A deep-learning model for intracranial aneurysm detection on CT angiography images in China: a stepwise, multicentre, early-stage clinical validation study

Bin Hu, Zhao Shi*, Li Lu*, Zhongchang Miao*, Hao Wang*, Zhen Zhou*, Fandong Zhang, Rongpin Wang, Xiao Luo, Feng Xu, Sheng Li, Xiangming Fang, Xiaodong Wang, Ge Yan, Fajin Lv, Meng Zhang, Qiu Sun, Guangbin Cui, Yubao Liu, Shu Zhang, Chengwei Pan, Zhibo Hou, Huiying Liang, Yuning Pan, Xiaoxia Chen, Xiaorong Li, Fei Zhou, U Joseph Schoepf, Akos Varga-Szemes, W Garrison Moore, Yizhou Yu†, Chunfeng Hu†, Long Jiang Zhang†, on behalf of the China Aneurysm AI Project Group‡*

Summary

Background Artificial intelligence (AI) models in real-world implementation are scarce. Our study aimed to develop a CT angiography (CTA)-based AI model for intracranial aneurysm detection, assess how it helps clinicians improve diagnostic performance, and validate its application in real-world clinical implementation.

Methods We developed a deep-learning model using 16 546 head and neck CTA examination images from 14 517 patients at eight Chinese hospitals. Using an adapted, stepwise implementation and evaluation, 120 certified clinicians from 15 geographically different hospitals were recruited. Initially, the AI model was externally validated with images of 900 digital subtraction angiography-verified CTA cases (examinations) and compared with the performance of 24 clinicians who each viewed 300 of these cases (stage 1). Next, as a further external validation a multi-reader multi-case study enrolled 48 clinicians to individually review 298 digital subtraction angiography-verified CTA cases (stage 2). The clinicians reviewed each CTA examination twice (ie, with and without the AI model), separated by a 4-week washout period. Then, a randomised open-label comparison study enrolled 48 clinicians to assess the acceptance and performance of this AI model (stage 3). Finally, the model was prospectively deployed and validated in 1562 real-world clinical CTA cases.

Findings The AI model in the internal dataset achieved a patient-level diagnostic sensitivity of 0·957 (95% CI 0·939–0·971) and a higher patient-level diagnostic sensitivity than clinicians (0·943 [0·921–0·961] *vs* 0·658 [0·644–0·672]; p<0·0001) in the external dataset. In the multi-reader multi-case study, the AI-assisted strategy improved clinicians' diagnostic performance both on a per-patient basis (the area under the receiver operating characteristic curves [AUCs]; 0·795 [0·761–0·830] without AI *vs* 0·878 [0·850–0·906] with AI; p<0·0001) and a per-aneurysm basis (the area under the weighted alternative free-response receiver operating characteristic curves; 0·765 [0·732–0·799] *vs* 0·865 [0·839–0·891]; p<0·0001). Reading time decreased with the aid of the AI model (87·5 s *vs* 82·7 s, p<0·0001). In the randomised open-label comparison study, clinicians in the AI-assisted group had a high acceptance of the AI model (92·6% adoption rate), and a higher AUC when compared with the control group (0·858 [95% CI 0·850–0·866] *vs* 0·789 [0·780–0·799]; p<0·0001). In the prospective study, the AI model had a 0·51% (8/1570) error rate due to poor-quality CTA images and recognition failure. The model had a high negative predictive value of 0·998 (0·994–1·000) and significantly improved the diagnostic performance of clinicians; AUC improved from 0·787 (95% CI 0·766–0·808) to 0·909 (0·894–0·923; p<0·0001) and patient-level sensitivity improved from 0·590 (0·511–0·666) to 0·825 (0·759–0·880; p<0·0001).

Interpretation This AI model demonstrated strong clinical potential for intracranial aneurysm detection with improved clinician diagnostic performance, high acceptance, and practical implementation in real-world clinical cases.

Funding National Natural Science Foundation of China.

Lancet Digit Health 2024; 6: e261–71

*Contributed equally

†Senior authors

‡Members listed at the end of the Article

For the Chinese translation of the abstract please see Online for appendix 1

Department of Radiology, Jinling Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, China (B Hu MS, Z Shi MS, Prof L J Zhang PhD); **Department of Radiology, the Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu, China** (L Lu MD, Prof C Hu MD); **Department of Medical Imaging, the First People's Hospital of Lianyungang, Lianyungang, Jiangsu, China** (Z Miao BS); **Deepwise Artificial Intelligence (AI) Lab, Deepwise, Beijing, China** (H Wang MS, Z Zhou PhD, F Zhang PhD, S Zhang PhD); **Department of Medical Imaging, Guizhou Province People's Hospital, Guiyang, Guizhou, China** (Prof R Wang MD); **Department of Radiology, Ma'anshan People's Hospital, Ma'anshan, Anhui, China** (X Luo BS); **Department of Medical Imaging, the Affiliated Suqian First People's Hospital of Nanjing Medical University, Suqian, Jiangsu, China** (F Xu MS); **Department of Radiology, People's Hospital, Hubei University of Medicine, Shiyan, Hubei, China** (S Li MS); **Department of Medical Imaging, the Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi, Jiangsu, China** (X Fang PhD); **Department of Radiology, General Hospital of Ningxia Medical University, Yinchuan, Ningxia, China** (X Wang MS);

Introduction

Intracranial aneurysm is a common cerebrovascular disease with a prevalence of 3·2% in the global population and 7·0% in China.[1,2] Aneurysmal rupture is the leading cause of subarachnoid haemorrhage. The case fatality after aneurysmal subarachnoid haemorrhage is up to 50%.[3–5] Early and accurate diagnosis could improve the clinical management and prognosis in patients with intracranial aneurysms and subarachnoid haemorrhage. CT angiography (CTA) has been considered the modality of choice by front-line clinicians for intracranial aneurysm detection.[6,7] The use of CTA continues to increase worldwide, especially in lower-income and middle-income countries, due to its rapidity

Department of Medical Imaging, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China (G Yan MD); Department of Radiology, the First Affiliated Hospital of Chongqing Medical University, Chongqing, China (Prof F Lv MD); Department of Radiology, People's Hospital of Sanya, Sanya, Hainan, China (M Zhang MD); Department of Radiology, Lanzhou University Second Hospital, Lanzhou, Gansu, China (Q Sun MS); Department of Radiology, Tangdu Hospital, Air Force Medical University (Fourth Military Medical University), Xi'an, Shaanxi, China (G Cui PhD); Medical Imaging Center, Shenzhen Hospital of Southern Medical University, Shenzhen, Guangdong, China (Y Liu MD); Institute of Artificial Intelligence, Beihang University, Beijing, China (C Pan PhD); Department of Radiology, Medical Imaging Center, Peking University Shougang Hospital, Beijing, China (Z Hou BS); Medical Big Data Center, Guangdong Provincial People's Hospital, Guangzhou Guangdong, China (H Liang MD); Department of Radiology, Ningbo First Hospital, Ningbo, Zhejiang, China (Y Pan MD); Department of Radiology, Third Center Medical Center, Chinese PLA General Hospital, Beijing, China (X Chen PhD); Department of Radiology, General Hospital of Southern Theater Command, PLA, Guangzhou, Guangdong, China (X Li MD); Department of Radiology, Central Hospital of Jilin City, Jilin, China (F Zhou MD); Division of Cardiovascular Imaging, Department of Radiology and Radiological Science, Medical University of South Carolina, Charleston, SC, USA (Prof U J Schoepf MD, Prof A Varga-Szemes PhD, W Garrison Moore BS); Department of Computer Science, The University of Hong Kong, Hong Kong Special Administrative Region, China (Prof Y Yu PhD)

Research in context

Evidence before this study

We searched PubMed from the inception of the database to Dec 20, 2022 for research articles with the search terms "artificial intelligence" OR "deep learning" OR "machine learning" AND "intracranial aneurysm" OR "cerebral aneurysm" OR "brain aneurysm" AND "CTA" OR "CT angiography" OR "computed tomography angiography" without language restrictions. We identified 15 studies concerning the development and validation of artificial intelligence (AI) models for intracranial aneurysm detection. However, most studies employed retrospective, single-centre, and case–control designs with relatively small and unbalanced samples. We found that only five publications had reported the AI-assisted diagnostic results, but involved only a small group of clinicians (less than eight) and did not appropriately address the underlying verification bias or spectrum bias of the validation dataset. There were no publications that attempted to bridge the development-to-implementation gap, report how the AI model helps clinicians improve diagnostic performance, and validate its performance in real-world clinical implementation.

Added value of this study

According to our literature search results, we developed and trained an accurate, robust AI model with highly clinically applicable potential based on the largest multicentre CTA dataset reported to date to the best of our knowledge. This is also the first study to our knowledge to evaluate an AI-assisted strategy for intracranial aneurysm detection in a multi-reader multi-case study with digital subtraction angiography as the reference standard and further explore its application potential in a prospective validation study. With clinicians from geographically different hospitals across China, this stepwise validation study demonstrated that the AI-assisted strategy could markedly improve the diagnostic performance of radiologists and interventional neurosurgeons regarding intracranial aneurysm detection. When implemented in real-world clinical practice, the AI model performed effectively as an additional diagnostic tool, acting as a second reader, with a high negative predictive value and significantly improving aneurysm detection performance.

Implications of all the available evidence

Our study presents an accurate, robust, and generalisable AI model for intracranial aneurysm detection. Radiologists and interventional neurosurgeons could significantly improve their diagnoses when assisted by this AI model. Effective implementation of this AI model has the promising potential to augment clinicians' diagnostic performance, reduce their workloads by reducing image reading time, and improve clinical practice related to aneurysm diagnosis.

and non-invasiveness. However, CTA interpretation by clinicians is challenging and time-consuming.[8] Intracranial aneurysms are often missed or misdiagnosed in routine clinical practice due to their relatively small size and the complexity of intracranial vasculature.[9,10] This is especially true for less experienced clinicians, resulting in high inter-reader and inter-study variability. The missed rate of small aneurysms (<5 mm in size) was up to 40% in a CTA study.[9] Given the lethal risk of underlying aneurysmal subarachnoid haemorrhage, additional methods are warranted to reduce CTA's diagnostic variability.

Artificial intelligence (AI), especially deep-learning algorithms, has shown promising potential in performing diagnostic tasks in medical imaging.[11,12] To build trust in the implementation of medical AI systems into clinical workflows, stronger standards for the process of AI model validation are required to demonstrate their impact on clinician performance and how models assist them with their tasks. Early-stage clinical evaluation before a large-scale clinical trial is crucial for assessing the true clinical performance of an AI system, evaluating its impact on human performance and workload, and ensuring its safety.[13,14] Thus, an adapted, stepwise, early-stage clinical implementation and evaluation is necessary for an AI model.

Several studies reported the competence of AI models for intracranial aneurysm detection.[15–17] The deep-learning algorithms applied to CTA have achieved better diagnostic performances than experienced clinicians. Nevertheless, to the best of our knowledge, deep-learning algorithms for intracranial aneurysm detection in real-world clinical implementation are still scarce. In this study, we developed a deep-learning algorithm for intracranial aneurysm detection from a large selection of CTA images. We assessed the overall performance of this AI model in an independent dataset and compared it with radiologists, who generate diagnostic reports as part of their regular work, and with interventional neurosurgeons, who employ the aneurysm diagnosis to guide its management. Then, a multi-reader multi-case study and a randomised open-label comparison study were performed to evaluate the AI-assisted diagnostic strategy of radiologists and interventional neurosurgeons in different hospitals across China. Finally, a prospective validation study was conducted to bridge the gap between theory and practice, demonstrating the impact of the AI model when it was deployed in a real-world clinical scenario.

Methods

Study design, datasets, and readers

An overview of the study design is presented in figure 1 and appendix 2 (p 11). This is an adapted, stepwise AI validation study involving 15 provinces across China (appendix 2 p 12). A total of 27 hospitals were involved: eight hospitals provided model training and internal validation data, seven hospitals provided model external

validation data, and 15 hospitals were involved in the reader study (three hospitals had previously provided training or validation data). We first retrospectively collected 16 546 head and neck CTA examinations from Jinling Hospital and seven other hospitals, which were randomly split into a training dataset with 14 613 CTA examinations from 12 817 patients, and an internal validation dataset with 1933 CTA examinations of 1700 patients. Then, we retrospectively collected two independent datasets (external validation dataset 1: 900 CTA examinations for a comparison study between clinicians and the AI [stage 1]; external validation dataset 2: 298 CTA examinations for a multi-reader multi-case study [stage 2] and a randomised open-label comparison study [stage 3]). All cases (examinations) were verified by digital subtraction angiography, the gold standard for detecting intracranial aneurysms due to its higher spatial resolution than CTA, with a reported sensitivity of 99%.[18] There was no overlap among patients or CTA examinations in the training dataset, internal validation dataset, or two gold-standard external validation datasets. Finally, a prospective validation dataset (stage 4) was conducted with 1562 head and neck CTA examinations. The detailed inclusion and exclusion criteria, ground truth labels, patient flow, distribution, and CT scanners of the dataset involved in this study are shown in appendix 2 (pp 2–3, 16, 19, 21). CTA examinations from patients who had a specific medical condition (moyamoya disease, arteriovenous malformations, arteriovenous fistula, arterial occlusion, or arterial dissection) were excluded in the retrospective datasets, and CTA examinations from patients who were younger than 18 years or had a previous aneurysm diagnosis were excluded from the prospective dataset.

In total, 120 certified clinicians with 2–32 years (median 8 years [IQR 4–15]) of clinical experience from 15 hospitals in ten provinces across China were recruited (appendix 2 p 20). Each hospital recruited eight clinicians including two resident radiologists, two attending radiologists, two senior radiologists, and two interventional neurosurgeons (appendix 2 p 4). Clinicians were assigned to different studies, with a hospital as the allocation unit. Hospitals were selected to balance the geographical distribution across China. Each clinician participated in only one of the studies and did not interpret the CTA examinations of their own institutions in stages 1–3. Eight radiologists who had participated in previous stages 1–3 were recruited in a prospective validation study (stage 4) to interpret routine CTA examinations consecutively performed in their own institution for 4 weeks.

Data collection protocols and the use of CTA images, radiological reports, and clinical information were approved by the ethical committee of Jinling Hospital and all other participating hospitals. Informed consent was waived for retrospectively collected CTA images, and written informed consent was obtained from patients whose CTA examinations were prospectively collected. All procedures followed the tenets of the Declaration of Helsinki. This study is reported in accordance with the DECIDE-AI reporting guidelines (appendix 2 p 29).

Correspondence to:
Prof Long Jiang Zhang, Department of Radiology, Jinling Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing 210002, China
kevinzhlj@nju.edu.cn

See **Online** for appendix 2

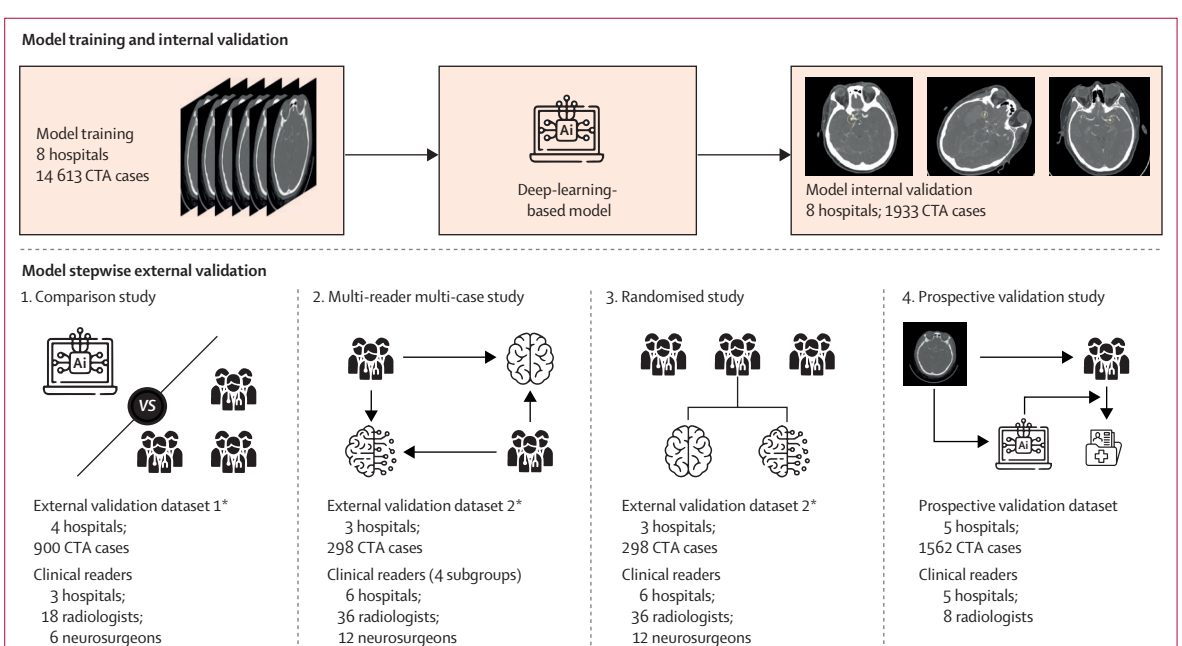

Figure 1: **Study profile**
CTA=CT angiography. *CTA images in both external validation dataset 1 and external validation dataset 2 were verified by digital subtraction angiography, the gold standard for detecting intracranial aneurysms.

Procedures

For deep-learning model training and internal validation, CTA images were obtained from eight hospitals. A dataset of 16 546 head and neck CTA examinations of 14 517 patients was collected, with 2104 (14·5%) of 14 517 patients having digital subtraction angiography reports. By a non-overlap patient-level splitting with a ratio of 15:2, patients were randomly divided into the training dataset and the internal validation dataset.

The model has a cascaded architecture with two major components. Further details are shown in appendix 2 (pp 5, 6, 13). The local fine-grained network captures local image details surrounding each intracranial aneurysm candidate to distinguish between the true positives and false positives predicted by the global context network. As well as the internal validation study, we carried out four-stage studies to comprehensively evaluate this AI model. CTA images in stage 1–3 were all anonymised, and no clinical information or other comparison images were provided for each CTA examination. CTA cases in stage 4 (not anonymised) would be further re-audited by more experienced radiologists to ensure the diagnosis safety. Consequently, the clinicians within the study did a research read rather than a clinical read, in which only the core result of an aneurysm was recorded. For detailed experimental settings, see the appendix 2 (p 7).

For stage 1, we adopted the external validation dataset 1 to compare the detection performance between the AI model alone and 18 radiologists and six interventional neurosurgeons from three hospitals. We randomly allocated the CTA cases to each clinician (n=300 each) with a hospital as the allocation unit. Clinicians diagnosed aneurysms relying on their clinical expertise. Finally, we had three subgroups with 7200 diagnostic results.

For stage 2, we used the external validation dataset 2 to perform a multi-reader multi-case study involving 36 radiologists and 12 interventional neurosurgeons from six geographically different hospitals. It was a cross-over design with a washout period of at least four weeks.[19] All clinicians interpreted each CTA case with and without the assistance of the AI model in a two-rounds cross-over reading process. CTA datasets were anonymised, cleansed, and randomised before each round of reading. A total of 28 608 diagnostic results were generated. Interviews were conducted in person with clinicians with specific questions to pinpoint the reading experience such as contributors to reading speed (recorded automatically) during the two diagnostic process rounds.

For stage 3, a randomised open-label comparison study was conducted using the external validation dataset 2. At this stage, 36 radiologists and 12 interventional neurosurgeons from six different hospitals across China were randomly assigned to two reading groups (control or AI-assisted) in a 1:1 ratio. The assignment was stratified based on their discipline and clinical experience. Clinicians in the control group diagnosed aneurysms relying on their clinical expertise. Clinicians in the AI-assisted group did not have access to the AI results by default unless they requested AI assistance. AI results were accessible with just one click. In this manner, we could measure the acceptance of the AI model among clinicians by recording diagnostic results with and without AI assistance in the AI-assisted group.

For stage 4, the AI model was prospectively deployed and validated between Sept 1, 2022, and Sept 28, 2022 in real-world clinical implementations using CTA examinations collected in five hospitals. The initial CTA interpretations and diagnoses were performed by eight radiologists (each CTA diagnostic report was handled by one radiologist in local hospitals) when reviewing CTA images in the local picture archiving and communication system. Meanwhile, the CTA images would be automatically transferred to a locally deployed AI system (appendix 2 pp 14–15) to output the AI results (which served as a second reader). Then, the final aneurysm-related diagnosis was recorded after one of the radiologists reviewed the AI results. Three independent diagnostic records of radiologists' initial diagnoses (without AI) and final diagnoses (with AI) and AI results for each CTA case were recorded.

Outcomes

The primary measure of performance for the AI model and the clinicians was the area under the receiver operating characteristic (ROC) curve (AUC) or the weighted alternative free-response ROC curve for intracranial aneurysm diagnosis. The primary comparison was the diagnostic performance of clinicians with AI assistance and without AI assistance. The secondary measures included sensitivity, specificity, positive predictive value, negative predictive value, and the average number of false positives in aneurysm diagnoses.

Statistical analysis

The AI model's performance and the clinicians' performance was evaluated based on patient-level and lesion-level metrics (appendix 2 p 8). The statistical analysis was performed according to the subgroups of patients with and without subarachnoid haemorrhage, clinician subgroups, and hospital subgroups. The reading time between clinicians when diagnosis was done with AI assistance and without AI assistance was compared.

According to the largest Youden index in the external datasets, the cutoff point (ie, diagnostic confidence score) was set at 4 or greater. ROC curves were constructed to calculate the AUC using MRMCaov (version 0.2.1) and pROC (version 1.18.0). The diagnostic sensitivity, specificity, positive predictive value, and negative predictive value were calculated by binarising confidence scores for each case. 95% CIs were calculated using the Clopper-Pearson method. The F_1 score was calculated as a harmonic mean of sensitivity and positive predictive value.

In the multi-reader multi-case study, a minimum sample size of 225 CTA cases was required to detect a mean AUC difference of 0·05 in the diagnostic accuracy of the preset 12 readers (alpha=0·05, beta=0·10), and sample size was deemed sufficient for the other three study stages with all post hoc statistical power higher than 0·90 (appendix 2 pp 9, 18). The Dorfman-Berbaum-Metz method with the empirical AUC estimation was used to analyse multi-reader diagnostic performance with AI assistance and without AI assistance.[20] By considering the clinical importance of each lesion, the area under the weighted alternative free-response ROC curve was calculated by the jackknife alternative free-response ROC analysis.[21] In the randomised open-label comparison study, the primary analysis was performed on an intention-to-treat basis. All randomly assigned cases were included in the primary analysis. Additionally, sensitivity analyses were also performed on the diagnosis which was determined without reference to the AI results in the AI-assisted group.

R version 4.13, OR-DBM MRMC version 2.52, and JAFROC version 4.2.1 were used for statistical analysis. Comparisons with a two-sided p value less than 0·05 were considered statistically significant.

For **OR-DBM MRMC** see https://perception.lab.uiowa.edu/or-dbm-mrmc-30-rc2

For **JAFROC** see https://github.com/dpc10ster/WindowsJAFROC

Role of the funding source

The study's funder played no role in the study design, data collection, data analysis, data interpretation, and writing or submitting of the manuscript.

Results

16 546 CTA examinations, collected from eight hospitals, were utilised for the development and internal validation of our model. For external validation, we collected two datasets (external validation datasets 1 and 2) from another seven hospitals, which included 900 (dataset 1)

	Training and internal validation datasets* (n=16 546)		External validation datasets† (n=1198)		Prospective dataset (n=1562)
	Training dataset	Internal validation dataset	External validation dataset 1	External validation dataset 2	
Hospitals	8	8	4	3	5
Patients	12 817	1700	900	298	1562
Mean age, years	61·0 (53·0–69·0)	61·0 (53·0–69·0)	59·0 (51·0–66·0)	62 (53·8–69·3)	62·0 (53·0–70·0)
Sex					
Female	5675 (44·3%)	734 (43·2%)	430 (47·8%)	120 (40·3%)	668 (42·8%)
Male	7142 (55·7%)	966 (56·8%)	470 (52·2%)	178 (59·7%)	894 (57·2%)
CTA	14 613	1933	900	298	1562
CT manufacturers	4	4	4	4	3
CT equipment	17	13	11	9	11
Positive CTA	5321 (36·4%)	699 (36·2%)	562 (62·4%)	140 (47·0%)	166 (10·6%)
Aneurysms	8432	1148	717	177	195
Aneurysm size					
≥10 mm	949 (11·3%)	111 (9·7%)	27 (3·8%)	10 (5·7%)	10 (5·1%)
5 to <10 mm	2935 (34·8%)	392 (34·1%)	210 (29·3%)	43 (24·3%)	39 (20·0%)
3 to <5 mm	2600 (30·8%)	359 (31·3%)	282 (39·3%)	82 (46·3%)	102 (52·3%)
<3 mm	1948 (23·1%)	286 (24·9%)	198 (27·6%)	42 (23·7%)	44 (22·6%)
Aneurysm location					
Internal carotid artery	3827 (45·4%)	507 (44·2%)	308 (43·0%)	95 (53·7%)	93 (47·7%)
Middle cerebral artery	1263 (15·0%)	169 (14·7%)	117 (16·3%)	28 (15·8%)	38 (19·5%)
Anterior cerebral artery	513 (6·1%)	69 (6·0%)	38 (5·3%)	6 (3·4%)	10 (5·1%)
Anterior communicating artery	1124 (13·3%)	153 (13·3%)	93 (13·0%)	23 (13·0%)	12 (6·2%)
Posterior communicating artery	752 (8·9%)	114 (9·9%)	95 (13·2%)	13 (7·3%)	17 (8·7%)
Posterior cerebral artery	256 (3·0%)	28 (2·5%)	22 (3·1%)	3 (1·7%)	7 (3·6%)
Vertebral artery	258 (3·1%)	42 (3·7%)	22 (3·1%)	7 (3·9%)	8 (4·1%)
Basilar artery	257 (3·0%)	46 (4·0%)	8 (1·1%)	1 (0·6%)	8 (4·1%)
Others‡	182 (2·2%)	20 (1·7%)	14 (1·9%)	1 (0·6%)	2 (1·0%)

Data are n or n (%) unless otherwise stated. The n numbers in the top row are the number of CTA examinations. CTA=CT angiography. *16 546 head or head and neck CTA examinations from 14 517 patients were retrospectively collected from eight hospitals, and 2104 (14·5%) patients received digital subtraction angiography, the gold standard for intracranial aneurysm diagnosis. By a non-overlap patient-level splitting, with a ratio of 15:2, they were randomly divided into the training dataset and internal validation dataset. †CTA examinations in two external validation datasets were all verified by digital subtraction angiography. The two external validation datasets were retrospectively collected from another seven independent hospitals which showed no overlap with the training and internal validation datasets. ‡Other locations including posterior inferior cerebellar artery and anterior choroid aneurysm.

Table 1: **Characteristics of the training, internal validation, external validation, and prospective datasets**

and 298 (dataset 2) digital subtraction angiography-verified CTA examinations. Additionally, we prospectively screened 1664 consecutive CTA examinations in five hospitals over 4 weeks and 1562 CTA examinations were finally included (appendix 2 p 16). The detailed demographic characteristics of the datasets used in this study are presented in table 1.

In the internal evaluation dataset, the AI standalone model achieved a patient-level sensitivity of 0·957 (95% CI 0·939–0·971), specificity of 0·797 (0·774–0·820), and lesion-level sensitivity of 0·866 (0·845–0·885). The sensitivities of the AI model stratified by aneurysm size were 0·910 (≥10 mm aneurysm size), 0·960 (5 to <10 mm), 0·954 (3 to <5 mm), and 0·675 (<3 mm). The average number of false positives per case was 0·272 (95% CI 0·252–0·293) with no significant difference between aneurysm-positive and aneurysm-negative CT examinations (0·298 *vs* 0·258 per case; p=0·058). The AI model took an average of 22·2 s (SD 16·2–28·2) to interpret one CTA examination and output its diagnosis.

For stage 1, the comparison of diagnostic performance between the AI model and clinicians is shown in table 2 and the appendix 2 (p 22). The patient-level sensitivity of the AI model (0·943 [95% CI 0·921–0·961]) was higher than that of the clinicians (18 radiologists and six interventional neurosurgeons, 0·658 [0·644–0·672]; p<0·0001). The specificity of the AI model (0·852 [0·810–0·888]) did not differ from that of clinicians (0·852 [0·838–0·865]; p=0·162). Similar findings were shown from all three participating hospitals (appendix 2 p 22). For the subgroups of patients with and without subarachnoid haemorrhage, no differences in diagnostic sensitivity were found for the AI model (0·953 [0·910–0·980] *vs* 0·939 [0·910–0·960], p=0·492); however, the clinicians had higher sensitivity in patients with subarachnoid haemorrhage than in patients without subarachnoid haemorrhage (0·745 [0·721–0·768] *vs* 0·621 [0·603–0·638], p<0·0001). The lesion-level sensitivity of the AI model (0·932 [0·911–0·949]) was higher than that of the clinicians (0·584 [0·571–0·597]; p<0·0001). The sensitivities of the AI model with respect to aneurysm size were 0·963 (26 of 27; ≥10 mm), 0·995 (209 of 210; 5 to <10 mm), 0·986 (278 of 282; 3 to <5 mm), and 0·783 (155/198; <3 mm). Tiny aneurysms (<3 mm) accounted for 43 (87·8%) of 49 of the aneurysms missed by the AI model. The AI model generated an average of 0·187 (95% CI 0·162–0·214) false positives per case when compared with digital subtraction angiography results. No significant difference in false positives was noted between aneurysm-positive versus aneurysm-negative CT examinations (0·194 *vs* 0·175 per case; p=0·470), and among patients with and without subarachnoid haemorrhage (0·206 *vs* 0·182 per case; p=0·467). The most common error recognition pattern was the intracranial arterial infundibulum (101 [60·1%] of 168 cases; appendix 2 p 17).

In the multi-reader multi-case study (stage 2), clinicians with the assistance of the AI model had higher AUC (0·878 [95% CI 0·850–0·906] *vs* 0·795 [0·761–0·830]; p<0·0001) and the area under weighted alternative free-response ROC (0·865 [0·839–0·891] *vs* 0·765 [0·732–0·799]; p<0·0001) than clinicians' standalone diagnoses. Similar results were observed across all four clinician subgroups (figure 2 and appendix 2 p 23). Among these subgroups, with AI assistance, the resident radiologists achieved the highest AUC improvement (0·100 [0·065–0·134]), followed by the interventional neurosurgeons (0·086 [0·048–0·123]), the senior radiologists (0·080 [0·044–0·115]), and the attending radiologists (0·065 [0·037–0·092]), with the individual improvement ranging from 0·014 to 0·172. The patient-level sensitivity with the AI assistance was 0·862 (0·854–0·871), which was higher than that of clinicians' standalone diagnoses (0·680 [0·668–0·691], p<0·0001). Decreased sensitivity was found in only two interventional neurosurgeons, while the other 46 clinicians obtained improved sensitivity ranging from 0·043 to 0·450, with an average value of 0·194 (0·164–0·224). AI-assisted diagnosis improved aneurysm detection, especially for aneurysms less than 5 mm (25·9% diagnostic improvement in sizes ranging from 3 to 5 mm and 31·8% diagnostic improvement in sizes

	Accuracy (95% CI)	p value	Sensitivity (95% CI)	p value	Specificity (95% CI)	p value	Positive predictive value (95% CI)	Negative predictive value (95% CI)	F_1
AI model									
All patients	0·909 (0·888–0·927)	..	0·943 (0·921–0·961)	..	0·852 (0·810–0·888)	..	0·908 (0·881–0·930)	0·911 (0·874–0·940)	0·928
Patients with SAH	0·944 (0·900–0·973)	..	0·953 (0·910–0·980)	..	0·778 (0·400–0·972)	..	0·988 (0·957–0·999)	0·467 (0·213–0·734)	0·974
Patients without SAH	0·900 (0·876–0·921)	0·064*	0·939 (0·910–0·960)	0·492*	0·854 (0·811–0·890)	0·627*	0·884 (0·850–0·913)	0·921 (0·885–0·949)	0·911
Clinicians									
All patients	0·731 (0·721–0·741)	<0·0001†	0·658 (0·644–0·672)	<0·0001†	0·852 (0·838–0·865)	0·162†	0·827 (0·814–0·839)	0·673 (0·657–0·689)	0·769
Patients with SAH	0·748 (0·725–0·770)	..	0·745 (0·721–0·768)	..	0·806 (0·695–0·889)	..	0·986 (0·977–0·993)	0·143 (0·110–0·180)	0·849
Patients without SAH	0·727 (0·715–0·738)	0·105*	0·621 (0·603–0·638)	<0·0001*	0·853 (0·839–0·866)	0·264*	0·834 (0·828–0·849)	0·654 (0·638–0·670)	0·712

AI=artificial intelligence. F_1=a weighted average of the positive predictive value and sensitivity. SAH=subarachnoid haemorrhage. *Diagnostic performance comparison between patients with and without SAH. †Diagnostic performance comparison between clinicians and the AI model.

Table 2: **Comparison of patient-level diagnostic performance between the AI model and clinicians in external validation dataset 1**

less than 3 mm; appendix 2 p 25). Per-hospital subgroup analysis revealed that all of the included hospitals achieved diagnostic improvement with AI assistance (appendix 2 p 23). The CTA interpretation time was reduced with AI assistance (87·5 s *vs* 82·7 s, p<0·0001) but varied among the different subgroups (appendix 2 p 26). Building on the insights from the interview, the following factors expedite diagnostic speed: when AI is not employed, clinicians tend to concentrate their attention on lesions within the Circle of Willis for the aneurysm detection task, and volume rendering images prove advantageous in identifying aneurysms suspected

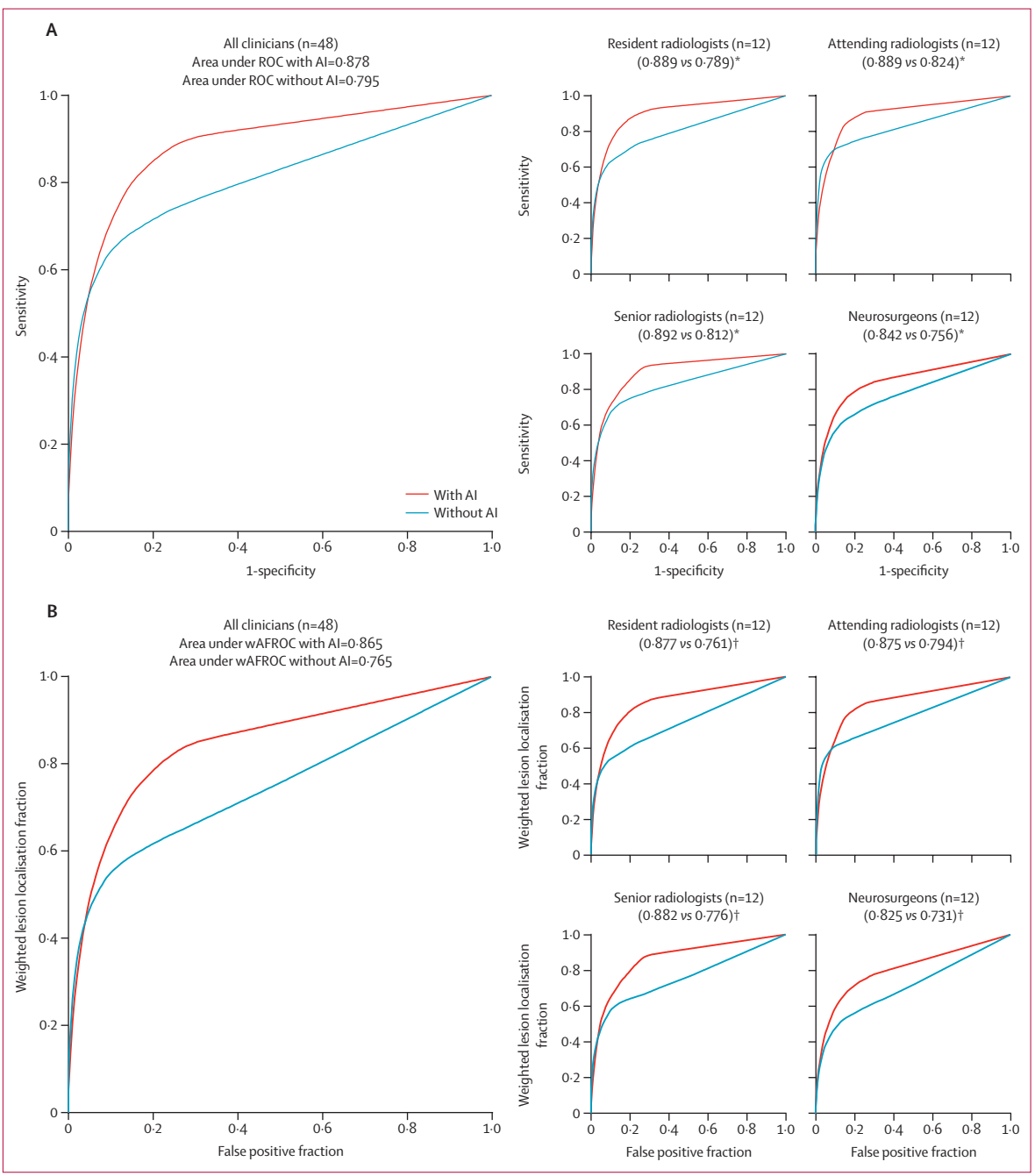

***Figure 2*: Primary comparison of the area under ROC curves (A) and wAFROC curves (B) in the multi-reader multi-case study**
ROC curves (A) and wAFROC curves (B) evaluating the diagnostic performance of clinicians with and without the AI model while interpreting the CTA images. AI=artificial intelligence. CTA=CT angiography. ROC=receiver operating characteristic. wAFROC=the weighted alternative free-response ROC. *Numbers in parentheses are areas under curves; the former is from reading with AI, and the latter is from reading without AI. †Numbers in parentheses are areas under the wAFROC curves; the former is from reading with AI, and the latter is from reading without AI.

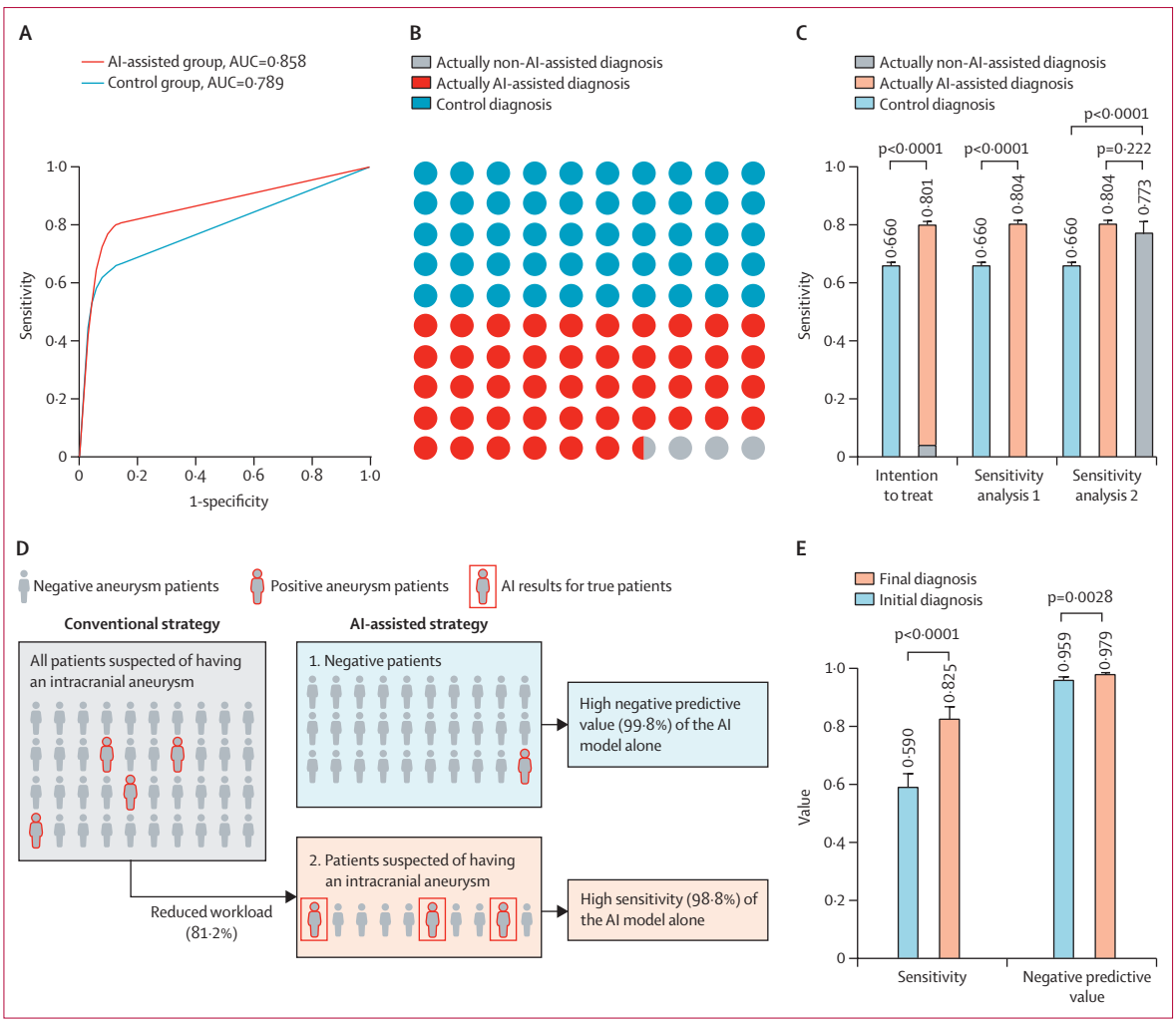

Figure 3: **Impact of the AI model in the randomised open-label comparison study and the prospective validation study**
(A–C) Randomised open-label comparison study results. (A) Receiver operating characteristic curves. (B) Proportion of diagnostic cases employing diverse methodologies. (C) Intention-to-treat and sensitivity analysis. (D–E) Prospective validation study results. (D) Impact of the proposed AI model on real-world clinical practice. (E) Improvement of the AI-assisted strategy. AI=artificial intelligence. AUC=area under the receiver operating characteristic curves.

of being an intracranial aneurysm and ruling out negative cases.

In stage 3, no clinical reader in the AI-assisted group fully switched to the control group. However, 531 (7·4%) of 7132 diagnostic results made by 17 (70·8%) of 24 clinicians in the AI-assisted group were determined without reference to the AI model. This implies that 92·6% of cases embraced the outcomes generated by AI (figure 3B). Among these 17 clinicians, seven (41·2%) reported that more than 10% (ranging from 12·8% to 37·3%) of their diagnoses were made without AI assistance (actually non-AI-assisted). Figure 3A and the appendix 2 (p 27) demonstrate the comparison between the AI-assisted group and the control group based on external validation dataset 2. The clinicians in the AI-assisted group achieved a significantly higher AUC (0·858 [95% CI 0·850–0·866]) than the clinicians in the control group (0·789 [0·780–0·799]; p<0·0001). The patient-level sensitivity in the AI-assisted group was higher than the control group (0·801 [0·787–0·815] *vs* 0·660 [0·644–0·676]; p<0·0001). In the sensitivity analysis which excluded these actually non-AI-assisted diagnoses in the AI-assisted group, the results before and after this exclusion were not different (0·801 [0·787–0·815] *vs* 0·804 [0·789, 0·818]; p=0·801). Of note, we observed that the patient-level sensitivity of the actually non-AI-assisted diagnosis (0·773 [0·719–0·821]) was higher than the control group (0·660 [0·644–0·676]; p<0·0001), and it was comparable with the actual AI-assisted results (0·804 [0·789–0·818]; p=0·222; figure 3C). These actually non-AI-assisted diagnoses were mostly found in negative CTA cases and positive CTA cases with aneurysms ≥5 mm.

In the prospective validation study (stage 4), the error rate due to poor-quality CTA images and failure to analyse by the AI model was 0·5% (eight of 1570 cases). In the initial diagnosis, eight radiologists reported 126 (8·1%) of 1562 aneurysm-related positive results in 1575 CTA examinations. After reviewing the AI results, 203 (13·0%) of 1562 positive results were finally reported, with a total of 82 new positive results from the clinicians and AI, and five new negative results. After reviewing the AI results, the AUC, patient-level sensitivity, lesion-level sensitivity, and negative predictive value of radiologists were improved from 0·787 (95% CI 0·766–0·808) to 0·909 (0·894–0·923; p<0·0001; AUC), 0·590 (0·511–0·666) to 0·825 (0·759–0·880; p<0·0001; patient-level sensitivity), 0·549 (0·476–0·620) to 0·764 (0·698–0·822; p<0·0001; lesion-level sensitivity), and 0·959 (0·947–0·969) to 0·979 (0·969–0·987; p=0·0028; negative predictive value; figure 3E). The median size of the newly detected aneurysms was 3·1 mm (IQR 2·6–3·6). The AI model alone had a high sensitivity of 98·8%, specificity of 81·2%, and negative predictive value of 99·8%. The data reveal that within the AI-assisted strategy, only 18·2% of cases required a double-check of their diagnosis, among whom 98·8% of the true positive aneurysm cases could be detected by the AI model (figure 3D). Appendix 2 (p 28) shows the detailed diagnostic performances of the AI model alone, the radiologists' initial diagnostic results, and the final diagnosis.

Discussion

In this study, we found the AI model alone achieved high diagnostic performance, which was independent of subarachnoid haemorrhage status and exceeded the performance of clinicians. The AI-assisted strategy improved clinicians' diagnostic performance in geographically different hospitals across China. More importantly, the use of the AI model was highly accepted and well implemented in real-world clinical practice as a second reader, as the majority of the clinicians with access to the AI results used these to assist with their diagnoses.

The first advantage of our AI model is its robustness and excellent diagnostic performance. Other available AI models based on CTA images for aneurysm detection have insufficient images for training and external validations.[15–17] However, our AI model is built on 16 546 CTA examinations with 2104 digital subtraction angiography-verified results and was externally validated in three independent, large-scale, and heterogeneous CTA datasets collected across China. Patients with and without subarachnoid haemorrhage were included in this study, and CTA images were generated by various CT scanners, ensuring the robustness and reproducibility of the AI model. Our AI model obtained a high diagnostic sensitivity for aneurysm detection, which showed superiority over clinicians, and it also compared favourably with recent meta-analysis results.[22] Compared with digital subtraction angiography results, our AI model produced substantially lower false positives than a previous study (ie, 0·187 per case compared with 13·8 per case in the study by Yang and colleagues).[17] Fewer false positives could reduce clinicians' workload and avoid blunting their responsiveness to true positive results.[23] This is one of three fundamental elements for deep-learning-based tools for intracranial aneurysm detection.[24]

The second advantage of our AI model is its great potential to improve the diagnostic performance of clinicians. Despite the evidence that some AI-based algorithms match or exceed the accuracy of human experts in preclinical studies, real-world medical practice often involves human-in-the-loop configurations, where humans actively collaborate with AI systems.[10] The multi-reader multi-case study design can effectively account for clinical readers' variability and the resulting added biases.[25,26] We first conducted a multi-reader multi-case study using AI detection for intracranial aneurysms. The results indicated a diagnostic improvement in both patient-level analysis (ROC) and lesion-level analysis (weighted alternative free-response ROC) when employing the AI-assisted strategy for both per-reader and per-hospital groups. The resident radiologists obtained the highest improvement with AI assistance, approaching the diagnostic level of senior radiologists. Only two interventional neurosurgeons (from the same hospital) showed decreased diagnostic sensitivity with AI assistance. One explanation could be their limited experience with AI products. Unlike a previous study which disclosed to the clinicians the accurate diagnostic performance of the AI model before the experiment,[27] the AI-assisted reading in our study was a dynamic process in which clinicians evaluated the performance of the AI model in each diagnosis alongside their own feedback. Moreover, the participating hospitals covered the southwestern and northwestern regions of China, indicating that this AI model could be used in lower resource settings to improve CTA-based aneurysm diagnosis variability.

The third advantage of our AI model is its usability. When a new AI model is implemented in clinical practice, a variety of qualities are desired for the system to garner user trust.[28] The most obvious component of AI trustworthiness is its diagnostic accuracy because users are unlikely to trust an AI model that has not been rigorously shown to give correct predictions.[10] In our randomised open-label comparison study, no clinician fully declined to use the AI results; only 7·4% of results were made without AI assistance in the AI-assisted group. Since all clinicians were not told about the true diagnostic performance beforehand, our AI model garnered trust based on its own reliable diagnostic performance. Once again, our results confirmed that the AI-assisted diagnoses achieved significantly higher sensitivity than non-AI-assisted diagnoses, and the superiority of the AI-assisted

strategy did not decrease substantially when excluding actually non-AI-assisted diagnoses. We observed that these actually non-AI-assisted diagnoses were commonly reported in negative CTA cases and positive CTA cases involving large aneurysms (≥5 mm) which could be considered straightforward cases. Once clinicians make a clear diagnosis in these cases, they might no longer seek AI assistance. Broadly speaking, implementing our AI model could improve aneurysm diagnosis with high acceptance by all clinicians and, based on our research, it could also provide novel insights into the features used for aneurysm diagnosis, especially for small aneurysms.

The fourth advantage is that our AI model can be effectively implemented into real-world clinical practice and utilised as a second reader. To address the core concerns of medical AI implementation, members of the AI community (researchers, developers, guideline steering teams, etc) have constantly advocated that any AI model validation ought to be conducted in the context of the model's intended clinical pathway.[29] In real-world clinical practice, aneurysm diagnosis is one of the CTA interpretation tasks with relatively low incidence compared with vascular stenosis. Using the AI model as a second reader could supplement the radiologists' diagnosis and reduce inaccurate diagnoses. The real-world results showed our AI model could markedly improve the aneurysm detection rate by identifying occult small aneurysms. This is very helpful for the diagnosis of small aneurysms and aneurysms in uncommon anatomical locations. Moreover, our AI model has a high negative predictive value (0·998 [0·994–1·000]), indicating it could reliably exclude true negative patients in a routine clinical setting and reduce radiologists' workload. This is consistent with our previous results which showed a negative predictive value of 0·990.[16] It is particularly relevant to optimise the clinical management of intracranial aneurysms, especially in the emergent clinical setting and high-patient volume hospitals.

Our study has some limitations. First, our current AI model is CTA modality dependent and mainly focused on intracranial aneurysm detection. Non-enhanced CT images showing subarachnoid haemorrhage were not included in the deep-learning algorithm in all datasets. However, our model's performance was independent of subarachnoid haemorrhage presence. Second, the AI model cannot identify aneurysm rupture complications like vasospasm and hydrocephalus, which are relevant to clinical management. Third, although a 4-week washout period and randomised ordering of CTA cases was applied for the multi-reader multi-case study, recall bias cannot be completely eliminated. Fourth, not all patients in the training, internal validation, and prospective datasets had the gold-standard digital subtraction angiography examination. Fifth, although our AI model showed superior results to a meta-analysis,[22] a few CTA cases with poor image quality or that failed to be analysed by the AI model were excluded, potentially leading to an overestimation of the model's performance. Sixth, it is important to note that the clinical reference standard employed in this study does not align with the standard of care for aneurysm diagnosis. In real-world clinical practice of a radiology department in the Chinese medical system, the final radiological report would be re-audited by more experienced senior radiologists. Additionally, the diagnoses times in this study might be too fast due to readings taking place in a research environment rather than a clinical environment. This can be observed in the results by a focus on the Circle of Willis, the dominant anatomical location of intracranial aneurysms, and the aid of volume rendering images significantly enhancing visual detection; however, a comprehensive evaluation through layer-by-layer analysis of CTA images remains essential for diagnosis.[30] Consequently, the comparative performance of the AI model is probably overinflated. Lastly, in our prospective study, we did not explore the influence of this AI model on patient follow-up care. A well designed prospectively randomised controlled study is warranted to evaluate the downstream effects of the AI model on the clinical pathway of intracranial aneurysms.

In conclusion, our proposed AI system achieved an accurate, robust, and generalisable performance for intracranial aneurysm detection, which was independent of the patients' subarachnoid haemorrhage status. This stepwise evaluation study demonstrated the diagnostic performance of clinicians' across geographically different hospitals in China was substantially improved with AI assistance. Future research should focus on addressing the identified limitations and conducting well designed prospective studies to evaluate the broader impact of AI models on the clinical pathway of intracranial aneurysms.

China Aneurysm AI Project Group members

Bin Hu, Bin Tan, Chengwei Pan, Chunfeng Hu, Changsheng Zhou, Fandong Zhang, Fajin Lv, Feidi Liu, Fei Zhou, Feng Chen, Feng Xu, Ge Yan, Guangbin Cui, Guang Ming Lu, Hongmei Gu, Huiying Liang, Li Lu, Long Jiang Zhang, Meng Zhang, Mingli Hou, Qiu Sun, Rongpin Wang, Rui Xu, Rui Zuo, Sheng Li, Shumin Tao, Shu Zhang, Weiwei Chen, Xue Chai, Wulin Wang, Xiangming Fang, Xiao Luo, Xiaodong Wang, Xiaorong Li, Xiaoxia Chen, Yan Gu, Yizhou Yu, Yongjian Dai, Yubao Liu, Yueqin Chen, Yuning Pan, Zhao Shi, Zhen Zhou, Zhibo Hou, Zhongchang Miao.

Contributors

LJZ initiated the project, organised a collaborative team, and provided study supervision. BH, ZS, CH, and LJZ conceived the study and its design. BH, ZS, ZH, HL, YP, XC, XL, and FZh collected the data for the model's training. BH, ZS, HW, ZZ, FDZ, SZ, CP, and YZY developed the network architectures, training, and validation setup. BH, LL, and HW collected the data for the model's external validation. LJZ, CH, XW, ZM, XF, RW, FX, XL, SL, GY, QS, MZ, GC, FJL, and YL organised the clinician team and supervised the implementation of quality control measures in stages 1–3. CH, ZM, XF, XW, RW, and LJZ organised the clinician team and supervised the implementation of quality control measures in stage 4. BH, ZS, LL, and HW co-wrote the manuscript. LJZ, UJS, AV-S, WM, and YY critically revised the manuscript, and all authors discussed the results and provided feedback regarding the manuscript. All authors had final responsibility for the decision to

submit for publication. BH, ZS, and LJZ verified the data and all authors had access to the data included in this study.

Declaration of interests

UJS receives research grants or support from Astellas, Bayer, Bracco, Elucid BioImaging, GE, Guerbet, Heartflow, Keya Medical, and Siemens; personal fees from Bayer and Guerbet; consultancy fees from Elucid BioImaging, GE, Guerbet, and Keya Medical; and speaker honorarium from HeartFlow and Siemens Healthineers. AV-S receives research grants from Siemens and is a consultant for Elucid Bioimaging. HW, ZZ, FDZ, and SZ are employees of the AI laboratory department at Deepwise. All other coauthors declare no competing interests.

Data sharing

The complete code for the deep-learning model training and validation is available at https://github.com/deepwise-code/CAIA. The datasets and annotations to develop this model are not publicly available due to our China Aneurysm AI Project (China Intracranial Aneurysm Artificial Intelligence Project) Group regulation restrictions. The de-identified two external datasets with digital subtraction angiography-verified results and the prospective validation dataset will be shared upon request to the corresponding author (LJZ). After approval by the China Aneurysm AI Project Group, these data could be shared through a secure online platform only for research purposes. To gain access to the data, requestors must sign a data access agreement and provide a methodologically viable proposal, and perform an analysis that achieves the proposal's aims.

Acknowledgments

This work was supported by the Key Projects of the National Natural Science Foundation of China (grant numbers 81830057 and 82230068), the Youth Fund of the National Natural Science Foundation of China (grant number 82102155), and the General Program of the National Natural Science Foundation of China (grant number 62076076). We sincerely thank all of the hospitals for providing CTA data and the participating radiologists and interventional neurosurgeons who generously donated their time and efforts for the reader studies.

References

1 Li MH, Chen SW, Li YD, et al. Prevalence of unruptured cerebral aneurysms in Chinese adults aged 35 to 75 years: a cross-sectional study. *Ann Intern Med* 2013; **159:** 514–21.
2 Vlak MHM, Algra A, Brandenburg R, Rinkel GJE. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. *Lancet Neurol* 2011; **10:** 626–36.
3 Claassen J, Park S. Spontaneous subarachnoid haemorrhage. *Lancet* 2022; **400:** 846–62.
4 van Gijn J, Kerr RS, Rinkel GJE. Subarachnoid haemorrhage. *Lancet* 2007; **369:** 306–18.
5 Korja M, Silventoinen K, Laatikainen T, Jousilahti P, Salomaa V, Kaprio J. Cause-specific mortality of 1-year survivors of subarachnoid hemorrhage. *Neurology* 2013; **80:** 481–86.
6 Hemphill JC 3rd, Greenberg SM, Anderson CS, et al. Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2015; **46:** 2032–60.
7 Thompson BG, Brown RD Jr, Amin-Hanjani S, et al. Guidelines for the management of patients with unruptured intracranial aneurysms: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2015; **46:** 2368–400.
8 Shi Z, Hu B, Schoepf UJ, et al. Artificial intelligence in the management of intracranial aneurysms: current status and future perspectives. *AJNR Am J Neuroradiol* 2020; **41:** 373–79.
9 Philipp LR, McCracken DJ, McCracken CE, et al. Comparison between CTA and digital subtraction angiography in the diagnosis of ruptured aneurysms. *Neurosurgery* 2017; **80:** 769–77.
10 Lian K, Bharatha A, Aviv RI, Symons SP. Interpretation errors in CT angiography of the head and neck and the benefit of double reading. *AJNR Am J Neuroradiol* 2011; **32:** 2132–35.
11 Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022; **28:** 31–38.
12 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; **18:** 500–10.
13 Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 2022; **28:** 924–33.
14 Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov* 2020; **6:** 45–47.
15 Park A, Chute C, Rajpurkar P, et al. Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA Netw Open* 2019; **2:** e195600.
16 Shi Z, Miao C, Schoepf UJ, et al. A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. *Nat Commun* 2020; **11:** 6090.
17 Yang J, Xie M, Hu C, et al. Deep learning for detecting cerebral aneurysms with CT angiography. *Radiology* 2021; **298:** 155–63.
18 Luo Z, Wang D, Sun X, et al. Comparison of the accuracy of subtraction CT angiography performed on 320-detector row volume CT with conventional CT angiography for diagnosis of intracranial aneurysms. *Eur J Radiol* 2012; **81:** 118–22.
19 Obuchowski NA, Bullen J. Multireader diagnostic accuracy imaging studies: fundamentals of design and analysis. *Radiology* 2022; **303:** 26–34.
20 Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol* 2008; **15:** 647–61.
21 Chakraborty DP. A brief history of free-response receiver operating characteristic paradigm data analysis. *Acad Radiol* 2013; **20:** 915–19.
22 Din M, Agarwal S, Grzeda M, Wood DA, Modat M, Booth TC. Detection of cerebral aneurysms using artificial intelligence: a systematic review and meta-analysis. *J Neurointerv Surg* 2023; **15:** 262–71.
23 Kallmes DF, Erickson BJ. Automated aneurysm detection: emerging from the shallow end of the deep learning pool. *Radiology* 2021; **298:** 164–65.
24 Shi Z, Zhang LJ. Three fundamental elements for deep learning-based computer-assisted diagnostic tools of intracranial aneurysms. *Radiology* 2021; **300:** E311.
25 Oakden-Rayner L, Gale W, Bonham TA, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet Digit Health* 2022; **4:** e351–58.
26 Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Health* 2021; **3:** e496–506.
27 Peng S, Liu Y, Lv W, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *Lancet Digit Health* 2021; **3:** e250–59.
28 Cutillo CM, Sharma KR, Foschini L, et al. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med* 2020; **3:** 47.
29 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; **1:** e271–97.
30 Brown RD Jr, Broderick JP. Unruptured intracranial aneurysms: epidemiology, natural history, management options, and familial screening. *Lancet Neurol* 2014; **13:** 393–404.