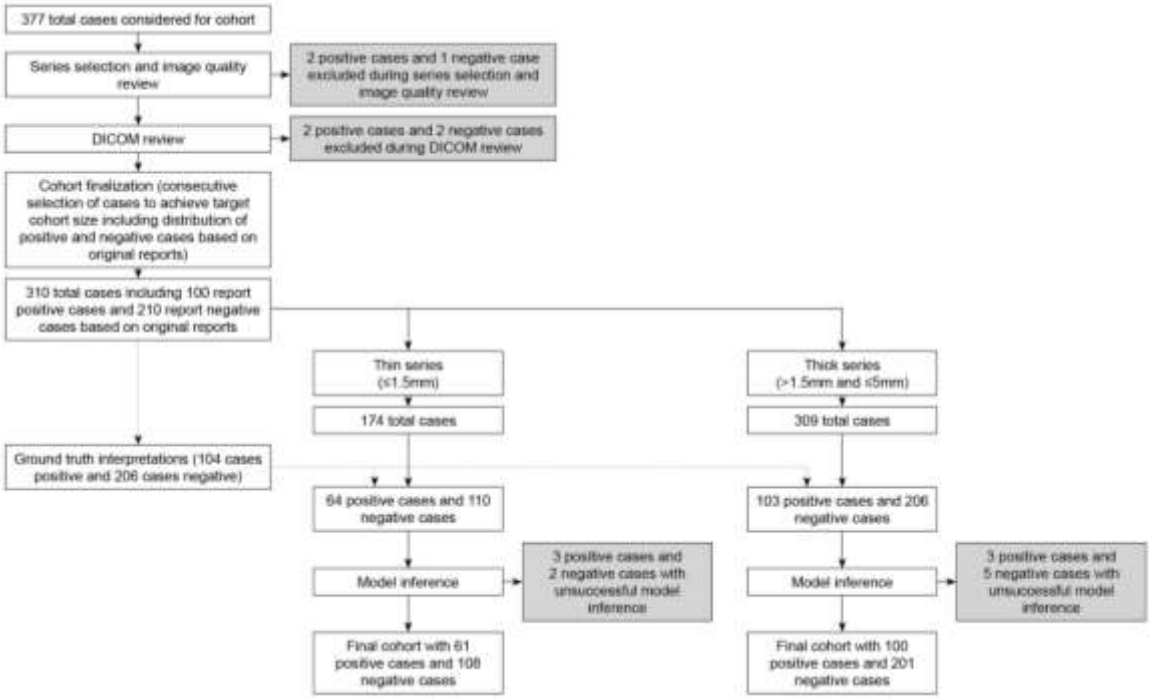
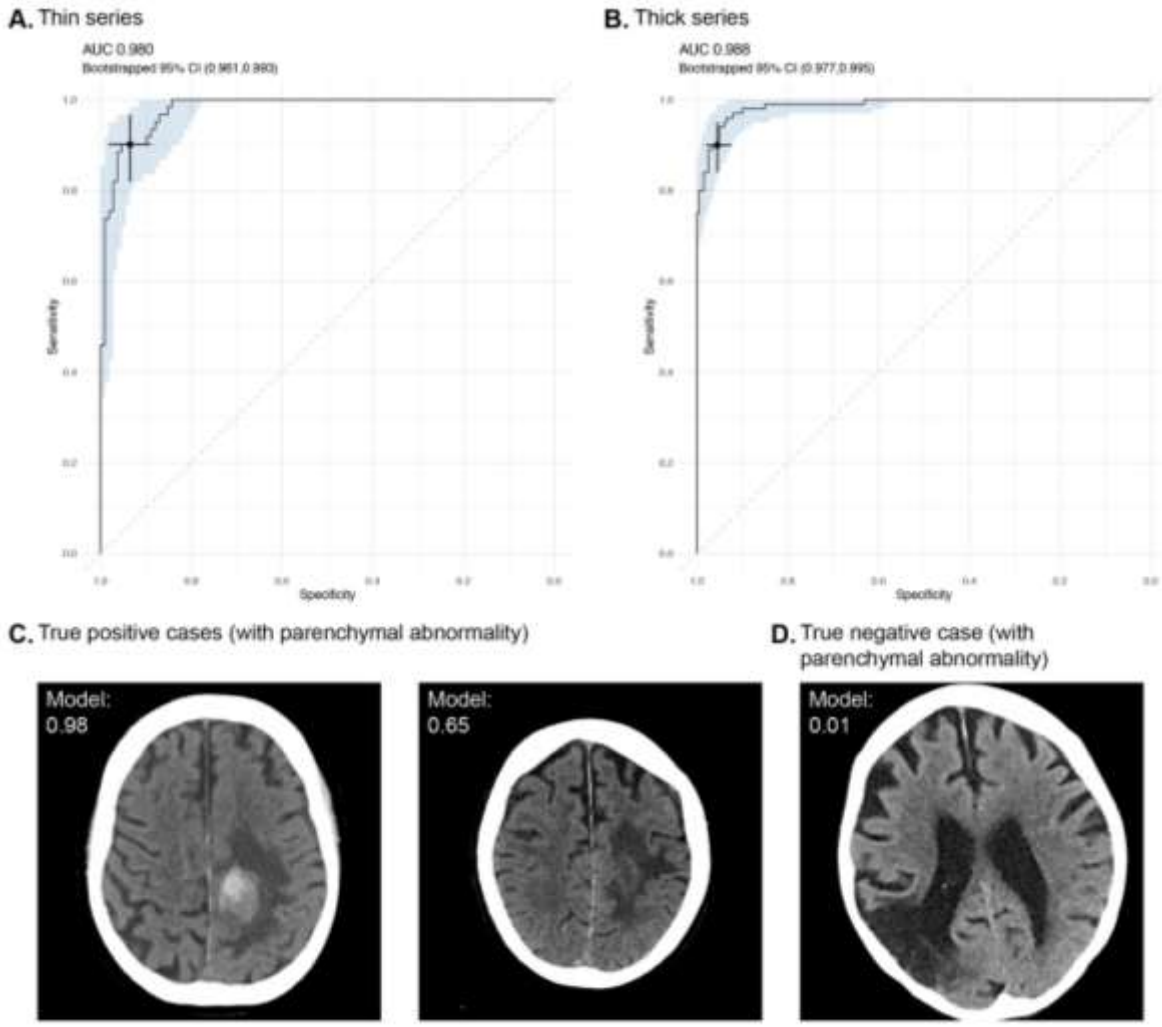


Figure 4: Performance for vasogenic edema. A and B, Receiver operating characteristic curves for the thin series (A) and thick series (B). The shaded region reflects the bootstrapped 95% CIs. The selected point on each graph reflects the performance at the operating points described in the text. C, D and E, Example images for true positive cases with parenchymal abnormality (C and D), and true negative case with parenchymal abnormality (E). The model classification score output is provided for each case.



SUPPLEMENTAL FILES

Online Supplemental Data: Demographic and technical breakdown of CT head cases for each finding.

	Mass effect				Vasogenic edema			
	Thin series		Thick series		Thin series		Thick series	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Total	495	137	596	174	61	108	100	201
Sex								
Female	238 (48.1%)	68 (49.6%)	274 (46.0%)	82 (47.1%)	28 (45.9%)	49 (45.4%)	45 (45.0%)	103 (51.2%)
Male	257 (51.9%)	69 (50.4%)	322 (54.0%)	92 (52.9%)	33 (54.1%)	59 (54.6%)	55 (55.0%)	98 (48.8%)
Age								
≤65 years	199 (40.2%)	52 (38.0%)	246 (41.3%)	74 (42.5%)	26 (42.6%)	52 (48.1%)	41 (41.0%)	94 (46.8%)
>65 years	296 (59.8%)	85 (62.0%)	350 (58.7%)	100 (57.5%)	35 (57.4%)	56 (51.9%)	59 (59.0%)	107 (53.2%)
Mean ± SD (years)	67.2 ± 17.0	67.1 ± 19.4	66.8 ± 16.5	65.3 ± 19.7	67.4 ± 16.9	64.5 ± 21.1	67.8 ± 15.2	62.9 ± 21.7
Ethnicity								
Hispanic	34 (6.9%)	16 (11.7%)	36 (6.0%)	15 (8.6%)	2 (3.3%)	8 (7.4%)	6 (6.0%)	18 (9.0%)
Not Hispanic	425 (85.9%)	113 (82.5%)	525 (88.1%)	151 (86.8%)	58 (95.1%)	95 (88.0%)	92 (92.0%)	176 (87.6%)
Prefer not to say / Decline	1 (0.2%)	2 (1.5%)	2 (0.3%)	2 (1.1%)	0 (0.0%)	0 (0.0%)	1 (1.0%)	0 (0.0%)
Unavailable	35 (7.1%)	6 (4.4%)	33 (5.5%)	6 (3.4%)	1 (1.6%)	5 (4.6%)	1 (1.0%)	7 (3.5%)
Race								
Asian	19 (3.8%)	3 (2.2%)	26 (4.4%)	3 (1.7%)	7 (11.5%)	4 (3.7%)	10 (10.0%)	4 (2.0%)
Black or African American	35 (7.1%)	12 (8.8%)	35 (5.9%)	12 (6.9%)	4 (6.6%)	6 (5.6%)	5 (5.0%)	13 (6.5%)
White	395 (79.8%)	106 (77.4%)	484 (81.2%)	144 (82.8%)	47 (77.0%)	83 (76.9%)	80 (80.0%)	159 (79.1%)
Other	25 (5.1%)	8 (5.8%)	28 (4.7%)	7 (4.0%)	1 (1.6%)	8 (7.4%)	2 (2.0%)	14 (7.0%)
Two or more	3 (0.6%)	2 (1.5%)	5 (0.8%)	2 (1.1%)	0 (0.0%)	1 (0.9%)	0 (0.0%)	4 (2.0%)
Declined	1 (0.2%)	2 (1.5%)	1 (0.2%)	2 (1.1%)	0 (0.0%)	1 (0.9%)	0 (0.0%)	1 (0.5%)
Unavailable	17 (3.4%)	4 (2.9%)	17 (2.9%)	4 (2.3%)	2 (3.3%)	5 (4.6%)	3 (3.0%)	6 (3.0%)
Manufacturer								
GE Healthcare	174 (35.2%)	44 (32.1%)	175 (29.4%)	41 (23.6%)	7 (11.5%)	12 (11.1%)	28 (28.0%)	56 (27.9%)
NeuroLogica	2 (0.4%)	0 (0.0%)	3 (0.5%)	0 (0.0%)	1 (1.6%)	4 (3.7%)	1 (1.0%)	5 (2.5%)
Philips	6 (1.2%)	0 (0.0%)	6 (1.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Siemens	309 (62.4%)	93 (67.9%)	289 (48.5%)	91 (52.3%)	52 (85.2%)	92 (85.2%)	53 (53.0%)	96 (47.8%)
Toshiba	4 (0.8%)	0 (0.0%)	123 (20.6%)	42 (24.1%)	1 (1.6%)	0 (0.0%)	18 (18.0%)	44 (21.9%)

Supplemental Table 1: Demographic and technical subgroup performance for identifying mass effect at operating point 0.221484 on thin series (≤1.5mm).

	Positive N	Negative N	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Overall	495	137	0.987 (0.979-0.993)	96.6 (94.9-98.2)	89.8 (84.7-94.2)
Sex					
Female	238	68	0.996 (0.990-0.999)	97.5 (95.4-99.2)	91.2 (83.8-97.1)
Male	257	69	0.980 (0.966-0.991)	95.7 (93.0-98.1)	88.4 (79.7-94.2)
Age					
≤65 years	199	52	0.992 (0.981-0.999)	97.5 (95.0-99.5)	92.3 (84.6-98.1)
>65 years	296	85	0.984 (0.973-0.993)	95.9 (93.6-98.0)	88.2 (81.2-94.1)
Ethnicity					
Hispanic	34	16	0.972 (0.912-1.000)	91.2 (82.4-100.0)	93.8 (81.2-100.0)
Not Hispanic	425	113	0.988 (0.980-0.994)	96.7 (94.8-98.1)	90.3 (85.0-94.7)
Prefer not to say / Decline	1	2	1.000 (1.000-1.000)	100.0 (100.0-100.0)	50.0 (0.0-100.0)
Unavailable	35	6	1.000 (1.000-1.000)	100.0 (100.0-100.0)	83.3 (50.0-100.0)
Race					
Asian	19	3	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Black or African American	35	12	0.988 (0.957-1.000)	94.3 (85.7-100.0)	91.7 (75.0-100.0)
White	395	106	0.986 (0.976-0.993)	96.2 (94.2-98.0)	88.7 (82.1-94.3)
Other	25	8	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Two or more	3	2	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Declined	1	2	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Unavailable	17	4	1.000 (1.000-1.000)	100.0 (100.0-100.0)	75.0 (25.0-100.0)
Manufacturer					
GE Healthcare	174	44	0.984 (0.967-0.995)	96.0 (93.1-98.3)	86.4 (75.0-95.5)
NeuroLogica	2	0	-	100.0 (100.0-100.0)	-
Philips	6	0	-	83.3 (42.9-100.0)	-
Siemens	309	93	0.991 (0.982-0.997)	97.4 (95.5-99.0)	91.4 (84.9-96.8)
Toshiba	4	0	-	75.0 (0.0-100.0)	-

Supplemental Table 2: Demographic and technical subgroup performance for identifying mass effect at operating point 0.160195 on thick series (>1.5mm and ≤5mm).

	Positive N	Negative N	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Overall	596	174	0.983 (0.975-0.991)	95.3 (93.5-96.8)	93.1 (89.1-96.6)
Sex					
Female	274	82	0.985 (0.973-0.993)	94.5 (91.6-97.1)	90.2 (82.9-96.3)
Male	322	92	0.982 (0.968-0.993)	96.0 (93.8-98.1)	95.7 (90.2-98.9)
Age					
≤65 years	246	74	0.990 (0.977-0.998)	95.5 (92.7-98.0)	97.3 (93.2-100.0)
>65 years	350	100	0.978 (0.965-0.988)	95.1 (92.9-97.1)	90.0 (84.0-95.0)
Ethnicity					
Hispanic	36	15	0.978 (0.922-1.000)	91.7 (80.6-100.0)	100.0 (100.0-100.0)
Not Hispanic	525	151	0.985 (0.976-0.992)	96.0 (94.3-97.5)	92.7 (88.1-96.7)
Prefer not to say / Decline	2	2	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Unavailable	33	6	0.985 (0.919-1.000)	87.9 (75.8-97.0)	83.3 (50.0-100.0)
Race					
Asian	26	3	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Black or African American	35	12	1.000 (1.000-1.000)	97.1 (91.4-100.0)	100.0 (100.0-100.0)
White	484	144	0.980 (0.969-0.989)	94.8 (92.8-96.7)	92.4 (87.5-96.5)
Other	28	7	1.000 (1.000-1.000)	96.4 (89.3-100.0)	100.0 (100.0-100.0)
Two or more	5	2	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Declined	1	2	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Unavailable	17	4	1.000 (0.912-1.000)	94.1 (82.4-100.0)	75.0 (25.0-100.0)
Manufacturer					
GE Healthcare	175	41	0.976 (0.954-0.993)	94.9 (91.4-97.7)	90.2 (80.5-97.6)
NeuroLogica	3	0	-	100.0 (100.0-100.0)	-
Philips	6	0	-	83.3 (42.9-100.0)	-
Siemens	289	91	0.989 (0.977- 0.996)	95.5 (92.7-97.9)	94.5 (89.0-98.9)
Toshiba	123	42	0.990 (0.977- 0.998)	95.9 (91.9-99.2)	92.9 (85.7-100.0)

Supplemental Table 3: Demographic and technical subgroup performance for identifying vasogenic edema at operating point 0.145352 on thin series (≤1.5mm).

	Positive N	Negative N	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Overall	61	108	0.980 (0.961-0.993)	90.2 (82.0-96.7)	93.5 (88.9-97.2)
Sex					
Female	28	49	0.990 (0.968-1.000)	92.9 (82.1-100.0)	93.9 (85.7-100.0)
Male	33	59	0.976 (0.942-0.997)	87.9 (75.8-97.0)	93.2 (86.4-98.3)
Age					
≤65 years	26	52	0.961 (0.915-0.990)	88.5 (73.1-100.0)	90.4 (82.7-98.1)
>65 years	35	56	0.993 (0.979-1.000)	91.4 (80.0-100.0)	96.4 (91.1-100.0)
Ethnicity					
Hispanic	2	8	1.000 (1.000-1.000)	100.0 (100.0-100.0)	87.5 (62.5-100.0)
Not Hispanic	58	95	0.981 (0.961-0.994)	91.4 (84.5-98.3)	93.7 (88.4-97.9)
Prefer not to say / Decline	0	0	-	-	-
Unavailable	1	5	0.800 (0.400-1.000)	0.0 (0.0-0.0)	100.0 (100.0-100.0)
Race					
Asian	7	4	0.929 (0.714-1.000)	71.4 (42.9-100.0)	75.0 (25.0-100.0)
Black or African American	4	6	1.000 (1.000-1.000)	75.0 (25.0-100.0)	100.0 (100.0-100.0)
White	47	83	0.981 (0.958-0.995)	93.6 (85.1-100.0)	92.8 (86.7-97.6)
Other	1	8	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Two or more	0	1	-	-	100.0 (100.0-100.0)
Declined	0	1	-	-	100.0 (100.0-100.0)
Unavailable	2	5	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Manufacturer					
GE Healthcare	7	12	0.964 (0.821-1.000)	71.4 (42.9-100.0)	91.7 (75.0-100.0)
NeuroLogica	1	4	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Siemens	52	92	0.985 (0.966-0.996)	92.3 (84.6-98.1)	93.5 (88.0-97.8)
Toshiba	1	0	-	100.0 (100.0-100.0)	-

Supplemental Table 4: Demographic and technical subgroup performance for vasogenic edema at operating point 0.145352 on thick series (>1.5mm and ≤5mm).

	Positive N	Negative N	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Overall	100	201	0.988 (0.977-0.995)	90.0 (84.0-96.0)	95.5 (92.5-98.0)
Sex					
Female	45	103	0.996 (0.989-1.000)	95.6 (88.9-100.0)	95.1 (90.3-99.0)
Male	55	98	0.981 (0.957-0.995)	85.5 (76.4-94.5)	95.9 (91.8-99.0)
Age					
≤65 years	41	94	0.988 (0.972-0.996)	87.8 (78.0-97.6)	94.7 (90.4-98.9)
>65 years	59	107	0.989 (0.971-0.999)	91.5 (83.1-98.3)	96.3 (92.5-99.1)
Ethnicity					
Hispanic	6	18	1.000 (1.000-1.000)	100.0 (100.0-100.0)	94.4 (83.3-100.0)
Not Hispanic	92	176	0.987 (0.974-0.995)	90.2 (83.7-95.7)	95.5 (92.0-98.3)
Prefer not to say / Decline	1	0	-	100.0 (100.0-100.0)	-
Unavailable	1	7	0.857 (0.571-1.000)	0.0 (0.0-0.0)	100.0 (100.0-100.0)
Race					
Asian	10	4	0.900 (0.700-1.000)	80.0 (50.0-100.0)	100.0 (100.0-100.0)
Black or African American	5	13	1.000 (1.000-1.000)	80.0 (40.0-100.0)	100.0 (100.0-100.0)
White	80	159	0.991 (0.982-0.997)	91.2 (83.8-97.5)	94.3 (90.6-97.5)
Other	2	14	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Two or more	0	4	-	-	100.0 (100.0-100.0)
Declined	0	1	-	-	100.0 (100.0-100.0)
Unavailable	3	6	1.000 (1.000-1.000)	100.0 (100.0-100.0)	100.0 (100.0-100.0)
Manufacturer					
GE Healthcare	28	56	0.985 (0.950-1.000)	89.3 (78.6-100.0)	98.2 (94.6-100.0)
NeuroLogica	1	5	1.000 (1.000-1.000)	100.0 (100.0-100.0)	80.0 (40.0-100.0)
Siemens	53	96	0.986 (0.970-0.995)	90.6 (81.1-98.1)	92.7 (87.5-97.9)
Toshiba	18	44	1.000 (1.000-1.000)	88.9 (72.2-100.0)	100.0 (100.0-100.0)