medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Title Page

Title: Evaluation of an artificial intelligence model for identification of intracranial hemorrhage subtypes on computed tomography of the head

Authors:
James M Hillis, MBBS DPhil*[1,2,3]
Bernardo C Bizzo, MD PhD*[1,3,4]
Isabella Newbury-Chaet, BSc[1]
Sarah F Mercaldo, PhD[1,3,4]
John K Chin, MD[1]
Ankita Ghatak, MSc[1]
Madeleine A Halle, BSc[1]
Eric L'Italien[1]
Ashley L MacDonald, BSc[1]
Alex S Schultz[1]
Karen Buch, MD[3,4]
John Conklin, MD[3,4]
Stuart Pomerantz, MD[1,3,4]
Sandra Rincon, MD[3,4]
Keith J Dreyer, DO PhD[1,3,4]
William A Mehan, MD MBA[3,4]

* 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

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

Date of revision: September 7 2023

Word count: 3159 (not including title, abstract, acknowledgment, references, tables, and figure legends).

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Key Points

Question: Does a commercial artificial intelligence model accurately identify intracranial hemorrhage subtypes on computed tomography (CT) of the head?

Findings: This retrospective study used non-contrast CT studies to compare artificial intelligence model outputs to consensus neuroradiologist interpretations. The model was provided with either thin (≤1.5mm) or thick (>1.5 and ≤5mm) series. The model detected each of acute subdural/epidural hematoma, acute subarachnoid hemorrhage, intra-axial hemorrhage and intraventricular hemorrhage with sensitivity and specificity greater than 80%.

Meaning: This artificial intelligence model could assist radiologists through its accurate detection of intracranial hemorrhage subtypes.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Importance: Intracranial hemorrhage is a critical finding on computed tomography (CT) of the head.

Objective: This study compared the accuracy of an AI model (Annalise Enterprise CTB) to consensus neuroradiologist interpretations in detecting four hemorrhage subtypes: acute subdural/epidural hematoma, acute subarachnoid hemorrhage, intra-axial hemorrhage and intraventricular hemorrhage.

Design: A retrospective standalone performance assessment was conducted on datasets of non-contrast CT head cases acquired between 2016 and 2022 for each hemorrhage subtype.

Setting: The cases were obtained from five hospitals in the United States.

Participants: The cases were obtained from patients aged 18 years or older. The positive cases were selected 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.

Interventions: 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 hemorrhage subtype. The neuroradiologists were provided with the

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

entire CT study. The AI model separately received thin (≤1.5mm) and/or thick (>1.5 and ≤5mm) axial series.

Results: The four cohorts included 571 cases for acute subdural/epidural hematoma, 310 cases for acute subarachnoid hemorrhage, 926 cases for intra-axial hemorrhage and 199 cases for intraventricular hemorrhage. The AI model identified acute subdural/epidural hematoma with area under the curve (AUC) 0.973 (95% confidence interval (CI), 0.958-0.984) on thin series and 0.942 (95% CI, 0.921-0.959) on thick series; acute subarachnoid hemorrhage with AUC 0.993 (95% CI, 0.984-0.998) on thin series and 0.966 (95% CI, 0.945-0.983) on thick series; intra-axial hemorrhage with AUC 0.969 (95% CI, 0.956-0.980) on thin series and 0.966 (95% CI, 0.953-0.976) on thick series; and intraventricular hemorrhage with AUC 0.987 (95% CI, 0.969-0.997) on thin series and 0.983 (95% CI, 0.968-0.994) on thick series. Each finding had at least one operating point with sensitivity and specificity greater than 80%.

Conclusions and Relevance: The assessed AI model accurately identified intracranial hemorrhage subtypes in this CT dataset. Its use could assist the clinical workflow especially through enabling triage of abnormal CTs.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Intracranial hemorrhage is a critical finding on computed tomography (CT) of the head and requires emergent medical attention. There are four key subtypes that occur in different regions of the brain: epidural/subdural hematoma, subarachnoid hemorrhage, intra-axial hemorrhage and intraventricular hemorrhage.[1,2] The faster detection of intracranial hemorrhage on head CT can enable sooner management and intervention including surgery.[3-5]

The use of artificial intelligence (AI) to identify intracranial hemorrhage on head CT has been proposed to assist clinical care, especially by triaging suspected cases for sooner interpretation by a radiologist. There have been at least thirteen computer assisted triage devices (CADt) cleared by the US Food and Drug Administration (FDA) for intracranial hemorrhage detection.[6] While these devices report subgroup performance for identification of different intracranial hemorrhage subtypes as an "intracranial hemorrhage",[7,8] the AI model assessed in the current study is the first FDA-cleared CADt device to output the identified intracranial hemorrhage subtype(s).[9] It specifically identifies acute subdural/epidural hematoma, acute subarachnoid hemorrhage, intra-axial hemorrhage and intraventricular hemorrhage.

This study calculated the performance of this AI model by comparing its outputs to consensus neuroradiologist interpretations on a cohort of head CT cases for each of the four findings. The AI 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

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

performance was also calculated for cases belonging to demographic and technical subgroups to determine the generalizability of the device.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Methods

Study design

This retrospective standalone model performance study was conducted using radiology cases from five hospitals within the Mass General Brigham (MGB) network between 2016 and 2022. It was approved by the MGB 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 (HIPAA). This report followed the Standards for Reporting Diagnostic Accuracy (STARD 2015) reporting guideline.

Case selection

The cohorts for each of the four intracranial hemorrhage subtypes were selected in a consecutive manner based on the original radiology reports. 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 cohort size for each of the positive and negative cases was based on powering calculations as described in the statistical analysis section below. When the cohort had equal numbers of positive and negative cases, the next negative cases were taken after each positive case; when there were unequal numbers of positive and negative cases, 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.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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 across multiple cohorts (e.g., both intra-axial hemorrhage and intraventricular hemorrhage).

All cases were deidentified and underwent an image quality review by an American Board of Radiology (ABR)-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 MGB.[10,11]

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

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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". 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, a "2+1" strategy was used: the first two neuroradiologists interpreted every case and a third neuroradiologist then interpreted cases with discrepant interpretations.

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.[12] 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.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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 Digital Imaging and Communications in Medicine (DICOM)-formatted CT head cases. It outputted a classification score between 0 and 1 for each of acute subdural/epidural hematoma, acute subarachnoid hemorrhage, intra-axial hemorrhage and intraventricular hemorrhage. A binary output for each of these findings could be derived using prespecified operating points.

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 acute subdural/epidural hematoma, acute subarachnoid hemorrhage, intra-axial hemorrhage and intraventricular hemorrhage 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 those operating points that have received US Food and Drug Administration 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).

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The AUCs, sensitivities and specificities were calculated as exploratory analyses for the subgroups of sex, age, race, ethnicity 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" and no data were imputed.

All confidence intervals (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 specificity was >80%.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Results

Acute subdural/epidural hematoma

A cohort of 571 CT head cases were selected for the acute subdural/epidural hematoma cohort. This cohort resulted in 423 thin series and 571 thick series for which the model could be evaluated (Supplementary Table 1).

Thin series

The model successfully performed inference on 409 (96.7%) thin series. This cohort for analysis included 185 (45.2%) women and 224 (54.8%) men; mean (SD) age was 67.0 (19.3) years; there were 308 (75.3%) positive cases and 101 (24.7%) negative cases (Table 1). The AI model identified acute subdural/epidural hematoma with AUC of 0.973 (95% CI: 0.958-0.984; Figure 1A) and, at an operating point of 0.060177, the sensitivity was 91.6% (95% CI: 88.3-94.5%) and the specificity was 87.1% (95% CI: 80.2-93.1%; Supplementary Table 2). These performances were broadly consistent across sex, age, ethnicity, race and manufacturer with all subgroups having sensitivity and specificity of at least 80% except for “unavailable” ethnicity specificity (66.7% for 6 cases) and Asian race specificity (75.0% for 4 cases; Supplementary Table 3).

Thick series

The model successfully performed inference on 539 (94.4%) thick series. This cohort for analysis included 241 (44.7%) women and 298 (55.3%) men; mean (SD) age was 66.8 (19.0) years; there were 401 (74.4%) positive cases and 138 (25.6%) negative cases (Table 1). The AI model identified acute subdural/epidural hematoma with AUC of 0.942 (95% CI: 0.921-0.959; Figure

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1B) and, at an operating point of 0.060177, the sensitivity was 82.5% (95% CI: 78.8-86.0%) and the specificity was 89.9% (84.8-94.2%; Supplementary Table 2). These performances were broadly consistent across sex, age, ethnicity, race and manufacturer with all subgroups having sensitivity and specificity of at least 80% except for age ≤65 years sensitivity (77.7% for 148 cases), Black or African American race sensitivity (78.6% for 14 cases), NeuroLogica manufacturer specificity (0.0% for 1 case; Supplementary Table 3).

Acute subarachnoid hemorrhage

A cohort of 310 CT head cases were selected for the acute subarachnoid hemorrhage cohort. This cohort resulted in 244 thin series and 309 thick series for which the model could be evaluated (Supplementary Table 4).

Thin series

The model successfully performed inference on 238 (97.5%) thin series. This cohort for analysis included 125 (52.5%) women and 113 (47.5%) men; mean (SD) age was 66.5 (19.5) years; there were 149 (62.6%) positive cases and 89 (37.4%) negative cases (Table 1). The AI model identified acute subarachnoid hemorrhage with AUC of 0.993 (95% CI: 0.984-0.998; Figure 2A) and, at an operating point of 0.060162, the sensitivity was 94.0% (95% CI: 89.9-97.3%) and the specificity was 96.6% (95% CI: 92.1-100.0%; Supplementary Table 5). These performances were consistent across sex, age, ethnicity, race and manufacturer with all subgroups having sensitivity and specificity of at least 80% (Supplementary Table 6).

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Thick series

The model successfully performed inference on 292 (94.5%) thick series. This cohort for analysis included 148 (50.7%) women and 144 (49.3%) men; mean (SD) age was 66.6 (18.7) years; there were 184 (63.0%) positive cases and 108 (37.0%) negative cases (Table 1). The AI model identified acute subarachnoid hemorrhage with AUC of 0.966 (95% CI: 0.945-0.983; Figure 2B) and, at an operating point of 0.020255, the sensitivity was 90.8% (95% CI: 86.4-94.6%) and the specificity was 90.7% (95% CI: 85.2-95.4%; Supplementary Table 5). These performances were consistent across sex, age, ethnicity, race and manufacturer with all subgroups having sensitivity and specificity of at least 80% (Supplementary Table 6).

Intra-axial hemorrhage

A cohort of 926 CT head cases were selected for the intra-axial hemorrhage cohort. This cohort resulted in 733 thin series and 925 thick series for which the model could be evaluated (Supplementary Table 7).

Thin series

The model successfully performed inference on 710 (96.9%) thin series. This cohort for analysis included 330 (46.5%) women and 380 (53.5%) men; mean (SD) age was 66.0 (18.4) years; there were 484 (68.2%) positive cases and 226 (31.8%) negative cases (Table 1). The AI model identified intra-axial hemorrhage with AUC of 0.969 (95% CI: 0.956-0.980; Figure 3A) and, at an operating point of 0.322700, the sensitivity was 93.2% (95% CI: 90.9-95.5%) and the specificity was 85.8% (95% CI: 81.0-90.3%; Supplementary Table 8). These performances were broadly

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

consistent across sex, age, ethnicity, race and manufacturer with all subgroups having sensitivity and specificity of at least 80% except for "two or more" race sensitivity (66.7% for 3 cases), "unavailable" ethnicity specificity (54.5% for 11 cases), Asian race specificity (72.7% for 11 cases) and "unavailable" race specificity (75.0% for 8 cases; Supplementary Table 9).

Thick series

The model successfully performed inference on 884 (95.6%) thick series. This cohort for analysis included 411 (46.5%) women and 473 (53.5%) men; mean (SD) age was 66.7 (18.0) years; there were 591 (66.9%) positive cases and 293 (33.1%) negative cases (Table 1). The AI model identified intra-axial hemorrhage with AUC of 0.966 (95% CI: 0.953-0.976; Figure 3A) and, at an operating point of 0.203600, the sensitivity was 93.2% (95% CI: 91.2-95.3%) and the specificity was 85.3% (95% CI: 80.9-89.1%; Supplementary Table 8). These performances were broadly consistent across sex, age, ethnicity, race and manufacturer with all subgroups having sensitivity and specificity of at least 80% except for "two or more" race sensitivity (75.0% for 4 cases), NeuroLogica manufacturer sensitivity (50.0% for 2 cases), "unavailable" ethnicity specificity (46.2% for 13 cases), Asian race specificity (72.7% for 11 cases) and "unavailable" race specificity (50.0% for 10 cases; Supplementary Table 9).

Intraventricular hemorrhage

A cohort of 199 CT head cases were selected for the intraventricular hemorrhage cohort. This cohort resulted in 159 thin series and 199 thick series for which the model could be evaluated (Supplementary Table 10).

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Thin series

The model successfully performed inference on 153 (96.2%) thin series. This cohort for analysis included 77 (50.3%) women and 76 (49.7%) men; mean (SD) age was 66.4 (20.4) years; there were 74 (48.4%) positive cases and 79 (51.6%) negative cases (Table 1). The AI model identified intraventricular hemorrhage with AUC of 0.987 (95% CI: 0.969-0.997; Figure 4A) and, at an operating point of 0.051859, the sensitivity was 90.5% (95% CI: 83.8-95.9%) and the specificity was 97.5% (95% CI: 93.7-100.0%; Supplementary Table 11). These performances were broadly consistent across sex, age, ethnicity, race and manufacturer with all subgroups having sensitivity and specificity of at least 80% except for NeuoLogica sensitivity (0.0% for 1 case; Supplementary Table 12).

Thick series

The model successfully performed inference on 187 (94.0%) thick series. This cohort for analysis included 92 (49.2%) women and 95 (50.8%) men; mean (SD) age was 67.0 (19.6) years; there were 91 (48.7%) positive cases and 96 (51.3%) negative cases (Table 1). The AI model identified intraventricular hemorrhage with AUC of 0.983 (95% CI: 0.968-0.994; Figure 4A) and, at an operating point of 0.051859, the sensitivity was 87.9% (95% CI: 81.3-93.4%) and the specificity was 97.9% (95% CI: 94.8-100.0%; Supplementary Table 11). These performances were consistent across sex, age, ethnicity, race and manufacturer with all subgroups having sensitivity and specificity of at least 80% (Supplementary Table 12).

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Discussion

This retrospective diagnostic study assessed the performance of an AI model in identifying acute subdural/epidural hematoma, acute subarachnoid hemorrhage, intra-axial hemorrhage and intraventricular hemorrhage on head CT. The model achieved sensitivity greater than 80% and specificity greater than 80% for all four findings. These results are consistent with other FDA-cleared intracranial hemorrhage CADt devices.[7,8,13-22] However, this model is the first FDA-cleared CADt device that outputs the subtype of intracranial hemorrhage.[9]

The output of the subtype of intracranial hemorrhage provides users with more explainability of the model outputs. As has been noted on an assessment of a similar model for pneumothorax identification, the FDA regulations for a CADt device only permit devices to output the binary classification performance (the AI model assessed here technically outputs a binary classification for each of the four subtypes).[23] A segmentation output could otherwise further assist with explainability by demonstrating the location of the identified intracranial hemorrhage.

The performance of the model was broadly consistent across sex, age, ethnicity, race and manufacturer subgroups with the vast majority achieving sensitivity and specificity above 80%. For the minority of subgroups that did not achieve a sensitivity or specificity of at least 80%, all except one had small sample sizes of less than 15. The sensitivity for the age ≤65 years subgroup for the thick series for acute subdural/epidural hematoma was 77.7% despite

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

including 148 cases; we note that the overall sensitivity was 82.5% and that these two values had overlapping 95% CIs.

When the model performance was calculated at the same operating point on both thin and thick series, it had better sensitivity on the thin series and better specificity on the thick series. This observation likely relates to the increased z-axis resolution on the thin series, which can facilitate better identification of hemorrhage especially for small hemorrhages that may only be present on a small number of slices. The thinner series, however, also have more noise that could lead to the model incorrectly identifying hemorrhage.[24] A consideration for clinical use may be the use of a lower operating point for thick series to balance out this difference.

As part of the study design, the cohorts of cases were selected based on the original radiology reports and subsequently interpreted in a consensus manner by neuroradiologists. The rationale for this approach is that the binary presence or absence of each finding is confirmed for every case based on the radiologic images alone, which matches what the model uses. It is, however, possible for the interpretations to change between the original reports and the neuroradiologist interpretations especially given the original reports can be aided by knowledge of the clinical situation and longitudinal radiology scans. All cohorts had a smaller number of positive cases based on the consensus neuroradiologist interpretations compared to the original reports. These cases would be considered ground truth negative for the analysis and, if the model identified them as positive, then they would have been recorded as false positive

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

363 cases with subsequently decreased specificity. The real-world specificity may therefore be

364 higher than what is recorded as part of this study.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Limitations

A key limitation of this study is that it is a retrospective study outside of the clinical workflow. As was noted in the similar assessment of a pneumothorax model, this study therefore establishes the accuracy of the model but does not assess its impact on the clinical workflow including for case prioritization and patient outcomes.[23] This initial step is necessary to ensure the model has the potential to provide clinical benefit. Further evaluation will be required moving forward to prove such benefit.

This study examines the ability of the model to identify each hemorrhage subtype independently. It does not consider the overlap of hemorrhage subtypes, which is common in the clinical environment; as an example, intraventricular hemorrhage occurs in 30 to 50% of patients with intra-axial hemorrhage.[5] This overlap may lead to redundancy especially when the model identifies a first hemorrhage subtype but then misses a second or third subtype. It is also possible that the model incorrectly identifies a second or third subtype when only a first subtype is present. Further research may consider the application of this model when multiple hemorrhage subtypes are present.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Conclusion

This diagnostic study assessed an AI model that accurately detected four subtypes of intracranial hemorrhage including acute subdural/epidural hematoma, acute subarachnoid hemorrhage, intra-axial hemorrhage and intraventricular hemorrhage. Its use in the clinical environment may lead to improved care and outcomes for patients with intracranial hemorrhage.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Acknowledgements

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

Funding

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. JMH, BCB, INC, SFM, JKC, AG, EL, ALM, ASS, KB, JC, SP, SR, KJD, WAM are employees of Mass General Brigham and/or Massachusetts General Hospital, which had received institutional funding from Annalise-AI for the study.

Access to data and data analysis

JMH had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. SFM performed the statistical analyses.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Dublin AB, French BN, Rennick JM. Computed tomography in head trauma. *Radiology*. Feb 1977;122(2):365-9. doi:10.1148/122.2.365

2. Mattiello JA, Munz M. Images in clinical medicine. Four types of acute post-traumatic intracranial hemorrhage. *N Engl J Med*. Feb 22 2001;344(8):580. doi:10.1056/NEJM200102223440806

3. Bullock MR, Chesnut R, Ghajar J, et al. Surgical management of acute subdural hematomas. *Neurosurgery*. Mar 2006;58(3 Suppl):S16-24; discussion Si-iv.

4. Lawton MT, Vates GE. Subarachnoid Hemorrhage. *N Engl J Med*. Jul 20 2017;377(3):257-266. doi:10.1056/NEJMcp1605827

5. Sheth KN. Spontaneous Intracerebral Hemorrhage. *N Engl J Med*. Oct 27 2022;387(17):1589-1596. doi:10.1056/NEJMra2201449

6. American College of Radiology. AI Central. Accessed August 19 2023, https://aicentral.acrdsi.org/

7. US Food and Drug Administration. K221456 (Rapid ICH). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf22/K221456.pdf

8. US Food and Drug Administration. K210209 (Viz ICH). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf21/K210209.pdf

9. US Food and Drug Administration. K223240 (Annalise Enterprise CTB Triage Trauma). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf22/K223240.pdf

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

10. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. Jul 2019;95:103208. doi:10.1016/j.jbi.2019.103208

11. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. Apr 2009;42(2):377-81. doi:10.1016/j.jbi.2008.08.010

12. 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;

13. US Food and Drug Administration. K200921 (qER). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf20/K200921.pdf

14. US Food and Drug Administration. K193351 (NinesAI). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf19/K193351.pdf

15. US Food and Drug Administration. K211179 (InferRead CT Stroke.AI). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf21/K211179.pdf

16. US Food and Drug Administration. K203260 (syngo.CT Brain Hemorrhage). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf20/K203260.pdf

17. US Food and Drug Administration. K190424 (HealthICH). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190424.pdf

18. US Food and Drug Administration. K182875 (DeepCT). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf18/K182875.pdf

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

19. US Food and Drug Administration. K192167 (CuraRad-ICH). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf19/K192167.pdf

20. US Food and Drug Administration. K221716 (Cina). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf22/K221716.pdf

21. US Food and Drug Administration. K221240 (BriefCase). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf22/K221240.pdf

22. US Food and Drug Administration. K201310 (Accipiolx). Accessed August 19 2023, https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201310.pdf

23. 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*. Dec 1 2022;5(12):e2247172. doi:10.1001/jamanetworkopen.2022.47172

24. Alshipli M, Kabir NA. Effect of slice thickness on image noise and diagnostic content of single-source-dual energy computed tomography. *Journal of Physics: Conference Series*. 2017/05/01 2017;851(1):012005. doi:10.1088/1742-6596/851/1/012005

Tables

Table 1: Demographic and technical breakdown of CT head cases for each finding.

	Acute subdural/epidural hematoma				Acute subarachnoid hemorrhage				Intra-axial hemorrhage				Intraventricular hemorrhage			
	Thin series		Thick series		Thin series		Thick series		Thin series		Thick series		Thin series		Thick series	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Total	308	101	401	138	149	89	184	108	484	226	591	293	74	79	91	96
Sex																
Female	131 (42.5%)	54 (53.5%)	172 (42.9%)	69 (50.0%)	87 (58.4%)	38 (42.7%)	97 (52.7%)	51 (47.2%)	214 (44.2%)	116 (51.3%)	258 (43.7%)	153 (52.2%)	40 (54.1%)	37 (46.8%)	46 (50.5%)	46 (47.9%)
Male	177 (57.5%)	47 (46.5%)	229 (57.1%)	69 (50.0%)	62 (41.6%)	51 (57.3%)	87 (47.3%)	57 (52.8%)	270 (55.8%)	110 (48.7%)	333 (56.3%)	140 (47.8%)	34 (45.9%)	42 (53.2%)	45 (49.5%)	50 (52.1%)
Age																
≤65 years	113 (36.7%)	44 (43.6%)	148 (36.9%)	61 (44.2%)	42 (28.2%)	53 (59.6%)	55 (29.9%)	59 (54.6%)	196 (40.5%)	102 (45.1%)	228 (38.6%)	127 (43.3%)	25 (33.8%)	32 (40.5%)	30 (33.0%)	41 (42.7%)
>65 years	195 (63.3%)	57 (56.4%)	253 (63.1%)	77 (55.8%)	107 (71.8%)	36 (40.4%)	129 (70.1%)	49 (45.4%)	288 (59.5%)	124 (54.9%)	363 (61.4%)	166 (56.7%)	49 (66.2%)	47 (59.5%)	61 (67.0%)	55 (57.3%)
Mean ± SD (years)	68.1 ± 19.0	63.8 ± 19.7	67.7 ± 18.4	64.0 ± 20.4	71.3 ± 16.8	58.5 ± 21.2	70.3 ± 16.8	60.3 ± 20.2	66.5 ± 18.1	64.8 ± 18.9	67.2 ± 17.6	65.7 ± 18.7	69.5 ± 19.2	63.5 ± 21.1	69.7 ± 18.1	64.4 ± 20.8
Ethnicity																
Hispanic	26 (8.4%)	16 (15.8%)	34 (8.5%)	17 (12.3%)	7 (4.7%)	11 (12.4%)	6 (3.3%)	11 (10.2%)	41 (8.5%)	18 (8.0%)	48 (8.1%)	22 (7.5%)	9 (12.2%)	9 (11.4%)	9 (9.9%)	9 (9.4%)
Not Hispanic	266 (86.4%)	79 (78.2%)	350 (87.3%)	115 (83.3%)	130 (87.2%)	70 (78.7%)	164 (89.1%)	89 (82.4%)	411 (84.9%)	196 (86.7%)	506 (85.6%)	257 (87.7%)	63 (85.1%)	68 (86.1%)	78 (85.7%)	85 (88.5%)
Prefer not to say / Decline	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (2.2%)	0 (0.0%)	2 (1.9%)	2 (0.4%)	1 (0.4%)	3 (0.5%)	1 (0.3%)	0 (0.0%)	1 (1.3%)	0 (0.0%)	1 (1.0%)
Unavailable	16 (5.2%)	6 (5.9%)	17 (4.2%)	6 (4.3%)	12 (8.1%)	6 (6.7%)	14 (7.6%)	6 (5.6%)	30 (6.2%)	11 (4.9%)	34 (5.8%)	13 (4.4%)	2 (2.7%)	1 (1.3%)	4 (4.4%)	1 (1.0%)
Race																
American Indian or Alaska Native	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (1.1%)	1 (0.5%)	1 (0.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Asian	17 (5.5%)	4 (4.0%)	16 (4.0%)	6 (4.3%)	7 (4.7%)	5 (5.6%)	8 (4.3%)	4 (3.7%)	35 (7.2%)	11 (4.9%)	39 (6.6%)	11 (3.8%)	2 (2.7%)	2 (2.5%)	2 (2.2%)	2 (2.1%)
Black or African American	14 (4.5%)	11 (10.9%)	14 (3.5%)	13 (9.4%)	4 (2.7%)	2 (2.2%)	6 (3.3%)	3 (2.8%)	30 (6.2%)	12 (5.3%)	39 (6.6%)	12 (4.1%)	7 (9.5%)	9 (11.4%)	8 (8.8%)	8 (8.3%)
Native Hawaiian or Other Pacific Islander	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.2%)	0 (0.0%)	2 (0.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
White or Caucasian	242 (78.6%)	73 (72.3%)	324 (80.8%)	105 (76.1%)	122 (81.9%)	67 (75.3%)	151 (82.1%)	84 (77.8%)	365 (75.4%)	179 (79.2%)	455 (77.0%)	239 (81.6%)	58 (78.4%)	61 (77.2%)	72 (79.1%)	80 (83.3%)
Other	18 (5.8%)	8 (7.9%)	25 (6.2%)	8 (5.8%)	4 (2.7%)	8 (9.0%)	4 (2.2%)	8 (7.4%)	29 (6.0%)	11 (4.9%)	29 (4.9%)	14 (4.8%)	5 (6.8%)	4 (5.1%)	5 (5.5%)	4 (4.2%)
Two or more	3 (1.0%)	1 (1.0%)	6 (1.5%)	2 (1.4%)	2 (1.3%)	0 (0.0%)	3 (1.6%)	1 (0.9%)	3 (0.6%)	3 (1.3%)	4 (0.7%)	4 (1.4%)	0 (0.0%)	1 (1.3%)	0 (0.0%)	0 (0.0%)
Declined	1 (0.3%)	1 (1.0%)	2 (0.5%)	1 (0.7%)	1 (0.7%)	1 (1.1%)	1 (0.5%)	2 (1.9%)	2 (0.4%)	2 (0.9%)	2 (0.3%)	2 (0.7%)	0 (0.0%)	1 (1.3%)	0 (0.0%)	1 (1.0%)
Unavailable	13 (4.2%)	3 (3.0%)	14 (3.5%)	3 (2.2%)	9 (6.0%)	5 (5.6%)	10 (5.4%)	5 (4.6%)	19 (3.9%)	8 (3.5%)	21 (3.6%)	10 (3.4%)	2 (2.7%)	1 (1.3%)	4 (4.4%)	1 (1.0%)
Manufacturer																
GE Healthcare	101	32	113	42	55	34	50	33	107	59	185	100	24	25	24	24

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

	(32.8%)	(31.7%)	(28.2%)	(30.4%)	(36.9%)	(38.2%)	(27.2%)	(30.6%)	(22.1%)	(26.1%)	(31.3%)	(34.1%)	(32.4%)	(31.6%)	(26.4%)	(25.0%)
NeuroLogica	1 (0.3%)	0 (0.0%)	2 (0.5%)	1 (0.7%)	0 (0.0%)	1 (1.1%)	0 (0.0%)	1 (0.9%)	1 (0.2%)	0 (0.0%)	2 (0.3%)	0 (0.0%)	1 (1.4%)	1 (1.3%)	1 (1.1%)	1 (1.0%)
Philips	1 (0.3%)	0 (0.0%)	1 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.2%)	0 (0.0%)	1 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Siemens	204 (66.2%)	68 (67.3%)	200 (49.9%)	67 (48.6%)	93 (62.4%)	54 (60.7%)	94 (51.1%)	52 (48.1%)	372 (76.9%)	166 (73.5%)	368 (62.3%)	162 (55.3%)	49 (66.2%)	53 (67.1%)	48 (52.7%)	49 (51.0%)
Toshiba	1 (0.3%)	1 (1.0%)	85 (21.2%)	28 (20.3%)	1 (0.7%)	0 (0.0%)	40 (21.7%)	22 (20.4%)	3 (0.6%)	1 (0.4%)	35 (5.9%)	31 (10.6%)	0 (0.0%)	0 (0.0%)	18 (19.8%)	22 (22.9%)

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figures

Figure 1: Performance for acute subdural/epidural hematoma. 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 operating point at the operating points described in the text. C, D and E, Example images for true positive (C), false negative (D) and false positive (E) cases. The model classification score output is provided for each case.

Figure 2: Performance for acute subarachnoid hemorrhage. 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 operating point at the operating points described in the text. C, D and E, Example images for true positive (C), false negative (D) and false positive (E) cases. The model classification score output is provided for each case.

Figure 3: Performance for intra-axial hemorrhage. 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 operating point at the operating points described in the text. C, D and E, Example images for true positive (C), false negative (D) and false positive (E) cases. The model classification score output is provided for each case.

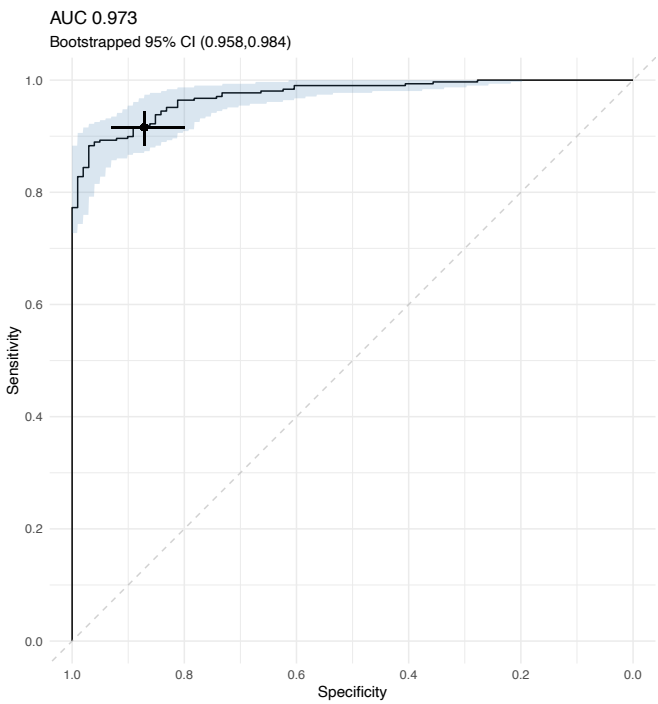
medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 4: Performance for intraventricular hemorrhage. 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 operating point at the operating points described in the text. C, D and E, Example images for true positive (C), false negative (D) and false positive (E) cases. The model classification score output is provided for each case.

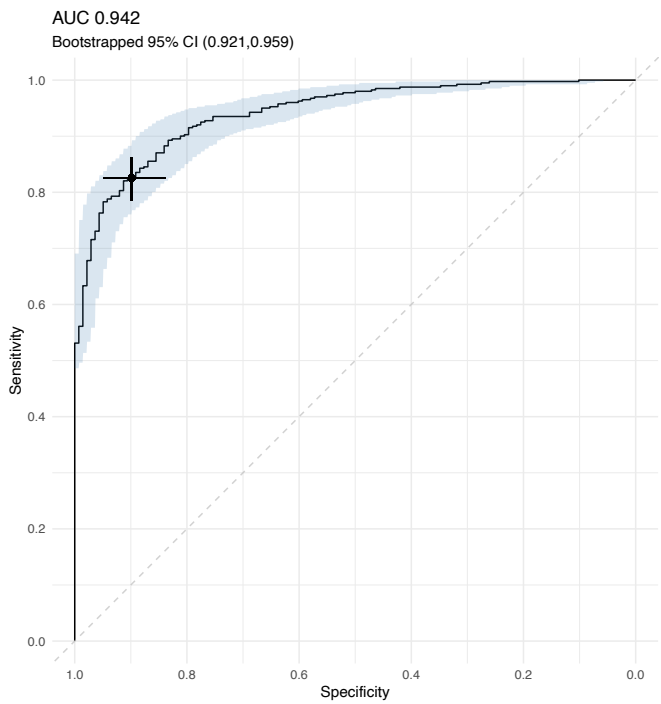
medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Figure 1: Acute subdural / epidural hematoma performance

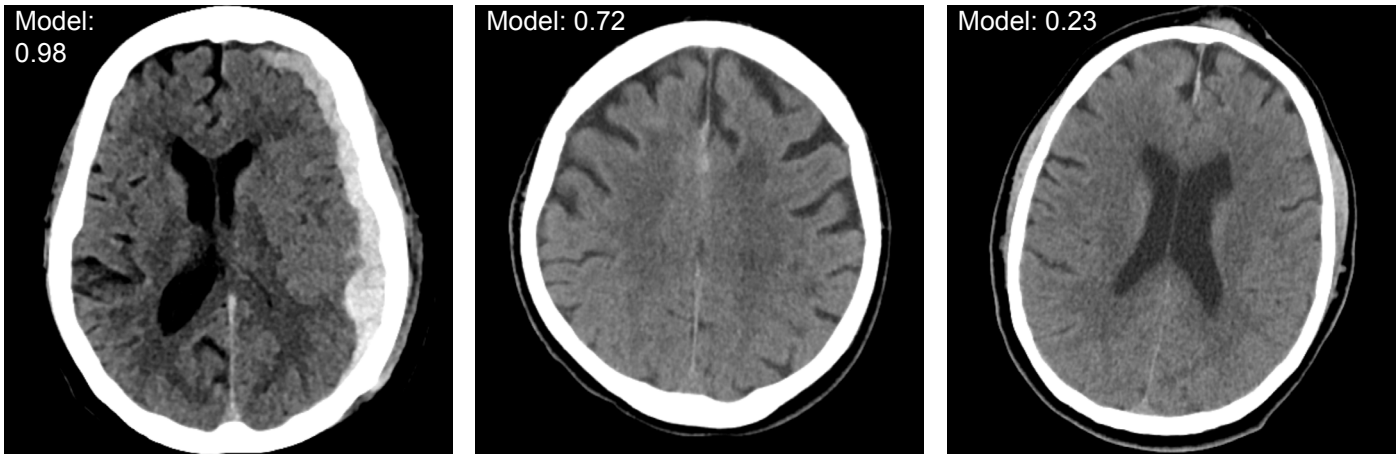
A. Thin series



B. Thick series

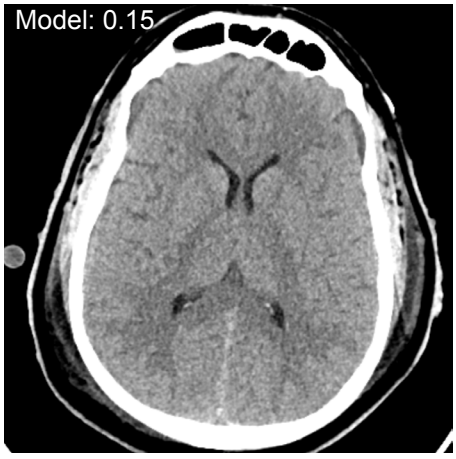


C. True positive cases



D. False negative case

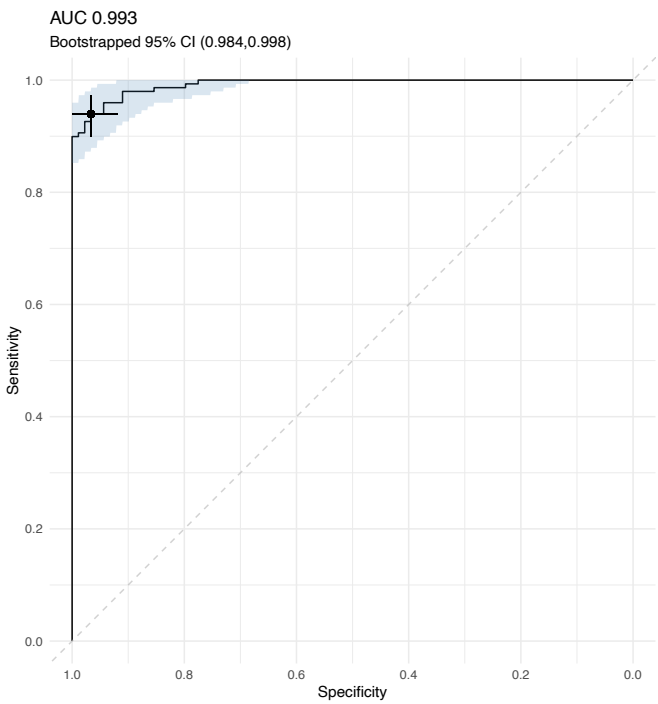
E. False positive case



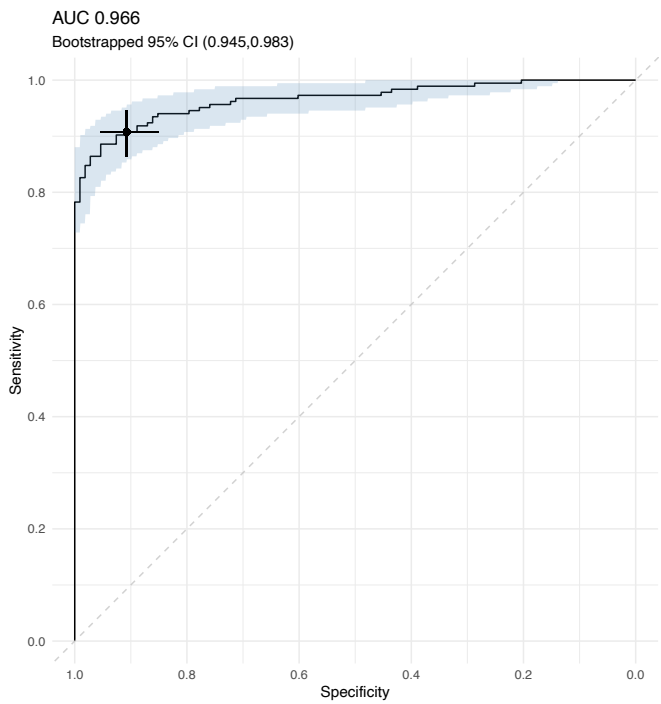
medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Figure 2: Acute subarachnoid hemorrhage performance

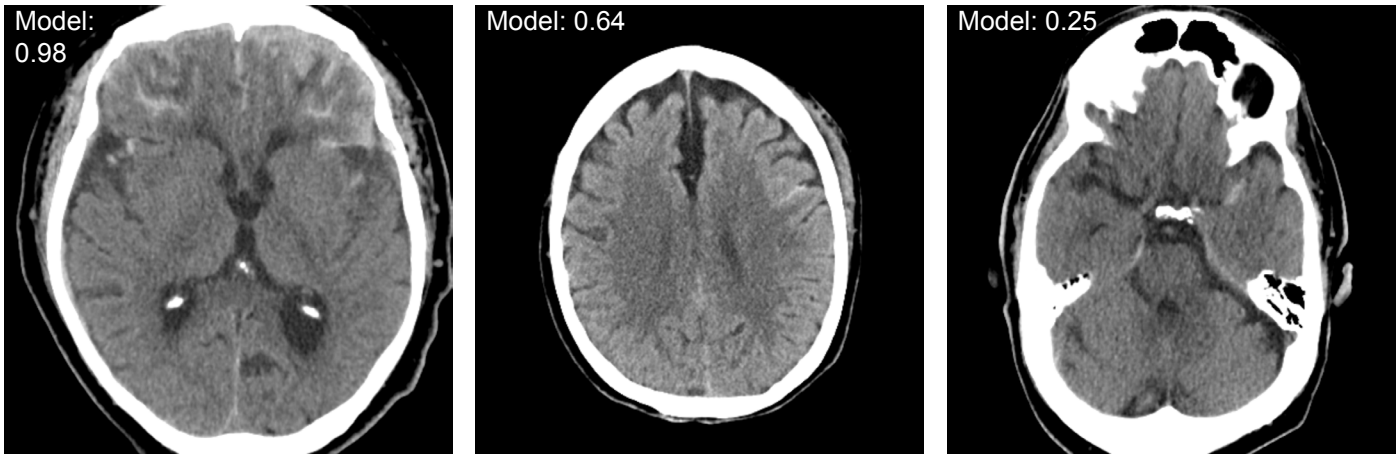
A. Thin series



B. Thick series

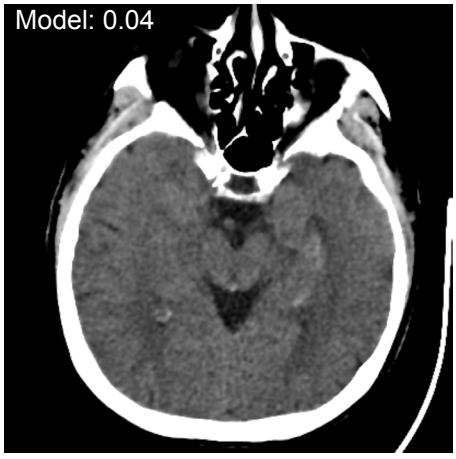


C. True positive cases



D. False negative case

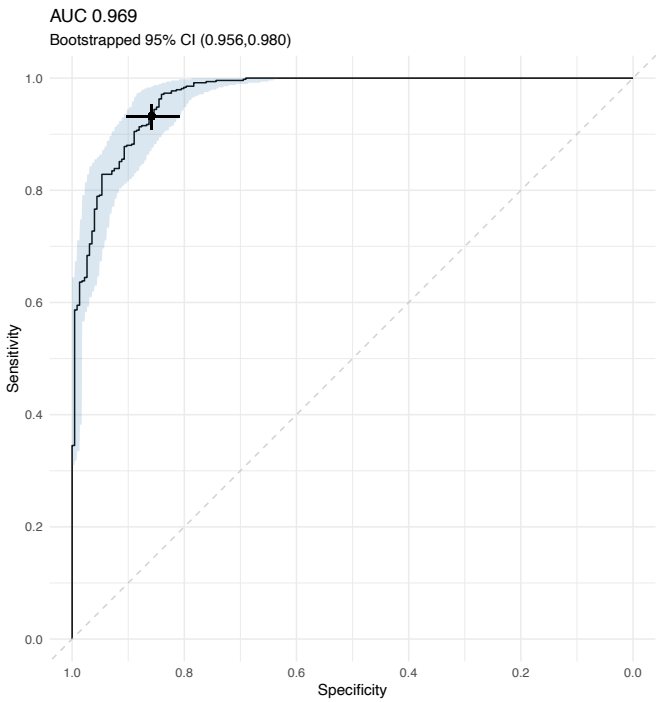
E. False positive case



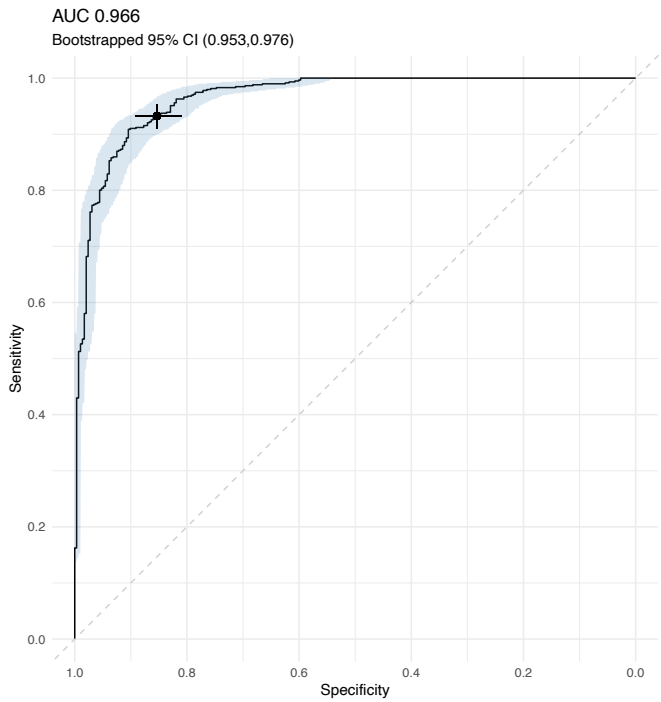
medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Figure 3: Intra-axial hemorrhage performance

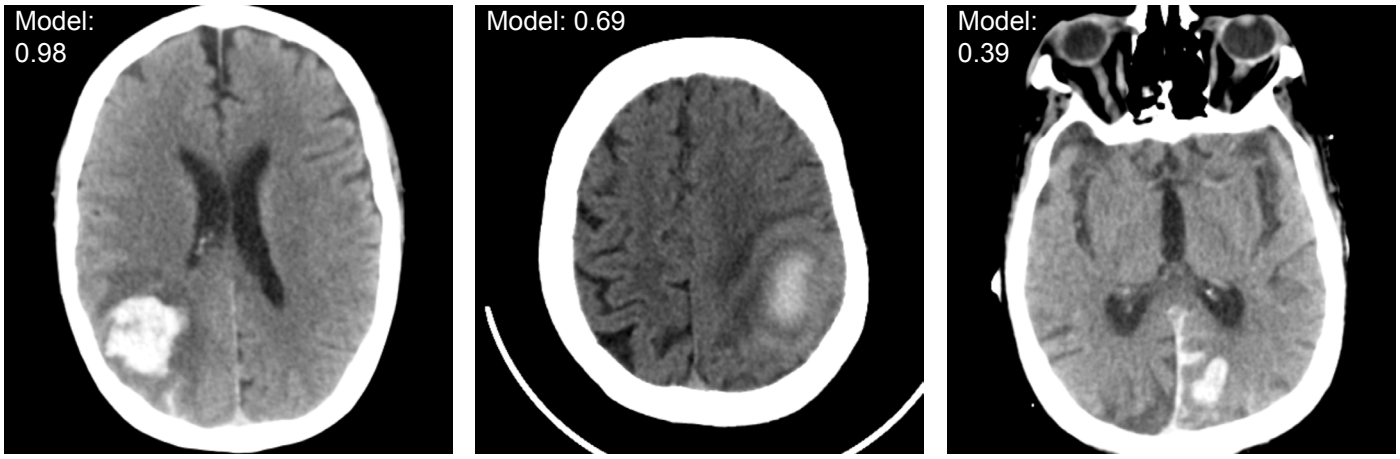
A. Thin series



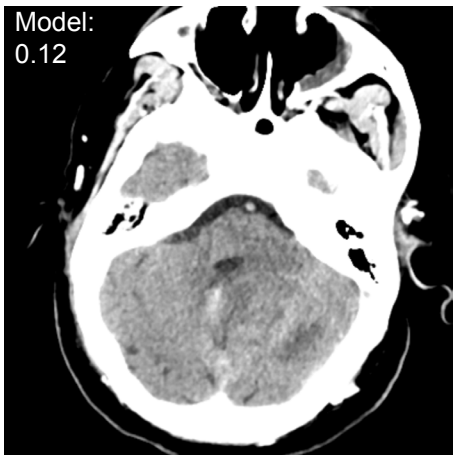
B. Thick series



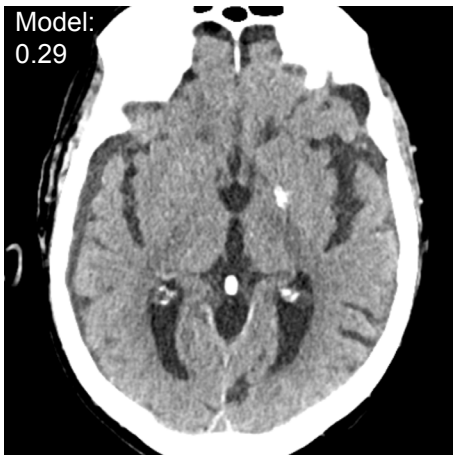
C. True positive cases



D. False negative case



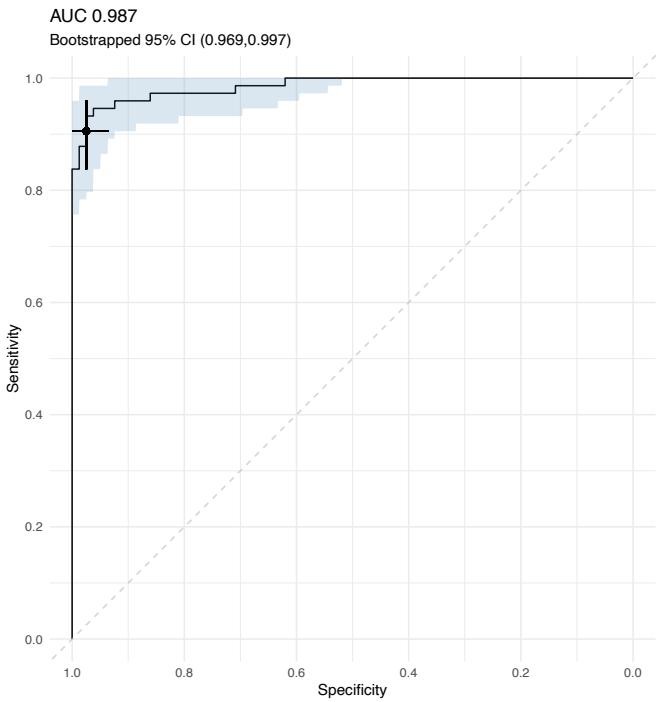
E. False positive case



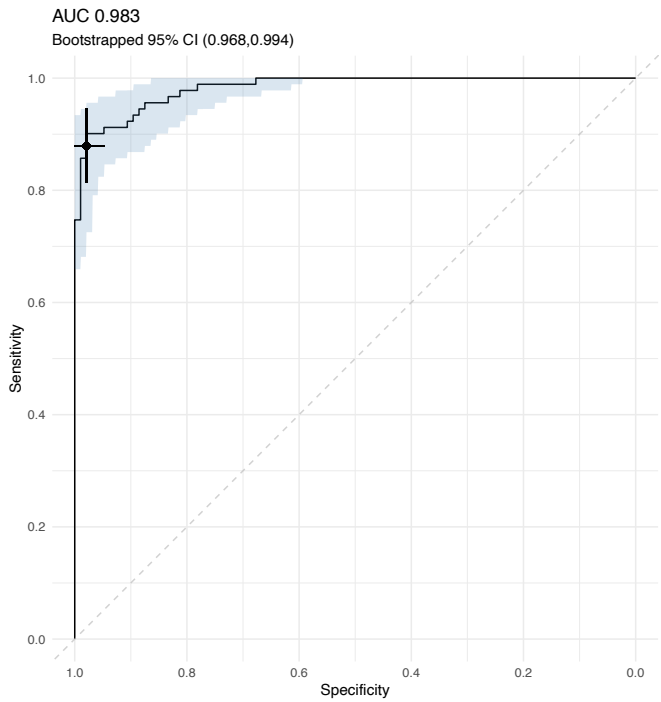
medRxiv preprint doi: https://doi.org/10.1101/2023.09.07.23295189; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Figure 4: Intraventricular hemorrhage performance

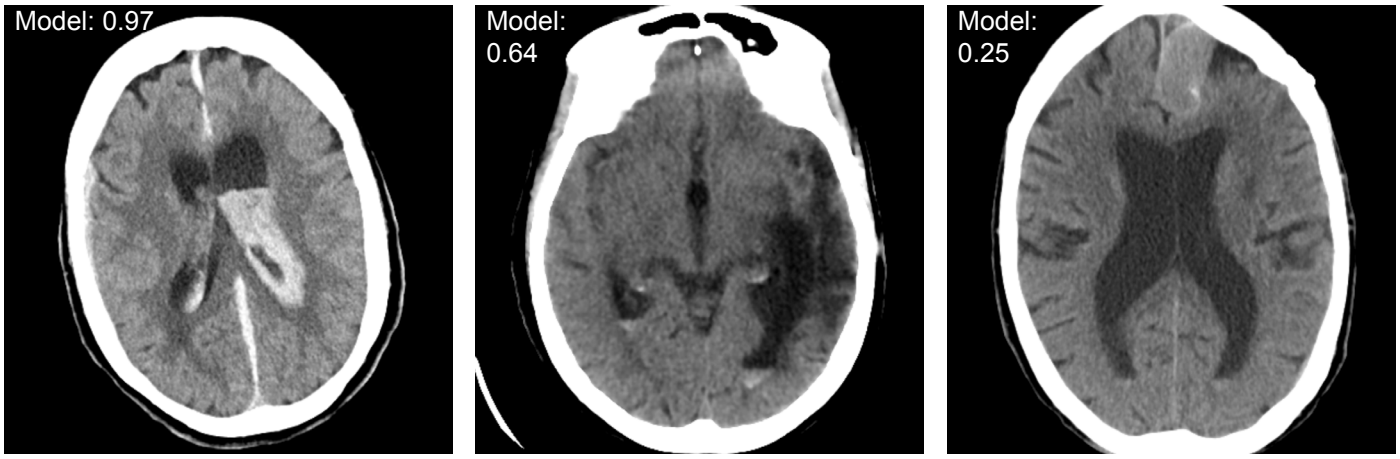
A. Thin series



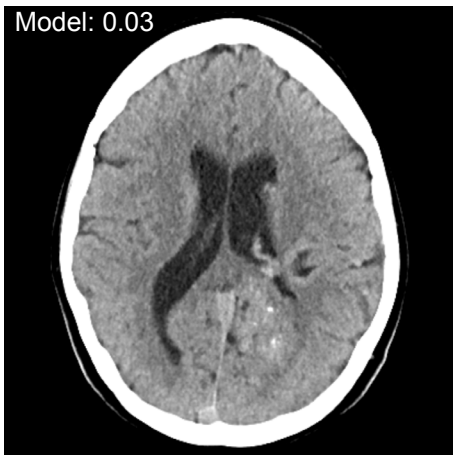
B. Thick series



C. True positive cases



D. False negative case



E. False positive case