Journal of Neuroradiology 52 (2025) 101388

Available online at
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

Original Article

AI-assisted detection of cerebral aneurysms on 3D time-of-flight MR angiography: User variability and clinical implications

Liang Liao [a,b,*], Ulysse Puel [a,c], Ophélie Sabardu [a], Oana Harsan [a], Luana Lopes De Medeiros [a], Wassim Abou Loukoul [a], René Anxionnat [a,c], Erwan Kerrien [b]

[a] Department of Diagnostic and Interventional Neuroradiology, CHRU Nancy, France
[b] INRIA, LORIA, CNRS, Université de Lorraine, Nancy, France
[c] IADI, INSERM U1254, Université de Lorraine, Nancy, France

ARTICLE INFO

Keywords:
Artificial intelligence
Magnetic resonance angiography
Cerebral aneurysm detection
AI assistance
User variability
Physician expertise

ABSTRACT

Background: The generalizability and reproducibility of AI-assisted detection for cerebral aneurysms on 3D time-of-flight MR angiography remain unclear. We aimed to evaluate physician performance using AI assistance, focusing on inter- and intra-user variability, identifying factors influencing performance and clinical implications.

Methods: In this retrospective study, four state-of-the-art AI models were hyperparameter-optimized on an in-house dataset (2019−2021) and evaluated via 5-fold cross-validation on a public external dataset. The two best-performing models were selected for evaluation on an expert-revised external dataset. Inclusion: saccular aneurysms without prior treatment. Five physicians, grouped by expertise, each performed two AI-assisted evaluations, one with each model. Lesion-wise sensitivity and false positives per case (FPs/case) were calculated for each physician−AI pair and AI models alone. Agreement was assessed using kappa. Aneurysm size comparisons used the Mann−Whitney U test.

Results: The in-house dataset included 132 patients with 206 aneurysms (mean size: 4.0 mm); the revised external dataset, 270 patients with 174 aneurysms (mean size: 3.7 mm). Standalone AI achieved 86.8 % sensitivity and 0.58 FPs/case. With AI assistance, non-experts achieved 72.1 % sensitivity and 0.037 FPs/case; experts, 88.6 % and 0.076 FPs/case; the intermediate-level physician, 78.5 % and 0.037 FPs/case. Intra-group agreement was 80 % for non-experts (kappa: 0.57, 95 % CI: 0.54−0.59) and 77.7 % for experts (kappa: 0.53, 95 % CI: 0.51−0.55). In experts, false positives were smaller than true positives (2.7 vs. 3.8 mm, $p < 0.001$); no difference in non-experts ($p = 0.09$). Missed aneurysm locations were mainly model-dependent, while true- and false-positive locations reflected physician expertise. Non-experts more often rejected AI suggestions and added fewer annotations; experts were more conservative and added more.

Conclusion: Evaluating AI models in isolation provides an incomplete view of their clinical applicability. Detection performance and patterns differ between standalone AI and AI-assisted use, and are modulated by physician expertise. Rigorous external validation is essential before clinical deployment.

Introduction

Cerebral aneurysm rupture accounts for 85 % of non-traumatic subarachnoid hemorrhages, with mortality rates reaching up to 44 % and 20 % of survivors experiencing permanent disabilities. With a 3.2 % prevalence in the general population, detecting unruptured intracranial aneurysms (UIAs) has become a critical clinical priority.[1,2] Although digital subtraction angiography remains the gold standard for aneurysm assessment, three-dimensional time-of-flight magnetic resonance angiography (3D TOF-MRA) is now the preferred noninvasive method for UIA detection and screening.[3] However, the rising number of annual scans strains radiology resources, contributing to physician fatigue and potential diagnostic errors.[4]

Automated aneurysm detection using convolutional neural networks —a type of deep learning within artificial intelligence (AI)—has shown promise in improving diagnostic performance on 3D TOF-MRA.[4–7]

Abbreviations: UIA, unruptured intracranial aneurysm; 3D TOF-MRA, three-dimensional time-of-flight magnetic resonance angiography; AI, artificial intelligence; TP, true positive; FN, false negative; FP, false positive; FPs/case, false positives per case; ACom, anterior communicating artery; MCA, middle cerebral artery

*Corresponding author at: Service de Neuroradiologie Diagnostique et Thérapeutique, CHRU Nancy, Hôpital Central, 29 Avenue du Maréchal de Lattre de Tassigny, 54035 Nancy, France.
E-mail address: l.liao@chru-nancy.fr (L. Liao).

https://doi.org/10.1016/j.neurad.2025.101388
Available online 10 September 2025

Recent studies demonstrate enhanced sensitivity when deep learning tools assist physicians, though most of this research has focused on CT angiography.[8–10] Data on physicians using AI models for UIA detection on 3D TOF-MRA remain limited.

Several critical gaps persist. Methodologically, concerns include the lack of multicenter and public datasets, as well as insufficient data on how different models perform on the same dataset—raising questions about generalizability and reproducibility. In terms of interpretation, variability in physician performance with AI tools, the influence of anatomical features on detection, and the overall clinical relevance of AI-assisted detection require further investigation. It also remains unclear whether achieving optimal performance with the use of AI tools requires expert-level neuroradiology training, or if non-expert physicians—particularly in peripheral centers—can attain comparable results using the same tools. Finally, how AI-assisted detection shapes physician behavior during interpretation is not well understood.

In this study, we aim to address these gaps by evaluating physician performance when using AI assistance. We examine inter-user and intra-user variability, the impact of anatomical features on detection, and the influence of physician expertise when using AI tools. We leverage an external public dataset with a well-defined reference standard and evaluate two different convolutional neural network models on the same dataset.

Materials and methods

Our institutional review board approved the study. Informed consent for the anonymous use of clinical imaging data was obtained through institutional general consent procedures, with no patients opting out.

Datasets and annotation

Two 3D TOF-MRA aneurysm datasets were used. The first comprised examinations collected consecutively at our institution (2019–2021) based on the following criteria: unruptured saccular aneurysms <20 mm, no prior treatment, and one examination per patient. An expert interventional neuroradiologist (14 years of experience, LL) annotated the in-house dataset. All images were acquired on a 3.0 T scanner (GE Healthcare, USA), with detailed sequence parameters reported previously.[11]

The second dataset was an external public dataset from Lausanne [12], collected on multi-vendor, multi-protocol MRI systems to reflect real-world clinical scenarios. The original annotation was verified by a senior neuroradiologist (>15 years of experience), external to our team. Scans were acquired on 1.5 T or 3.0 T MRI systems from Philips and Siemens; acquisition details are provided in Di Noto et al.[12] In both datasets, aneurysms were annotated as spheres.

Our in-house dataset and the external dataset (with original annotation) overlapped with our previous methodological study focused on technical development and evaluation of an object detection-based model.[11] The present work expands the clinical scope, establishing a revised reference standard for the external dataset and evaluating physician performance with AI assistance, focusing on user variability and expertise-related differences. Analyses used non-overlapping training and test sets, ensuring no data leakage.

Model selection and hyperparameter optimization

We selected four state-of-the-art AI models[13]: nnU-Net,[14] based on a segmentation approach, and nnDetection,[15] SCPM-Net,[16] and the Assis model,[11] based on object detection. The latter two required hyperparameter optimization. Hyperparameter options were compared using a 5-fold cross-validation strategy on the in-house dataset: the dataset was divided into five subsets—four used for training and one for performance estimation—iteratively, resulting in five assessments for statistical comparison of the hyperparameter options. The optimal settings—initial learning rate of 0.01, batch size of 32, and 200 epochs—were selected based on sensitivity and false-positive rate. Full methodological details are provided in our previous work.[11]

To assess model performance and the robustness of the original annotation in the external dataset, we applied automatic detection to the external dataset while maintaining the same hyperparameters, again following a 5-fold cross-validation approach, with each model trained from random initialization. In each fold, the model was trained on 80 % of cases and tested only on the held-out 20 %, producing out-of-fold predictions. For each model, these out-of-fold predictions were aggregated across the five folds to compute performance metrics. Based on mean lesion-wise sensitivity and false positives per case across folds, the two best-performing models, nnDetection and Assis, were selected for further analysis.

Reference standard establishment

A review of AI-generated annotations on the external dataset revealed numerous previously unlabeled aneurysms, raising concerns about the original annotation quality. To address this, we established a more robust reference standard for the external dataset. Two interventional neuroradiology experts (LL and RA, with 14 and 32 years of experience) independently conducted a blinded review of all AI-generated annotations, including false negatives, using 3D Slicer software.[17] Disagreements were resolved by consensus. This revised annotation served as the reference standard. The two selected AI models were then evaluated against this reference standard by re-scoring their out-of-fold detections; no retraining was performed. This produced two final AI-generated annotation sets for standalone evaluation. In the next phase of the study, the annotation tool used by physicians displayed 3D TOF-MRA images overlaid with the AI-generated annotations from either model, serving as initial proposed detections.

Additionally, we annotated the anatomical location of each aneurysm in the reference standard, categorized by theoretical rupture risk based on the literature.[18,19] Locations included the anterior communicating artery (ACom), cavernous segment, ophthalmic segment, supraophthalmic segment (posterior communicating artery, anterior choroidal artery, carotid termination), middle cerebral artery (MCA), pericallosal, and posterior circulation (basilar tip, posterior inferior cerebellar artery). Furthermore, we documented morphological features such as vascular loops, adjacent perforators, and atheromatous or irregular parent arteries.

Evaluation of physician performance in AI-assisted aneurysm detection

Five physicians participated in the study, grouped by expertise level: two non-expert radiologists (4 and 5 years of experience in general radiology), two expert physicians (an interventional neuroradiologist and a vascular neurosurgeon, each with 8 years of experience), and one intermediate-level physician (a general radiologist training in interventional neuroradiology, with 6 years of total experience).

A senior interventional neuroradiologist (LL) trained all participants on the 3D Slicer annotation interface. Physicians could validate or reject AI-generated annotations and annotate any additional aneurysms they believed the AI had missed. The final detection decision was physician-driven, overriding AI-generated annotations; we define this outcome as AI-assisted annotation. Additionally, physicians reported aneurysm location and relevant morphological features.

Each physician evaluated the external dataset twice with AI assistance—once using AI-generated annotations from nnDetection and once from the Assis model—with a 3-week washout period between evaluations. This resulted in ten AI-assisted annotation sets, in addition to the two AI-generated annotation sets, each compared against the reference standard to assess detection performance.

Code and annotations availability

The code developed and used in this study, the reference-standard annotations, our custom 3D Slicer annotation plug-in, and the evaluation scripts (including fold assignments) are publicly available at https://gitlab.inria.fr/yassis/DeepAneDet.

Statistical analysis and interpretation of detection errors

Predicted spheres were matched one-to-one with reference standard spheres if their intersection over union exceeded 10 %, a commonly used threshold for volumetric object detection tasks in the literature.[15] In cases of multiple overlaps, the pair with the highest score was retained. Matched pairs were classified as true positives (TP). Unmatched predictions were false positives (FP), and unmatched reference spheres were false negatives (FN). A detailed review of all FP and FN cases was conducted for each user to identify anatomical factors contributing to errors.

Lesion-wise sensitivity and false positives per case (FPs/case) were calculated for each AI-assisted annotation and each AI-generated annotation. Agreement was assessed using Cohen's kappa (pairwise) and Fleiss' kappa (multi-rater). To compare agreement levels between physician groups, a jackknife procedure was applied by systematically excluding each aneurysm, recalculating kappa, and estimating variance for a Z-test comparison.

Aneurysm size comparisons between FN, FP, and TP detections were performed using the Mann−Whitney U test. At the physician group level, each aneurysm was assigned a single label across both models, ensuring independent statistical comparisons.

All analyses were conducted using R version 4.4.2, with p-values < 0.05 considered statistically significant.

Results

Patient and aneurysm characteristics

Our in-house dataset included 132 patients (mean age: 56 ± 12 years [SD]; 75 females, 57 males) with 206 aneurysms (mean size: 4.0 ± 2.3 mm [SD]).

The external dataset initially comprised 296 examinations. After applying the same inclusion criteria as the in-house dataset, 270 patients remained for analysis (mean age: 52 ± 14 years [SD]; 159 females, 111 males), with 164 aneurysms. Annotation revision, with disagreements in 7.7 % of cases resolved by consensus, identified 23 previously unlabeled aneurysms and reclassified 13 annotated cases as vessel surface irregularities, increasing the total number of aneurysms from 164 to 174 (mean size: 3.7 ± 2.2 mm [SD]). 140 patients had aneurysms, and 130 were healthy individuals.

Aneurysm location details are provided in Table 1, and the flow diagram of dataset processing, model selection, and evaluation framework is shown in Fig. 1.

Table 1
Dataset characteristics.

Characteristic	In-house dataset	Revised external dataset
No. of examinations	132	270
No. of male patients	57	111
No. of female patients	75	159
Mean age (years)	56 ± 12	52 ± 14
No. of aneurysms	206	174
Mean aneurysm size (mm)	4.0 ± 2.3	3.7 ± 2.2
Aneurysm size range (mm)	1.2−19.6	1.2−18.5
No. of aneurysms by location		
Middle cerebral artery	54 (26.2)	51 (29.3)
Anterior communicating artery	34 (16.5)	35 (20.1)
Ophthalmic segment	28 (13.6)	29 (16.7)
Cavernous segment	26 (12.6)	20 (11.5)
Supra-ophthalmic segment	34 (16.5)	20 (11.5)
Pericallosal	14 (6.8)	10 (5.8)
Posterior circulation	16 (7.8)	9 (5.2)
No. of examinations by magnetic field strength		
1.5 T	0	51 (18.9)
3.0 T	132 (100)	219 (81.1)

Note. Numbers in parentheses indicate percentages. Data are presented as mean ± standard deviation.

Performance and agreement of AI models and physician−AI pairs

The pooled standalone performance of both AI models demonstrated a sensitivity of 86.8 % with 0.58 FPs/case. With AI assistance, non-experts achieved a sensitivity of 72.1 % with 0.037 FPs/case, experts reached 88.6 % with 0.076 FPs/case, and the intermediate-level physician attained 78.5 % with 0.037 FPs/case. Full details are provided in Table 2.

Agreement analysis showed an inter-AI model agreement of 39.4 % between nnDetection and the Assis model, with a Cohen's kappa of −0.37 (95 % CI: −0.44 to −0.30; $p < 0.001$). Excluding FP cases, agreement increased to 86.2 %, with a Cohen's kappa of 0.42 (95 % CI: 0.24 −0.60; $p < 0.001$). Intra-user agreements, comparing the results of the same user with both models, are summarized in Table 3. Intra-group agreement was 80 % in the non-expert group (Fleiss' kappa: 0.57, 95 % CI: 0.54−0.59; $p < 0.001$) and 77.7 % in the expert group (Fleiss' kappa: 0.53, 95 % CI: 0.51−0.55; $p < 0.001$), indicating moderate agreement and supporting homogeneous expertise within each group. The difference in kappa values (0.04) had a jackknife-estimated standard error of 0.047 ($Z = -0.74$, $p = 0.46$), showing no statistically significant difference in agreement levels between the two groups.

Influence of aneurysm characteristics on detection

In the pooled standalone detection results of both AI models, FN aneurysms had a significantly smaller mean diameter than TP aneurysms (2.5 ± 1.4 mm vs. 3.7 ± 2.0 mm, $p < 0.001$). This size difference was also present when AI models were used by physicians. Among non-experts, FNs were smaller than TPs (2.9 ± 2.0 mm vs. 3.9 ± 2.1 mm, $p < 0.001$), and a similar difference was observed in experts (2.6 ± 2.3 mm vs. 3.8 ± 2.1 mm, $p < 0.001$).

Regarding FP aneurysms, in the expert group, FPs were significantly smaller than TPs (2.7 ± 1.2 mm vs. 3.8 ± 2.1 mm, $p < 0.001$). However, in non-experts, no significant difference was observed (3.2 ± 2.7 mm vs. 3.9 ± 2.1 mm, $p = 0.09$). These detection error patterns across expertise levels are illustrated in Fig. 2.

The most frequent FN locations varied between AI models in standalone detection. For the nnDetection model, FNs were most common in the cavernous segment (25 %, 5/20) and MCA (13.7 %, 7/51), whereas for the Assis model, they were in the ophthalmic (20.7 %, 6/29) and cavernous segments (20 %, 4/20). These patterns remained consistent across all expertise levels when the models were used by physicians.

The most frequent TP locations also differed between models. For nnDetection, TPs were most common in the supra-ophthalmic (100 %, 20/20) and ophthalmic segments (89.7 %, 26/29), while for the Assis model, they were in the posterior circulation (100 %, 9/9) and supra-ophthalmic segment (95 %, 19/20). However, when used by physicians, TP locations varied by expertise level regardless of the AI model: in non-experts and the intermediate-level physician, TPs were most frequent in the ACom and supra-ophthalmic segment, whereas in experts, they were in the supra-ophthalmic and pericallosal locations. FP locations were also expertise-dependent. In non-experts, they were most common in the ACom and supra-ophthalmic segment, while in experts, they were in the ACom and posterior circulation. Details of these findings are summarized in Table 4.

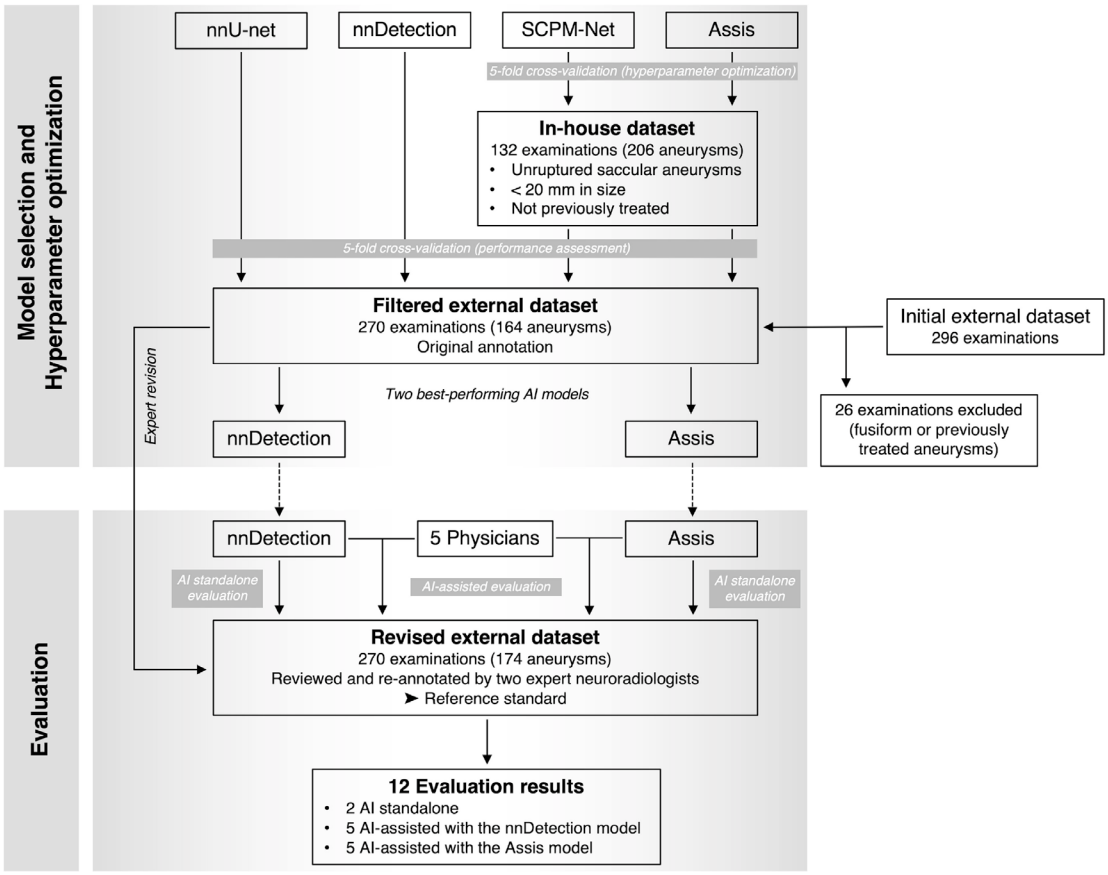

Fig. 1. Flow diagram of dataset processing, model selection, and evaluation framework.

Table 2
Performance of AI models (standalone) and physician−AI pairs.

Reader & Model	Sensitivity	FPs/case
AI standalone performance	86.8 (302/348)	0.58 (313/540)
nnDetection model	86.8 (151/174)	0.63 (170/270)
Assis model	86.8 (151/174)	0.53 (143/270)
Non-expert physicians with AI assistance	72.1 (502/696)	0.037 (40/1080)
Radiologist 1 with nnDetection	66.7 (116/174)	0.037 (10/270)
Radiologist 1 with Assis model	68.9 (120/174)	0.029 (8/270)
Radiologist 2 with nnDetection	74.1 (129/174)	0.048 (13/270)
Radiologist 2 with Assis model	78.7 (137/174)	0.033 (9/270)
Expert physicians with AI assistance	88.6 (617/696)	0.076 (82/1080)
Interventional neuroradiologist with nnDetection	87.9 (153/174)	0.089 (24/270)
Interventional neuroradiologist with Assis model	90.8 (158/174)	0.089 (24/270)
Vascular neurosurgeon with nnDetection	86.2 (150/174)	0.063 (17/270)
Vascular neurosurgeon with Assis model	89.1 (155/174)	0.063 (17/270)
Intermediate-level radiologist with AI assistance	78.5 (273/348)	0.037 (20/540)
Intermediate-level radiologist with nnDetection	77.6 (135/174)	0.048 (13/270)
Intermediate-level radiologist with Assis model	79.3 (138/174)	0.026 (7/270)

Note. Sensitivity is expressed as a percentage. Numbers in parentheses indicate the number of aneurysms or cases. Values are pooled across physicians for expertise-level groups and across AI models for the intermediate-level radiologist. FPs/case = false positives per case.

Physician interpretation and decision-making

Among aneurysms correctly detected by AI models but incorrectly rejected by physicians, FN cases accounted for 19.9 % (120/604) in non-experts, 4.8 % (29/604) in experts, and 11.3 % (34/302) in the intermediate-level physician. In non-experts, the most frequent cause was misinterpretation of the aneurysm's relationship to the curvature of MCA dividing branches (20.8 %, 25/120) (Fig. 3), while in experts, it was irregularity of the cavernous segment due to atheroma or dysplasia (24 %, 7/29). For FNs neither detected by AI nor added by physicians, the most frequently missed location was pericallosal in non-experts (20 %, 8/40) and the intermediate-level physician (25 %, 5/20), while for experts, it was the cavernous segment (20 %, 16/80).

The total number of annotations added by physicians was 25 for non-experts, 74 for experts, and 6 for the intermediate-level physician. Among these, FPs accounted for 48 % (12/25) in non-experts and 48.6 % (36/74) in experts. In experts, the most frequent location for incorrectly added FPs was the ACom (12.2 %, 9/74). Conversely, the most frequent location for correctly added TPs was the pericallosal artery (20 % of all pericallosal aneurysms, 8/40)—which

Table 3
Agreement analysis.

Comparison	Agreement	Kappa	95 % CI	*p*-value
Inter-AI model agreement				
nnDetection vs Assis model	39.4	−0.37	−0.44 to −0.30	< 0.001
nnDetection vs Assis model (excluding FPs)	86.2	0.42	0.24–0.60	< 0.001
nnDetection vs Assis model (FPs only)	10.6	−0.79	−0.86 to −0.72	< 0.001
Intra-user agreement (nnDetection vs Assis model)				
Radiologist 1	80.9	0.61	0.52–0.71	< 0.001
Radiologist 2	79.4	0.58	0.49–0.68	< 0.001
Interventional neuroradiologist	80.6	0.55	0.44–0.66	< 0.001
Vascular neurosurgeon	78.7	0.52	0.41–0.63	< 0.001
Intermediate-level radiologist	81.3	0.62	0.53–0.72	< 0.001
Intra-group agreement				
Non-expert group	80.0	0.57	0.54–0.59	< 0.001
Expert group	77.7	0.53	0.51–0.55	< 0.001

Note. Agreement is expressed as a percentage. Cohen's kappa was used for pairwise comparisons, and Fleiss' kappa for multi-rater agreement. Intra-user agreement compares the results of each physician using nnDetection and Assis models. Abbreviations: 95 % CI = 95 % confidence interval, FPs = false positives.

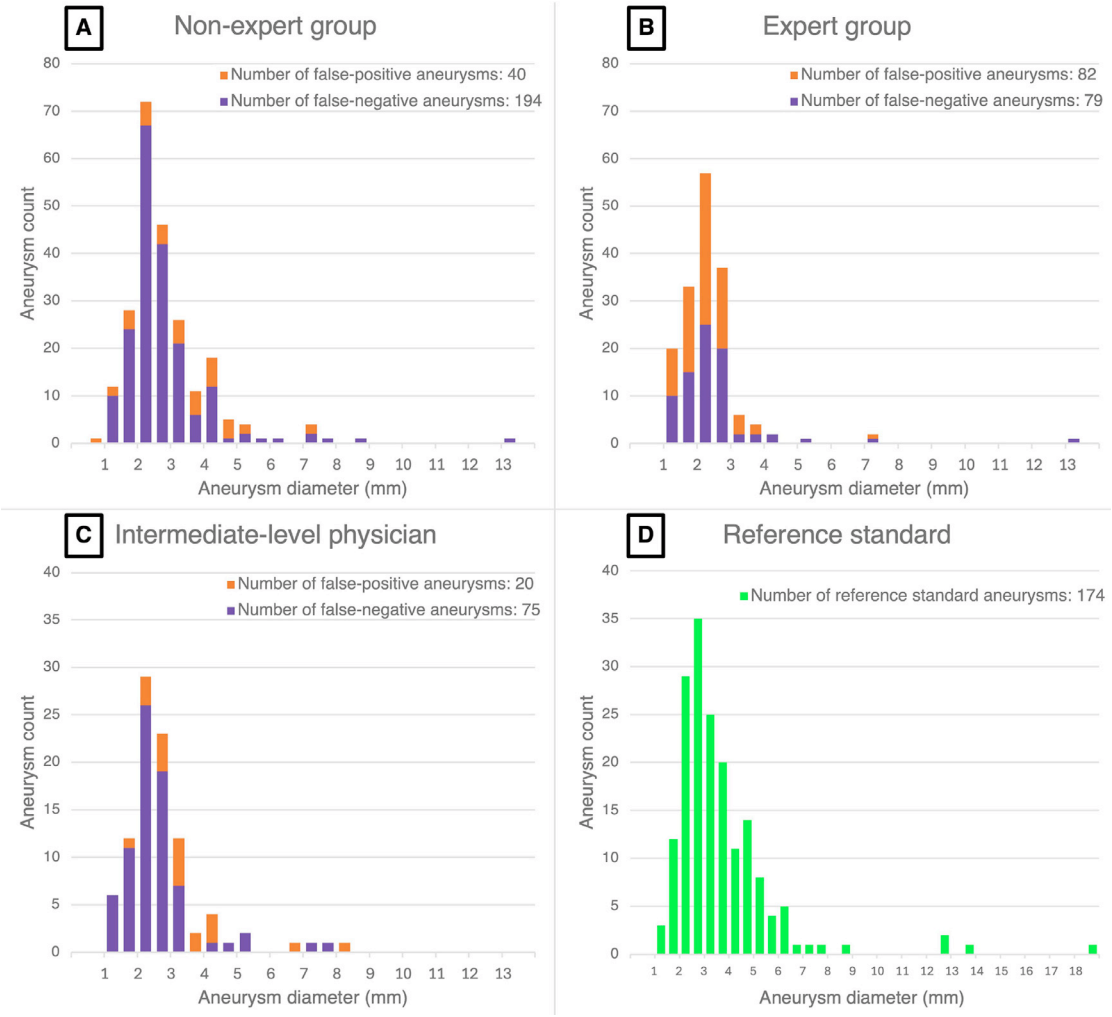

Fig. 2. Stacked histogram of detection errors (false positives and false negatives) by aneurysm size across expertise levels using AI-assistance. **(A)** and **(B)** show results for the non-expert and expert groups, respectively. The aneurysm count in each group represents the pooled total from four evaluations, performed by two physicians using both the nnDetection and Assis models. **(C)** shows results for the intermediate-level physician, where the aneurysm count is pooled from two evaluations, one per AI model. **(D)** shows the size distribution of all reference standard aneurysms in the revised external dataset used for evaluation.

was also the most frequently missed location by non-experts when AI failed to detect them (Fig. 4). The likelihood of physicians compensating for an aneurysm missed by AI by correctly adding a new detection was 14.1 % (13/92) for non-experts and 41.3 % (38/92) for experts.

For FPs detected by AI and accepted by physicians, non-experts most often misidentified branch origins or infundibula as aneurysms in the ophthalmic or supra-ophthalmic carotid segment (27.6 %, 8/29), while experts most commonly misclassified dysplasia at the A1–A2 junction (17.4 %, 8/46).

Table 4
Aneurysm size and common locations in detection outcomes.

Reader & Model	FN size	FN locations (Top 2)	FP size	FP locations (Top 2)	TP size	TP locations (Top 2)
AI standalone						
nnDetection	2.4 ± 1.2	Cavernous, 25 % (5/20) MCA, 13.7 % (7/51)	3.3 ± 2.5	...	3.8 ± 2.2	Supra-oph, 100 % (20/20) Oph, 89.7 % (26/29)
Assis model	2.5 ± 1.5	Oph, 20.7 % (6/29) Cavernous, 20 % (4/20)	3.9 ± 2.5	...	3.6 ± 1.7	Posterior, 100 % (9/9) Supra-oph, 95 % (19/20)
With AI assistance						
NExp + nnDet	2.9 ± 1.6	MCA, 35.3 % (36/102) Cavernous, 30 % (12/40)	3.1 ± 3.4	ACom (6) Supra-oph (6)	4.0 ± 2.3	ACom, 78.6 % (55/70) Supra-oph, 75 % (30/40)
NExp + AM	2.8 ± 2.5	Cavernous, 42.5 % (17/40) Oph, 31 % (18/58)	3.4 ± 0.9	ACom (4) Supra-oph (4)	3.9 ± 1.9	ACom, 92.9 % (65/70) Supra-oph 77.5 % (31/40)
Exp + nnDet	2.6 ± 1.9	Cavernous, 25 % (10/40) MCA, 15.7 % (16/102)	2.5 ± 1.0	ACom (14) Posterior (10)	3.8 ± 2.2	Supra-oph, 100 % (40/40) Pericallosal, 100 % (18/18)
Exp + AM	2.6 ± 2.9	Cavernous, 25 % (10/40) Oph, 13.8 % (8/58)	2.9 ± 1.4	ACom (10) Posterior (8)	3.8 ± 2.0	Supra-oph, 97.5 % (39/40) Pericallosal, 94.4 % (17/18)
IntRad + nnDet	2.5 ± 1.0	Cavernous, 40 % (8/20) MCA, 21.6 % (11/51)	3.2 ± 1.6	ACom (8) Supra-oph (3)	4.0 ± 2.3	ACom, 85.7 % (30/35) Supra-oph, 80 % (16/20)
IntRad + AM	2.6 ± 1.2	Cavernous, 35 % (7/20) Oph, 27.6 % (8/29)	3.6 ± 1.3	ACom (2) Supra-oph (2)	3.9 ± 2.3	ACom, 91.4 % (32/35) Supra-oph, 90 % (18/20)

Note. Aneurysm size (mm) is reported as mean ± standard deviation. The two most frequent locations are listed for false negatives (FN), false positives (FP), and true positives (TP), with percentages indicating their proportion within each category. Numbers in parentheses indicate the absolute number of aneurysms. Values are pooled across physicians for expertise-level groups. For AI standalone detection outcomes, false-positive locations are not listed, because they can occur anywhere, including outside the vascular tree. Abbreviations: NExp = Non-experts, Exp = Experts, IntRad = Intermediate-level radiologist, nnDet = nnDetection model, AM = Assis model, MCA = Middle cerebral artery, Oph = Ophthalmic, ACom = Anterior communicating artery, Supra-oph = Supra-ophthalmic.

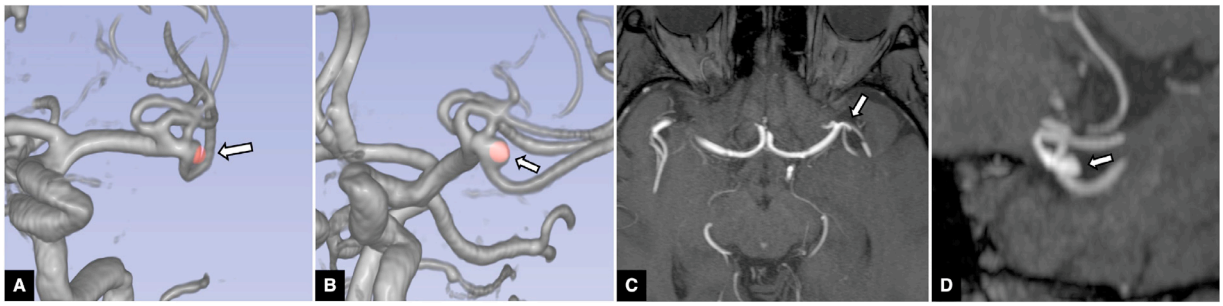

Fig. 3. Example of a 3.1 mm false-negative aneurysm at the division of the middle cerebral artery, detected by AI but rejected by one of the non-expert radiologists. **(A)** Volume-rendered 3D TOF-MRA shows the aneurysm flagged by AI (arrow). The anteroposterior view only partially reveals the aneurysm's long axis and overlaps with the lower division branch, potentially misleading an inexperienced physician. **(B)** Correct oblique view exposes the aneurysm neck and its full long axis (arrow). **(C)** Axial 2D maximum intensity projection shows the aneurysm (arrow), which is difficult to distinguish from adjacent branches. **(D)** Oblique coronal reformation clearly delineates the aneurysm (arrow).

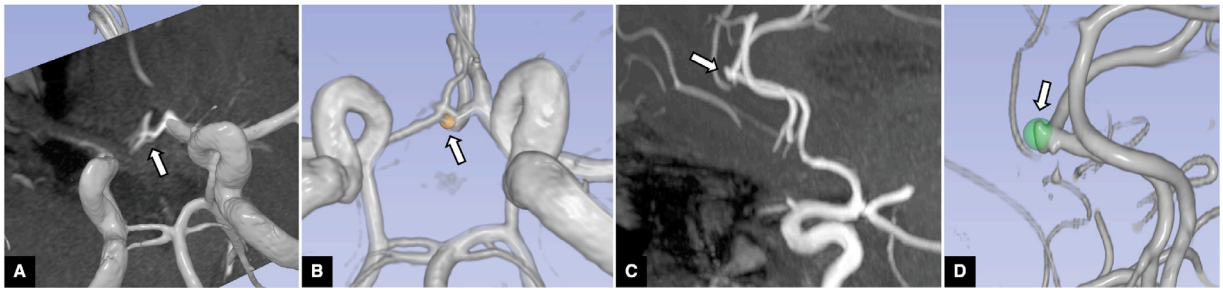

Fig. 4. Examples of aneurysms added by expert physicians when they believe AI missed them. **(A, B)** A 2 mm false-positive aneurysm is incorrectly added at the anterior communicating artery. **(A)** Volume-rendered 3D TOF-MRA with an embedded axial cross-sectional plane reveals a fenestration of the anterior communicating artery complex (arrow). **(B)** Annotation is placed on the smaller channel of the fenestration, which the physician mistakenly identifies as an aneurysm (arrow). **(C, D)** A 2.5 mm true-positive pericallosal aneurysm is correctly added by the expert. **(C)** Sagittal 2D maximum intensity projection highlights the small aneurysm missed by AI (arrow). **(D)** Volume-rendered 3D TOF-MRA shows the correctly placed annotation of this aneurysm (arrow).

Discussion

The generalizability and reproducibility of AI models for UIA detection on 3D TOF-MRA—whether used in standalone mode or as assistance to physicians—remain unclear, as does the influence of anatomical variables on detection performance and its clinical relevance. Our findings show that evaluating AI models in isolation provides an incomplete picture of their practical applicability. Detection performance and patterns differ significantly between standalone AI and AI-assisted use, and are further modulated by physician expertise.

We initiated this study with the premise—demonstrated in prior work by Sohn et al.[8]—that AI-assisted detection can improve sensitivity without substantially increasing FPs compared to unaided physician performance. Given this, we focused on user variability and the impact of physician expertise, rather than reevaluating unaided performance. This approach was further supported by a posteriori analysis of the external dataset's original annotation: the expert neuroradiologist from the external dataset's team (independent of our group), working without AI assistance, achieved 82.7 % sensitivity and 0.07 FPs/case when retrospectively compared to our revised reference standard.[12] Compared to the performance of our expert physician−AI pairs (88.6 % sensitivity, 0.076 FPs/case), sensitivity was higher with AI assistance, while the FP rate remained comparable.

To evaluate AI integration in UIA detection, we designed a structured framework incorporating several methodological strengths. We selected two state-of-the-art AI models with comparable performance to minimize bias and used an external public dataset with a robust, expert-revised reference standard. Physicians of varying expertise levels each performed two AI-assisted evaluations—one with each model—allowing us to examine how expertise and decision-making interact with AI support. Consistent agreement patterns within groups validated this categorization, enabling a focused analysis of the influence of expertise on detection outcomes.

A marked reduction in FPs/case (from 0.58 to 0.05) was observed when AI was used by physicians rather than in standalone mode, highlighting the human ability to filter out incorrect AI-generated detections and maintain diagnostic coherence. A similar trend was reported by Sohn et al.,[8] though with smaller magnitude (FPs/case from 0.12 to 0.06). However, in our study, AI-assisted detection by non-expert physicians resulted in lower sensitivity (72.1 %) compared with standalone AI (86.8 %), whereas pairing AI with expert physicians slightly improved sensitivity to 88.6 %. This discrepancy was not described in previous reports. In Sohn et al.'s study on 3D TOF-MRA,[8] AI-assisted detection by non-expert physicians slightly outperformed standalone AI (94.8 % vs 92.3 %). Several differences may account for this divergence. First, our evaluation used an external public dataset with heterogeneous scanners and protocols, while Sohn et al. relied on a single-center internal dataset, and their non-expert readers were from that same institution, which may have made AI suggestions easier to validate. Second, Sohn et al. excluded cases with annotator disagreement, all of which involved aneurysms <2 mm, effectively removing the most challenging small lesions. Third, their dataset was heavily imbalanced anatomically, with 63 % of aneurysms clustered around the dural ring, limiting location diversity and potentially inflating sensitivity compared with cohorts containing broader distributions. In a different modality, CT angiography, Park et al. reported that AI assistance brought a resident's performance very close to standalone AI (91 % vs 94 %).[9] This may reflect the higher spatial resolution of CT angiography, particularly advantageous for small lesions, whereas flow-related artifacts and slab boundary effects degrade effective resolution and detectability on 3D TOF-MRA.

Regarding aneurysm size, small aneurysms (<3 mm) remain challenging to detect, as previously reported,[4,20] and our results confirm that all physician−AI pairs—regardless of expertise—were more likely to miss them. Expert physicians, however, had greater difficulty deciding whether to confirm small aneurysms flagged by AI, resulting in more frequent FPs, while non-experts more often misclassified larger aneurysms. For aneurysm location, FN locations appeared to depend primarily on the AI model rather than physician expertise, likely reflecting the tendency of physicians to miss the same aneurysms that AI fails to detect (85.9 % for non-experts, 58.7 % for experts). In contrast, TP and FP locations were more influenced by physician expertise, with each group showing specific locations they were more inclined to confirm.

Non-experts more frequently rejected AI-proposed detections (19.9 % incorrectly discarded) and rarely added new annotations (only 25). Experts, by contrast, were more conservative, with fewer incorrect rejections (4.8 %) and more added annotations (74 in total). The intermediate-level physician demonstrated behavior between these two extremes (11.3 % incorrect rejections and 6 added annotations). Non-experts most often discarded aneurysms at the MCA division (20.8 %), where the curvature and superposition of branches may require reorientation of volume-rendered images and multiplanar review—skills they may be less familiar with. Experts more frequently rejected cavernous segment aneurysms (24 %), possibly because of their lower perceived risk. Among FPs in the expert group, the ACom was the most common site for added or accepted detections (29.3 %, 24/82), likely reflecting the higher perceived bleeding risk of small aneurysms in this location.[21]

These findings suggest a practical strategy for non-experts: AI-proposed detections should be rejected only when the physician is reasonably confident the findings are not aneurysms. If uncertainty remains, accepting an AI suggestion may be safer. Given their already low FP rate (0.037), such an approach is unlikely to lead to clinically significant increases in unnecessary follow-up. Moreover, since non-experts rarely add new annotations—and nearly half of those are FPs—focusing on reducing incorrectly discarded AI-proposed detections may offer more clinical benefit than seeking additional undetected aneurysms. This recommendation is less critical for experts, who already demonstrate strong, balanced performance when using AI support.

This study has several limitations. First, it was retrospective in design. Second, we did not assess interpretation time per examination, as this would be more meaningful when comparing unaided and AI-assisted performance by the same physician—beyond the scope of this study. Third, the external public dataset was of moderate size and included relatively few aneurysms in certain locations—particularly the pericallosal and posterior circulation—limiting more detailed statistical analysis of the influence of location on detection performance. Fourth, all five physicians who used the AI models were from a single institution, which may limit the generalizability of user behavior. Finally, only one intermediate-level physician participated; although this physician completed evaluations using both AI models, the findings may be less generalizable for this experience level.

Conclusion

For UIA detection on 3D TOF-MRA, AI model evaluation should incorporate physician use of AI assistance rather than rely solely on standalone AI performance, as detection outcomes differ significantly between the two approaches. Usage guidelines tailored to physician expertise are also warranted. These findings highlight the need for rigorous external validation—akin to medical device evaluation—prior to clinical deployment. Future work should focus on retraining models with the revised reference standard, with particular emphasis on anatomically challenging regions, and on prospective evaluation in real-world workflows to confirm clinical utility.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Human and animal right

The authors declare that the work described has not involved experimentation on humans or animals.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s) and/or volunteers

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Disclosure of funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

Acknowledgments

The authors would like to thank Mr. Youssef ASSIS for his contribution to the technical development of the object detection-based model described in our previous methodological study, which served as the foundation for the clinical evaluation conducted in this work.

References

1. 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*. 2022. https://doi.org/10.1136/jnis-2022-019456. jnis-2022-019456.
2. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de RNK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *Lancet Neurol*. 2009;8:635–642. https://doi.org/10.1016/S1474-4422(09)70126-7.
3. Sailer AMH, Wagemans BAJM, Nelemans PJ, de Graaf R, van Zwam WH. Diagnosing intracranial aneurysms with MR angiography: systematic review and meta-analysis. *Stroke*. 2014;45:119–126. https://doi.org/10.1161/STROKEAHA.113.003133.
4. Faron A, Sichtermann T, Teichert N, Luetkens JA, Keulers A, Nikoubashman O, et al. Performance of a deep-learning neural network to detect intracranial aneurysms from 3D TOF-MRA compared to human readers. *Clin Neuroradiol*. 2019. https://doi.org/10.1007/s00062-019-00809-w.
5. Sichtermann T, Faron A, Sijben R, Teichert N, Freiherr J, Wiesmann M. Deep learning-based detection of intracranial aneurysms in 3D TOF-MRA. *AJNR Am J Neuroradiol*. 2019;40:25–32. https://doi.org/10.3174/ajnr.A5911.
6. Ueda D, Yamamoto A, Nishimori M, Shimono T, Doishita S, Shimazaki A, et al. Deep learning for MR angiography: automated detection of cerebral aneurysms. *Radiology*. 2019;290:187–194. https://doi.org/10.1148/radiol.2018180901.
7. Claux F, Baudouin M, Bogey C, Rouchaud A. Dense, deep learning-based intracranial aneurysm detection on TOF MRI using two-stage regularized U-Net. *J Neuroradiol*. 2023;50:9–15. https://doi.org/10.1016/j.neurad.2022.03.005.
8. Sohn B, Park K-Y, Choi J, Koo JH, Han K, Joo B, et al. Deep learning-based software improves clinicians' detection sensitivity of aneurysms on brain TOF-MRA. *AJNR Am J Neuroradiol*. 2021;42:1769–1775. https://doi.org/10.3174/ajnr.A7242.
9. Park A, Chute C, Rajpurkar P, Lou J, Ball RL, Shpanskaya K, et al. Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA Netw Open*. 2019;2:e195600. https://doi.org/10.1001/jamanetworkopen.2019.5600.
10. Yang J, Xie M, Hu C, Alwalid O, Xu Y, Liu J, et al. Deep learning for detecting cerebral aneurysms with CT angiography. *Radiology*. 2021;298:155–163. https://doi.org/10.1148/radiol.2020192154.
11. Assis Y, Liao L, Pierre F, Anxionnat R, Kerrien E. Intracranial aneurysm detection: an object detection perspective. *Int J Comput Assist Radiol Surg*. 2024;19:1667–1675. https://doi.org/10.1007/s11548-024-03132-z.
12. Di Noto T, Marie G, Tourbier S, Alemán-Gómez Y, Esteban O, Saliou G, et al. Towards automated brain aneurysm detection in TOF-MRA: open data, weak labels, and anatomical knowledge. *Neuroinformatics*. 2023;21:21–34. https://doi.org/10.1007/s12021-022-09597-0.
13. Timmins KM, van der Schaaf IC, Bennink E, Ruigrok YM, An X, Baumgartner M, et al. Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: the ADAM challenge. *NeuroImage*. 2021;238:118216. https://doi.org/10.1016/j.neuroimage.2021.118216.
14. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203–211. https://doi.org/10.1038/s41592-020-01008-z.
15. Baumgartner M., Jaeger P.F., Isensee F., Maier-Hein K.H. nnDetection: a self-configuring method for medical object detection. arXiv:210600817 [Cs, Eess] 2021;12905:530–9. https://doi.org/10.1007/978-3-030-87240-3_51.
16. Luo X, Song T, Wang G, Chen J, Chen Y, Li K, et al. SCPM-Net: an anchor-free 3D lung nodule detection network using sphere representation and center points matching. *Med Image Anal*. 2022;75:102287. https://doi.org/10.1016/j.media.2021.102287.
17. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30:1323–1341. https://doi.org/10.1016/j.mri.2012.05.001.
18. Etminan N, Brown RD, Beseoglu K, Juvela S, Raymond J, Morita A, et al. The unruptured intracranial aneurysm treatment score. *Neurology*. 2015;85:881–889. https://doi.org/10.1212/WNL.0000000000001891.
19. Greving JP, Wermer MJH, Brown RD, Morita A, Juvela S, Yonekura M, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. *Lancet Neurol*. 2014;13:59–66. https://doi.org/10.1016/S1474-4422(13)70263-1.
20. Štepán-Buksakowska IL, Accurso JM, Diehn FE, Huston J, Kaufmann TJ, Luetmer PH, et al. Computer-Aided diagnosis improves detection of small intracranial aneurysms on MRA in a clinical setting. *Am J Neuroradiol*. 2014;35:1897–1902. https://doi.org/10.3174/ajnr.A3996.
21. Bijlenga P, Ebeling C, Jaegersberg M, Summers P, Rogers A, Waterworth A, et al. Risk of rupture of small anterior communicating artery aneurysms is similar to posterior circulation aneurysms. *Stroke*. 2013;44:3018–3026. https://doi.org/10.1161/STROKEAHA.113.001667.