NEUROINTERVENTION REVIEW

pISSN 2093-9043
eISSN 2233-6273

Neurointervention 2025;20:52-65
https://doi.org/10.5469/neuroint.2025.00283

Methodological Challenges in Deep Learning-Based Detection of Intracranial Aneurysms: A Scoping Review

Bio Joo, MD, PhD

Department of Radiology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea

---

Artificial intelligence (AI), particularly deep learning, has demonstrated high diagnostic performance in detecting intracranial aneurysms on computed tomography angiography (CTA) and magnetic resonance angiography (MRA). However, the clinical translation of these technologies remains limited due to methodological limitations and concerns about generalizability. This scoping review comprehensively evaluates 36 studies that applied deep learning to intracranial aneurysm detection on CTA or MRA, focusing on study design, validation strategies, reporting practices, and reference standards. Key findings include inconsistent handling of ruptured and previously treated aneurysms, underreporting of coexisting brain or vascular abnormalities, limited use of external validation, and an almost complete absence of prospective study designs. Only a minority of studies employed diagnostic cohorts that reflect real-world aneurysm prevalence, and few reported all essential performance metrics, such as patient-wise and lesion-wise sensitivity, specificity, and false positives per case. These limitations suggest that current studies remain at the stage of technical validation, with high risks of bias and limited clinical applicability. To facilitate real-world implementation, future research must adopt more rigorous designs, representative and diverse validation cohorts, standardized reporting practices, and greater attention to human-AI interaction.

Key Words: Artificial intelligence; Deep learning; Intracranial aneurysm; Methodology

---

INTRODUCTION

Intracranial aneurysms are focal dilations of cerebral arteries that affect approximately 3 to 7 percent of the general population.[1,2] Although the annual risk of rupture for an unruptured intracranial aneurysm is relatively low, averaging around 1 percent, the consequences can cause subarachnoid hemorrhage (SAH), a life-threatening hemorrhagic stroke with high morbidity and mortality rates.[3] Given these serious outcomes, early detection of intracranial aneurysms is crucial for guiding timely and appropriate intervention.

With the substantial advancement of artificial intelligence (AI) in recent years, particularly in deep learning, a growing body of research has demonstrated its potential applicability in the medical field. To date, the U.S. Food and Drug Administration has approved over 1,000 AI- and machine learning-enabled medical devices.[4] However, compelling evidence of the clinical benefits or widespread adoption of AI in routine clinical practice, beyond controlled research settings, remains limited.[4] This phenomenon has been described as the "AI chasm," a term introduced by Keane and Topol[5] to highlight the disconnect between the strong performance of AI algorithms in research environments and their limited impact in real-world clinical applications. Several factors

Received: March 29, 2025 **Revised:** May 5, 2025
Accepted: May 12, 2025

Correspondence to:
Bio Joo, MD, PhD
Department of Radiology, Gangnam Severance Hospital, Yonsei University College of Medicine, 211 Eonju-ro, Gangnam-gu, Seoul 06273, Korea
Tel: +82-2-2019-3510 **Fax:** +82-2-3462-5472
E-mail: yonnebio@yuhs.ac

contribute to this gap, including limited generalizability, challenges associated with clinical integration, the inherent lack of explainability of AI algorithms, insufficient user knowledge, ethical/legal considerations, and cultural resistance to adopting new technologies.[6-8]

In the application of AI for the detection of intracranial aneurysms, concerns regarding bias and limited generalizability have been prominently addressed. A meta-analysis by Din et al.,[9] which assessed the diagnostic performance of AI in detecting intracranial aneurysms on magnetic resonance angiography (MRA), computed tomography angiography (CTA), or digital subtraction angiography (DSA), reported a pooled sensitivity of 91.2% and an area under the receiver operating characteristic curve of 0.936. However, the authors noted that most of the included studies exhibited a high risk of bias and poor generalizability, as evaluated by the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool.[10] They further emphasized that this lack of generalizability remains a major obstacle to the widespread clinical adoption of AI algorithms for intracranial aneurysm detection. Similarly, another systematic review and meta-analysis published in 2023, which evaluated the diagnostic accuracy of deep learning-based algorithms for detecting intracranial aneurysms on CTA, reported a high pooled sensitivity of 0.87 for aneurysms larger than 3 mm. Nonetheless, this review also identified substantial concerns regarding risk of bias and methodological limitations across the included studies.[11] Consequently, users of these AI-based diagnostic tools—particularly clinicians—are increasingly interested not only in their diagnostic accuracy, but also in identifying the patient populations most likely to benefit from their use, the specific clinical settings in which the algorithms have been validated, the inherent limitations of each model, and, ultimately, their clinical utility. Moreover, there is growing interest in understanding how far these AI algorithms have progressed toward integration into routine clinical practice.

Although previous review articles have offered valuable insights into the methodology of existing studies through systematic evaluation using the evaluation tools such as QUADAS-2, the relatively broad criteria of these instruments—particularly in the patient selection domain—may hinder readers from fully appreciating the nuances of study design limitation and to identify the sources of bias and limited generalizability in research on AI-based detection of intracranial aneurysms. For example, in the systemic review by Din et al.,[9] 95% (41 out of 43) of studies were rated as having high or unclear concerns regarding applicability in the patient selection domain. However, the specific aspects of study design that contributed to these ratings remain difficult to discern. Moreover, general evaluation tools may not adequately capture the distinct clinical considerations specifically relevant to intracranial aneurysm detection.

In this context, the aim of this scoping review is to provide a comprehensive overview of study methodologies in published research on deep learning-based detection of intracranial aneurysms on CTA or MRA, with a particular focus on study populations, validation strategies, reporting practices, and reference standards, rather than on their diagnostic performance. By identifying and synthesizing the methodological limitations of existing studies, this review may help shape future approaches to algorithm development and contribute to ongoing efforts to establish their clinical utility.

METHODS

A systematic literature search was conducted in January 2025 using the Embase, MEDLINE, and Web of Science databases. The search used the following query: ('artificial intelligence' OR 'deep learning' OR 'computer assisted' OR 'computer aided' OR 'AI' OR 'automated detection' OR 'automatic detection') AND ('intracranial aneurysm' OR 'intracranial aneurysms' OR 'cerebral aneurysm' OR 'cerebral aneurysms'). Search results were screened by evaluating their titles and abstracts. The exclusion criteria were as follows: articles that were not original research; studies focused on rupture risk prediction, treatment outcome prediction, aneurysm segmentation, technological development or image quality, hemodynamics, or unrelated subjects (e.g., genomics, treatment simulation); studies comparing diagnostic performance with and without AI assistance; studies primarily based on DSA; studies on computer-aided diagnosis using conventional machine learning published before 2018; studies with inadequate reporting that compromised credibility; and publications not in English (Fig. 1).

In the 36 studies that were finally selected, 14 key questions pertaining to study population selection, validation of diagnostic performance, reporting methodology, and the reference standard used were systematically investigated (Table 1).

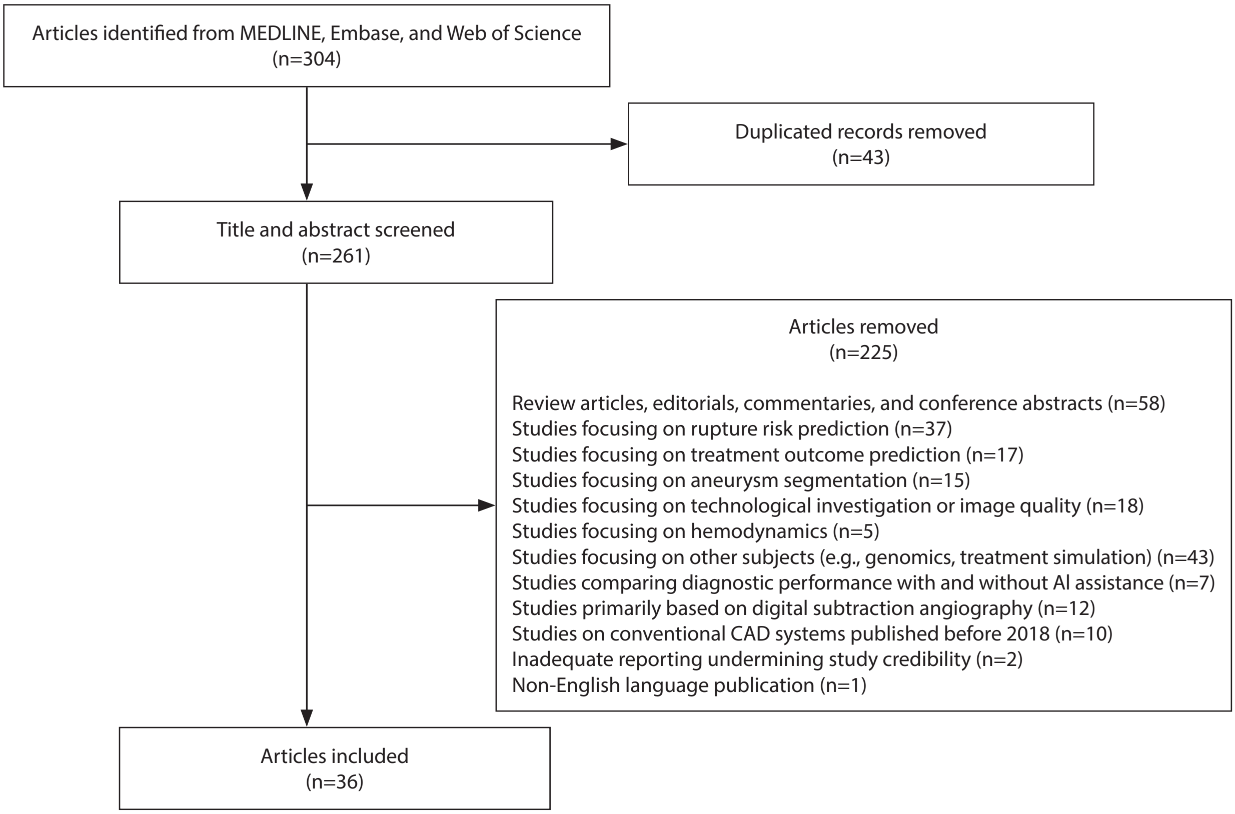
Articles identified from MEDLINE, Embase, and Web of Science
(n=304)
Duplicated records removed
(n=43)
Title and abstract screened
(n=261)
Articles removed
(n=225)
Review articles, editorials, commentaries, and conference abstracts (n=58)
Studies focusing on rupture risk prediction (n=37)
Studies focusing on treatment outcome prediction (n=17)
Studies focusing on aneurysm segmentation (n=15)
Studies focusing on technological investigation or image quality (n=18)
Studies focusing on hemodynamics (n=5)
Studies focusing on other subjects (e.g., genomics, treatment simulation) (n=43)
Studies comparing diagnostic performance with and without AI assistance (n=7)
Studies primarily based on digital subtraction angiography (n=12)
Studies on conventional CAD systems published before 2018 (n=10)
Inadequate reporting undermining study credibility (n=2)
Non-English language publication (n=1)
Articles included
(n=36)

Fig. 1. Flowchart of articles screened and included in this review. AI, artificial intelligence; CAD, computer-aided diagnosis.

Table 1. Methodological evaluation criteria: 14 key questions on study population selection, validation, reporting methodology, and the reference standard used

Category	Question
Study population selection	1. Did the authors exclude aneurysms of certain sizes from the study population?
	2. Did the authors exclude aneurysms located in specific regions from the study population?
	3. Did the authors exclude ruptured aneurysms in the study population?
	4. Did the authors exclude subjects who had previously undergone clipping or coil embolization for intracranial aneurysms?
	5. Did the authors exclude or address concurrent pathological findings in the brain parenchyma or intracranial vessels (e.g., brain tumor, hematoma, arterial stenosis/occlusion, or vascular malformation)?
Validation of diagnostic performance	6. Was the study conducted prospectively or retrospectively?
	7. Was geographically external validation conducted?
	8. Did the test set or external validation set include more than 100 cases?
	9. Did the authors include examinations in the external validation set that were acquired using vendors different from those used in the training set, or was the number of scanners used in the external validation set at least 3?
	10. Was the external validation based on a diagnostic cohort that reflects the true prevalence of intracranial aneurysms?
Reporting methodology	11. Did the authors report both patient-wise sensitivity and lesion-wise sensitivity?
	12. Did the authors report patient-wise specificity?
	13. Did the authors report the number of false positives per case?
Reference standard	14. What was the reference standard used in the study?

Table 2. Results of the methodological evaluation of included studies

Author	Year	Q1. Size	Q2. Location	Q3. Ruptured aneurysm	Q4. Treated aneurysm	Q5. Concurrent findings	Q6. Retro-spective or prospective	Q7. External validation	Q8. Test size >100	Q9. Multi-vendors	Q10. Diagnostic cohort	Q11. Sensitivity	Q12. Specificity	Q13. FP/case	Q14. Reference standard	Total score
Nakao et al.[12]	2018	0	1	Unclear	0	Unclear	0	0	1	-	-	0	0	1	Consensus	3
Stember et al.[13]	2019	0	1	Unclear	Unclear	Unclear	0	0	0	-	-	0	0	0	Consensus	1
Ueda et al.[14]	2019	1	1	Unclear	Included but not tested	Unclear	0	1	1	1	0	0	0	1	Consensus	6
Sichtermann et al.[15]	2019	1	1	Unclear	0	Unclear	0	0	5-fold CV	-	-	0	0	1	Consensus	3
Shi et al.[16]	2020	1	1	1	0	AVM, MMD, occlusion	0	1	1	1	1	1	1	1	DSA, consensus	10
Joo et al.[17]	2020	1	1	0	1	Tumor, hematoma	0	1	1	1	0	0	1	0.5	Consensus	7.5
Chen et al.[18]	2020	1	1	0	Unclear	Unclear	0	0	0	-	-	0	0	1	DSA	3
Dai et al.[19]	2020	1	1	Unclear	Included but not tested	Unclear	0	0	1	-	-	0	0	1	Consensus	4
Shahzad et al.[20]	2020	1	1	1	0	Unclear	0	0	1	-	-	0	0	1	Consensus	5
Joo et al.[21]	2021	0	1	0	0	Tumor, hematoma	0	0	1	-	-	1	1	1	Consensus	5
Pennig et al.[22]	2021	1	1	1	0	Unclear	0	0	1	-	-	0	0	1	Consensus	5
Yang et al.[23]	2021	1	1	1	0	Unclear	0	0	1	-	-	0	0	1	Consensus	5
Terasaki et al.[24]	2022	0	1	Unclear	Unclear	Unclear	0	0	1	-	-	0	0	1	Consensus	3
Wei et al.[25]	2022	1	1	Unclear	0	Unclear	0	1	1	Unclear	Unclear	1	Unclear	0.5	DSA	5.5
Lehnen et al.[26]	2022	1	1	1	Unclear	Major competing pathologies (hemorrhage)	0	1	1	1	0	0	1	1	Consensus	8
You et al.[27]	2022	1	1	0	0	AVM, MMD	0	0	1	-	-	0	0	1	DSA	4
Heit et al.[28]	2022	0	1	0	0	Unclear	0	1	0	1	0	0	0	0.5	Consensus	3.5
Ham et al.[29]	2023	1	1	1	0	Unclear	0	1	1	Unclear	0	0	1	1	Consensus	7
Tajima et al.[30]	2023	1	1	0	0	Unclear	0	1	0	1	0	0	0	1	Consensus	5
Liu et al.[31]	2023	1	1	Unclear	0	Tumor	0	0	0	-	-	0	0	0	DSA	2
Claux et al.[32]	2023	1	1	0	Unclear	Occlusion	0	0	0	-	-	0	0	1	DSA	3

Table 2. Continued

Author	Year	Q1. Size	Q2. Location	Q3. Ruptured aneurysm	Q4. Treated aneurysm	Q5. Concurrent findings	Q6. Retro-spective or prospective	Q7. External validation	Q8. Test size >100	Q9. Multi-vendors	Q10. Diagnostic cohort	Q11. Sensitivity	Q12. Specificity	Q13. FP/case	Q14. Reference standard	Total score
Ishihara et al.[33]	2023	1	1	0	Unclear	Unclear	0	1	1	Unclear	0	0	0	1	Consensus	5
Wang et al.[34]	2023	1	1	1	0	AVM, MMD, tumor	0	0	1	-	-	1	0	1	DSA, consensus	6
Zhou et al.[35]	2023	1	1	Unclear	Unclear	Stenosis	0	1	0	1	0	0	0	1	Consensus	5
Colasurdo et al.[36]	2023	1	1	0	Unclear	Unclear	0	1	1	1	0	0	1	0.5	Consensus	6.5
Hu et al.[37]	2024	1	1	1	0	AVM, MMD, occlusion, dissection	1	1	1	1	1	1	1	1	DSA	11
Adamchic et al.[38]	2024	1	1	0	Unclear	Major competing pathologies	0	1	1	1	1	1	1	0.5	Consensus	8.5
Bizjak et al.[39]	2024	1	1	1	1	Unclear	0	1	1	1	0	0	1	1	Consensus	9
Li et al.[40]	2024	1	1	0	0	Unclear	0	1	1	1	0	1	1	0	Consensus	7
You et al.[41]	2024	1	1	1	0	AVM, MMD	0	1	1	1	0	1	0	1	DSA, consensus	8
Wei et al.[42]	2024	1	1	0	Unclear	AVM, MMD	0	0	1	-	-	1	1	0.5	DSA, consensus	5.5
De Toledo et al.[43]	2024	1	1	0	Unclear	Unclear	0	1	1	Unclear	1	0	1	0	Consensus	6
Goertz et al.[44]	2025	1	1	1	0	Unclear	0	0	1	-	-	0	0	1	Consensus	5
Zhuo et al.[45]	2025	1	1	Unclear	0	Deformity, AVF, extremely tortuous vessel	0	1	1	1	1	1	1	0.5	Consensus	8.5
Hu et al.[46]	2025	1	1	1	0	AVM, MMD, occlusion	0	1	1	Unclear	1	1	1	1	DSA, consensus	9
Schmidt et al.[47]	2025	1	1	0	0	AVM	0	1	1	Unclear	1	0	0	0	Consensus	5

FP, false positive; CV, cross validation; AVM, arteriovenous malformation; MMD, moyamoya disease; DSA, digital subtraction angiography; AVF, arteriovenous fistula.

RESULTS

The results of the methodological evaluation of the included studies, based on the 14 key questions, are summarized in Table 2.[12-47] The scoring system was designed such that a higher total score indicates a lower risk of bias and concerns regarding generalizability. A detailed description of the scoring criteria is provided in Supplementary Material 1. Fig. 2 illustrates an upward trend in the mean scores of the included studies over their publication years. Items that could not be quantitatively assessed were excluded from the total score calculation.

Study Population Selection

Q1. Exclusion criteria based on aneurysm size

Among the 36 studies reviewed, 5 studies excluded intracranial aneurysms based on size criteria. Nakao et al.[12] excluded aneurysms smaller than 2 mm in diameter, while Stember et al.[13] and Heit et al.[28] excluded those measuring less than 3 mm. In contrast, 2 studies excluded large aneurysms: Joo et al.[21] excluded those exceeding 25 mm in diameter, and Terasaki et al.[24] excluded aneurysms larger than 15 mm. The remaining studies did not specify any exclusion criteria related to aneurysm size.

Q2. Exclusion criteria based on aneurysm location

No study explicitly listed specific aneurysm location as an exclusion criterion. However, in 2 studies, aneurysms located in the posterior circulation were not included in the test dataset.[18,33]

Q3. Rupture status of aneurysms in study populations

The inclusion or exclusion of ruptured aneurysms in the testing set was evaluated in the included studies. The training set was not considered in this assessment. Among the 36 studies, 12 included ruptured aneurysms in their study populations. Of these, 3 studies specifically focused on detecting aneurysms in patients with aneurysmal SAH.[20,22,24] Among the remaining studies, 3 clearly stated in the title or study aim that only unruptured aneurysms were investigated.[21,28,33] Eleven studies indicated in the inclusion/exclusion criteria or discussion that ruptured aneurysms were not included. The remaining 10 studies did not clearly state whether ruptured aneurysms were included in the study population.

Q4. Exclusion of previously treated aneurysms

This assessment examined whether aneurysms previously treated with coil embolization or surgical clipping were excluded from the testing of the AI algorithms. In studies that included external validation, the evaluation was based on the population used in the external validation set. Twenty-one studies excluded previously treated aneurysms from

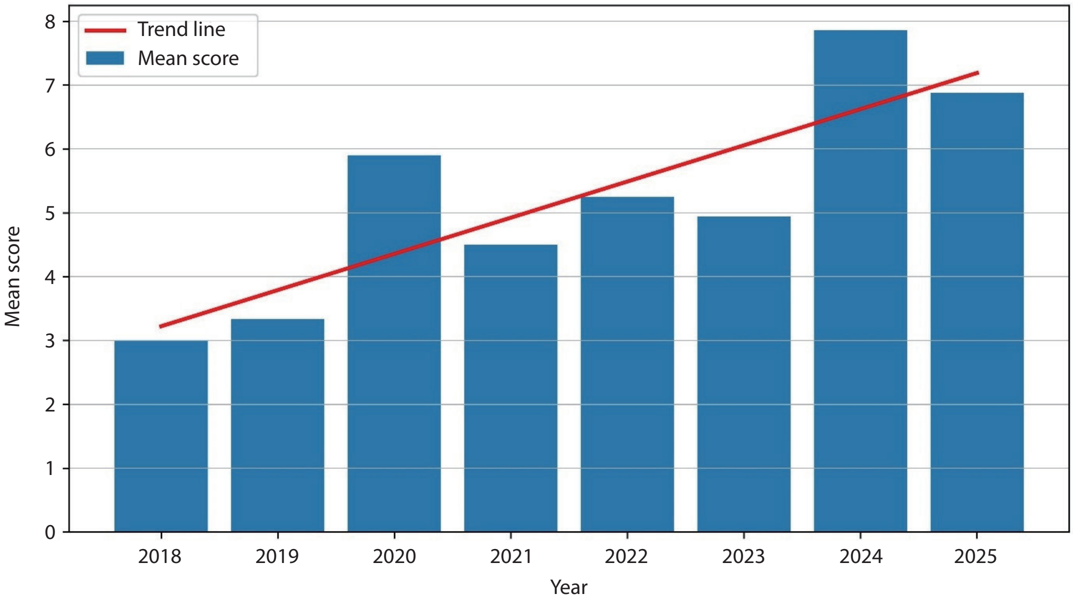

Fig. 2. Mean score of included studies by year.

their study populations. In 2 studies, patients with treated aneurysms were included, but the treated aneurysms themselves were not used in the testing process.[14,19] Only 2 studies explicitly stated that treated aneurysms were included in the study population for evaluating the diagnostic performance of their algorithms.[17,39] The remaining 11 studies did not specify whether treated aneurysms were included in the study population.

Q5. Consideration of coexisting brain parenchymal or vascular abnormalities

This assessment evaluated whether the studies addressed the presence of coexisting brain parenchymal or intracranial vascular abnormalities in their study populations. Sixteen studies explicitly reported the exclusion of specific abnormalities. The most frequently cited were vascular malformations such as arteriovenous malformations or fistulas (n=9), moyamoya disease (n=7), arterial occlusion or stenosis (n=5), and tumors (n=4). These exclusion criteria were not mutually exclusive; individual studies often excluded more than 1 type of abnormality. The remaining 20 studies did not clearly address the presence of such coexisting abnormalities.

Validation of Diagnostic Performance

Q6. Prospective or retrospective approach

Only 1 study utilized a prospective design. In that study, a portion of the external validation was conducted retrospectively. However, in 1 external dataset, the model was prospectively applied and evaluated in 1,562 real-world clinical CTA cases. All other studies were conducted retrospectively.

Q7. Implementation of external validation

In this assessment, an external validation dataset was defined as data obtained from institutions geographically distinct from those used for model training. Consequently, temporally independent data from the same institution—referred to as temporal validation—were not considered external validation in this process.[48] Studies that evaluated the diagnostic performance of a preexisting AI model without further model development were also regarded as having conducted external validation. However, even in such cases, if the test dataset was clearly derived from the same institution as the training dataset, it was not classified as external validation. Furthermore, even in studies described as 'multicenter study,' if the data used for validation did not originate from institutions entirely separate from those providing the training data, the validation was not deemed external.

According to the defined criteria, 20 studies performed external validation to evaluate the diagnostic performance of the AI model. Among these, 11 assessed the performance of preexisting AI models or commercially available software.[25,26,28,30,33,36,38,40,43,46,47] In contrast, 16 studies did not conduct external validation. Of these, 2 studies utilized a preexisting AI model but appeared to have tested it on data acquired from the same institution as the training set.[22,44] One study employed 5-fold cross-validation without incorporating an independent test set.[15]

Q8. Size of the test or external validation set

The size of the independent test set, or the external validation set if applicable, was evaluated and categorized based on whether it included more than 100 cases. Twenty-six studies included more than 100 cases in the test set or, when external validation was performed, in both the test set and the external validation set. Among the studies that conducted external validation, 2 included more than 100 cases in only 1 of the 2 datasets. The study by Ham et al.[29] included 15 cases in the internal test set but used 113 cases for external validation, obtained from the open-source Aneurysm Detection and segMentation (ADAM) challenge database. In contrast, the study by Ueda et al.[14] included 521 cases in the internal test set but only 67 cases in the external validation set. In 7 studies, the number of cases in the test set or in both the test set and the external validation set was fewer than 100, with 1 study including only 10 cases in the test set.[31] As noted above, 1 study did not employ an independent test set or external validation set; therefore, the current assessment was not applicable.[15]

Q9. Use of multiple vendors or scanners in external validation sets

This analysis was restricted to studies that conducted external validation. Although external validation enhances the assessment of a model's generalizability, its value may be diminished if the scanners used are the same as those in the training set. Therefore, this assessment evaluated whether the scanners used in the external validation set were different from those used during training, or whether at least 3 different scanners were employed. This criterion was used to assess whether the AI model could reasonably claim generalizability.

Among the 20 studies that conducted external validation, 14 employed sufficiently different scanners in their external validation process, while this information was unclear in 6 studies. Of these 6, 5 utilized a preexisting AI model for which the scanners used during model development were not specified. In the remaining study, the external validation dataset was derived from the ADAM challenge database, and details regarding the scanners used were incomplete.[29]

Q10. Representation of real-world aneurysm prevalence in external validation cohorts

Evaluating diagnostic performance using a diagnostic cohort that reflects the true prevalence of intracranial aneurysms, rather than relying on datasets with a predetermined number of positive and negative cases, is generally considered a more reliable approach and may provide a closer approximation of real-world performance.[49,50]

This assessment also focused exclusively on studies that conducted external validation. A dataset was considered a diagnostic cohort only when the study population was recruited consecutively or randomly, without prior knowledge of aneurysm prevalence. Based on this definition, 7 studies were identified as using a diagnostic cohort in their external validation. Twelve studies did not meet this criterion. In 1 study, the recruitment method for the dataset was not clearly described; however, it was likely non-consecutive or non-random, as 179 out of 212 patients were diagnosed with aneurysms.[25]

Reporting Methodology

Questions 11 to 13 were applied only to the performance of the AI model when used in a standalone setting. If a study employed a paired-design to compare user performance between AI-assisted and AI-unassisted interpretations, the metrics from that comparison were not included in this assessment.

Q11. Sensitivity metrics: patient-wise and lesion-wise

Of the 36 included studies, 11 reported both patient-wise and lesion-wise sensitivity, while the remaining 25 studies reported only 1 of the 2.

Q12. Patient-wise specificity

Fifteen studies reported patient-wise specificity. In contrast, 21 studies did not report this metric, primarily because they included only aneurysm-positive examinations, making the calculation of specificity impossible. Of these, 3 studies reported specificity based on vessel-level analysis rather than patient-wise analysis.[13,28,33]

Q13. Reporting of false positives per case

This assessment evaluated whether the authors reported the number of false positives per case or presented a free-response receiver operating characteristic curve. Twenty-four studies explicitly reported this information. In 7 studies, the metric was not directly stated but could be inferred from the reported number of false positive detections and the total number of cases. In the remaining 5 studies, the information was not available.

Reference Standards

Q14. The reference standard used

The reference standard used across the included studies was categorized as either DSA or radiologist consensus. If a study primarily relied on radiologist consensus but incorporated DSA findings when available, it was classified as using consensus.

Six studies used DSA as the reference standard, including only cases with available DSA results. In 5 studies, both DSA and radiologist consensus were used based on the dataset. The remaining 25 studies primarily relied on radiologist consensus as their reference standard.

DISCUSSION

This scoping review provides a comprehensive overview of study methodologies in published research on deep learning–based detection of intracranial aneurysms using CTA or MRA. Rather than focusing on their diagnostic performance, particular attention was given to study population selection, validation strategies, reporting practices, and reference standards. This approach aims to clarify the current state of the field and highlight methodological factors that may impede the demonstration of clinical utility.

Easing exclusion criteria to reduce selection bias and better reflect real-world clinical variability may pose substantial practical challenges, including the need for larger and more heterogeneous datasets, increased complexity in model development, and potential declines in model performance. These difficulties may have led many studies to favor more restricted datasets, even at the expense of clinical generalizability. Although explicit exclusion of aneurysms based on size or location was relatively uncommon (Q1 and Q2) among the included studies, more than half of the studies either excluded ruptured or previously treated aneurysms, or failed to clearly specify how these cases were handled (Q3 and Q4). Notably, only 2 studies explicitly reported the inclusion of treated aneurysms. Given that real-world clinical scenarios frequently involve ruptured or previously treated aneurysms, such exclusions or ambiguities may limit the generalizability and clinical applicability of these AI models. Additionally, some studies focused specifically on aneurysm detection in the context of aneurysmal SAH, where at least 1 ruptured aneurysm per case can be presumed. Taken together, the inclusion/exclusion criteria across studies vary considerably—with some excluding ruptured aneurysms, others including only ruptured cases, and many not specifying rupture status—highlighting the need for caution when interpreting and comparing findings across the literature.

Another important concern was the limited consideration of concurrent findings (Q5). Sixteen studies explicitly excluded specific coexisting abnormalities, while 20 did not clearly address this issue. It is possible that studies which did not list coexisting findings as exclusion criteria did so intentionally to include all such cases. For example, 1 study did not mention coexisting abnormalities in the exclusion criteria but later reported in the results that some false-positive findings of the model were due to misclassification of arteriovenous malformations or fistulas as aneurysms.[25] Even so, the lack of explicit reporting makes it difficult to assess the robustness of AI models in the presence of other brain or cerebrovascular abnormalities. Providing clearer information about the inclusion of such cases would improve clinical relevance of these studies.

Robust clinical verification of the performance of a diagnostic AI model requires external validation using a clinical cohort that accurately reflects the characteristics of the target patient population.[50] However, only 20 out of 36 studies reported conducting geographically external validation (Q7). This number further decreases to 9 when excluding the 11 studies that evaluated preexisting algorithms or commercially available software, indicating that only a minority of studies validated their own AI models using independently recruited external datasets. Even among studies that performed external validation, the generalizability of the models remains a concern due to limitations in dataset size and scanner diversity (Q8 and Q9). Of the 20 studies, only 10 used external datasets comprising more than 100 cases with sufficient variability in scanner types. Taken together, there remains a limited body of well-designed research that can robustly support the generalizability of AI models for intracranial aneurysm detection.

The lack of prospective studies has long been recognized as a major limitation in the development and clinical implementation of AI models for intracranial aneurysm detection.[9] Among the 36 articles included in this review, only 1 study validated its model using a prospective cohort (Q6). Given the methodological challenges of conducting prospective studies, the use of retrospective diagnostic cohorts that approximate real-world clinical settings may serve as a practical alternative for model evaluation, especially when cases are not selected through convenience sampling. Convenience sampling, which selects disease-positive and disease-negative cases separately rather than consecutively, can distort the spectrum of the diseased and non-diseased states in the dataset and introduce spectrum bias, limiting the model's applicability to real-world patient populations.[50] However, only 7 studies employed such cohorts for model testing, suggesting that this approach remains underutilized (Q10).

For AI models designed to detect intracranial aneurysms, it is essential to evaluate diagnostic performance from multiple perspectives to ensure both technical accuracy and clinical relevance. To this end, the following 4 evaluation metrics should be reported in combination: patient-wise sensitivity, lesion-wise sensitivity, patient-wise specificity, and the number of false positives per case. Each metric captures a distinct yet complementary aspect of model performance. Lesion-wise sensitivity measures how accurately the model detects individual aneurysms, which is particularly important in patients with multiple lesions. However, this metric alone does not indicate whether the model detects any aneurysm in a given patient—an aspect more directly reflected by patient-wise sensitivity, which aligns closely with clinical decision-making. Conversely, patient-wise sensitivity may overestimate overall performance by ignoring missed lesions when at least 1 is identified, underscoring the need for lesion-wise evaluation to ensure thorough detection. Patient-wise specificity indicates how reliably the model identifies aneurysm-negative patients, helping to reduce unnecessary follow-ups and false-positive burdens—issues not fully captured by vessel- or segment-level specificity. The number of false positives per case is also critical, as excessive false alarms can hinder workflow and reduce diagnostic confidence. Just as lesion-wise sensitivity and patient-wise sensitivity offer complementary perspectives on sensitivity, false positives per case and patient-wise specificity together provide a more complete understanding of specificity.[11,51] Despite their importance, only 4 of the 36 included studies explicitly reported all 4 metrics (Q11 to 13). Notably, 21 studies did not report patient-wise specificity; among these, 16 included only aneurysm-positive cases, making it impossible to calculate this metric. The definitions of the metrics also varied across studies. For example, some reported vessel-wise specificity, instead of patient-wise specificity, by assessing aneurysms at the level of individual vascular segments. The definition of false positives per case also varied. Some studies counted false positives on a per-patient basis, while others used lesion-level counts as the numerator. The denominator likewise differed, with some studies including all examinations regardless of aneurysm status, and others including only aneurysm-negative cases. These inconsistencies underscore the need for standardized reporting practices to ensure accurate interpretation and meaningful comparison across studies.

DSA is widely regarded as the gold standard imaging modality for diagnosing intracranial aneurysms. Previous reviews have noted that many studies used radiologist consensus rather than DSA as the reference standard, consistent with our findings that only 6 studies employed DSA as the reference standard (Q14). However, due to the risk of procedure-related morbidity, DSA is typically reserved for cases where the potential benefit justifies the risk, particularly for treatment planning or when noninvasive imaging results are inconclusive or inconsistent.[52] As a result, restricting the study population to patients who underwent DSA can be an additional source of spectrum bias, as it may not reflect the broader clinical population where the AI model was intended to be applied. Therefore, it is important to recognize that the choice between DSA and radiologist consensus as the reference standard entails a trade-off between diagnostic certainty and the representativeness of the target population. This decision should be guided by the specific clinical scenario and the target population in which the AI model is intended to be applied. Selecting different reference standards for different purposes can be a practical approach, as demonstrated in a previous study where DSA was used for evaluating diagnostic accuracy, while radiologist consensus was used for assessing generalizability in external validation.[42]

One important yet often overlooked issue is the diagnostic gray zone in identifying intracranial aneurysms. Even with the gold standard DSA, uncertainty may remain as to whether a bulging contour represents a true aneurysm, an infundibulum, or a tortuous vessel, as the determination ultimately relies on visual interpretation of a complex 3-dimensional structure. This challenge is further exacerbated in studies relying on convenience sampling, which often yield a study population where aneurysm cases are distinctly abnormal and control cases are distinctly normal, thereby limiting the representativeness of the target population encountered in real-world clinical practice. Addressing these gray areas warrants careful consideration, and the application of uncertainty quantification may offer a practical approach that could enhance both research and clinical use of AI in this field.[53,54] In addition, reporting inter- and intrarater variability of features annotated for the reference standard may be considered to address this issue, as recommended in the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guideline.[55]

In general, the integration of AI tools into clinical practice necessitates a stepwise evaluation process, beginning with the validation of technical performance, followed by the assessment of clinical performance.[50,56] Typically, technical performance is assessed in case-control studies, while clinical performance uses diagnostic cohort designs.[54] Reflecting this progression, the studies reviewed showed a trend toward improved methodology over time (Fig. 2). Earlier studies focused on technical performance using only aneurysm-positive cases or convenience samples, whereas recent studies more often adopted cohort designs and external validation. Notably, only 1 study employed a prospective design, highlighting that clinical validation of AI models for intracranial aneurysm detection remains in its early stages.

Beyond technical and clinical performance, human-AI interaction is an important factor to consider for the successful integration of AI into clinical practice. This interaction involves several dimensions, including interface design, explainability, trust, fairness, adaptability, and accountability.[4] Inadequate attention to these aspects has been associated with increased cognitive burden and digital fatigue among users.[4,57] However, formal conceptual frameworks for evaluating the quality of human-AI interaction are still lacking.[58] In the specific context of AI models for intracranial aneurysm detection, comparing diagnostic performance with and without AI assistance could serve as a proxy for assessing human-AI interaction; however, this was not within the scope of the present review. Moreover, the included studies provided insufficient information to enable a meaningful evaluation of human-AI interaction. Nevertheless, medical algorithmic auditing may offer a useful approach to improving human-AI interaction by systematically identifying and characterizing algorithmic errors, as demonstrated in a recent study.[46] Such efforts can enhance our understanding of the limitations and failure modes of AI models and offer insights to guide the development and validation of future AI systems in this context.[59]

This review has several limitations. First, it excluded AI models designed for rupture risk prediction, aneurysm segmentation, or treatment outcome prediction. Second, it focused specifically on studies utilizing CTA and MRA, while excluding those based on DSA. As DSA is both invasive and considered the gold standard for diagnosing intracranial aneurysms, its clinical application differs substantially from that of CTA or MRA. Therefore, in the context of evaluating the potential utility of AI models for identifying patients who may be appropriate candidates for DSA, this review was limited to studies using CTA and MRA. In addition, some studies did not provide sufficient information regarding the imaging scanners used in their datasets. As a result, the assessment of scanner or vendor diversity may be subject to some limitations. Another limitation of this review is that all included studies were evaluated using a single methodological standard, irrespective of their position within the AI development and implementation framework.[50,56] This uniform approach may not fully reflect the contextual and methodological differences among studies and may overlook nuances in study design and intended application stage. Nevertheless, the objective was not to critique individual studies, but to provide an overview of current trends in the field.

In conclusion, this scoping review provides a comprehensive methodological overview of deep learning–based AI studies for intracranial aneurysm detection using CTA and MRA. Rather than emphasizing diagnostic performance, the review focused on methodological factors—such as population selection, validation strategies, reporting practices, and reference standards—within the context of bias and generalizability concerns. Common limitations included inconsistent handling of ruptured or treated aneurysms, incomplete reporting of coexisting pathologies, limited use of external validation, non-random or non-consecutive sampling, underreporting of key performance metrics, insufficient scanner diversity, and a near-absence of prospective validation. These findings indicate that most studies remain at the stage of technical performance evaluation, with a high risk of bias and poor generalizability, reflecting limited progress toward clinical performance assessment. To support real-world implementation, future research will require more rigorous study designs, representative validation cohorts, standardized reporting, and greater attention to human-AI interaction.

SUPPLEMENTARY MATERIALS

Supplementary material related to this article can be found online at https://doi.org/10.5469/neuroint.2025.00283.

Fund
None.

Ethics Statement
This article was exempted from the review by the institutional ethics committee. This article does not include any information that may identify the person.

Conflicts of Interest
The author has no conflicts to disclose.

Author Contributions
Concept and design: BJ. Analysis and interpretation: BJ. Data collection: BJ. Writing the article: BJ. Critical revision of the article: BJ. Final approval of the article: BJ. Overall responsibility: BJ.

ORCID
Bio Joo: https://orcid.org/0000-0001-7460-1421

REFERENCES

1. Vlak MH, Algra A, Brandenburg R, Rinkel GJ. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. *Lancet Neurol* 2011;10:626-636
2. Li MH, Chen SW, Li YD, Chen YC, Cheng YS, Hu DJ, et al. Prevalence of unruptured cerebral aneurysms in Chinese adults

aged 35 to 75 years: a cross-sectional study. *Ann Intern Med* 2013;159:514-521

3. UCAS Japan Investigators. The natural course of unruptured cerebral aneurysms in a Japanese cohort. *N Engl J Med* 2012;366:2474-2482
4. Park SH, Langlotz CP. Crucial role of understanding in human-artificial intelligence interaction for successful clinical adoption. *Korean J Radiol* 2025;26:287-290
5. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *NPJ Digit Med* 2018;1:40
6. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics (Basel)* 2023;13:2760
7. Jussupow E, Spohrer K, Heinzl A. Identity threats as a reason for resistance to artificial intelligence: survey study with medical students and professionals. *JMIR Form Res* 2022;6:e28750
8. Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, et al. An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. *Eur Radiol* 2021;31:8797-8806
9. 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;15:262-271
10. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al.; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529-536
11. Bizjak Ž, Špiclin Ž. A systematic review of deep-learning methods for intracranial aneurysm detection in CT angiography. *Biomedicines* 2023;11:2921
12. Nakao T, Hanaoka S, Nomura Y, Sato I, Nemoto M, Miki S, et al. Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. *J Magn Reson Imaging* 2018;47:948-953
13. Stember JN, Chang P, Stember DM, Liu M, Grinband J, Filippi CG, et al. Convolutional neural networks for the detection and measurement of cerebral aneurysms on magnetic resonance angiography. *J Digit Imaging* 2019;32:808-815
14. 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
15. 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
16. Shi Z, Miao C, Schoepf UJ, Savage RH, Dargis DM, Pan C, et al. A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. *Nat Commun* 2020;11:6090
17. Joo B, Ahn SS, Yoon PH, Bae S, Sohn B, Lee YE, et al. A deep learning algorithm may automate intracranial aneurysm detection on MR angiography with high diagnostic performance. *Eur Radiol* 2020;30:5785-5793
18. Chen G, Wei X, Lei H, Liqin Y, Yuxin L, Yakang D, et al. Automated computer-assisted detection system for cerebral aneurysms in time-of-flight magnetic resonance angiography using fully convolutional network. *Biomed Eng Online* 2020;19:38
19. Dai X, Huang L, Qian Y, Xia S, Chong W, Liu J, et al. Deep learning for automated cerebral aneurysm detection on computed tomography images. *Int J Comput Assist Radiol Surg* 2020;15:715-723
20. Shahzad R, Pennig L, Goertz L, Thiele F, Kabbasch C, Schlamann M, et al. Fully automated detection and segmentation of intracranial aneurysms in subarachnoid hemorrhage on CTA using deep learning. *Sci Rep* 2020;10:21799
21. Joo B, Choi HS, Ahn SS, Cha J, Won SY, Sohn B, et al. A deep learning model with high standalone performance for diagnosis of unruptured intracranial aneurysm. *Yonsei Med J* 2021;62:1052-1061
22. Pennig L, Hoyer UCI, Krauskopf A, Shahzad R, Jünger ST, Thiele F, et al. Deep learning assistance increases the detection sensitivity of radiologists for secondary intracranial aneurysms in subarachnoid hemorrhage. *Neuroradiology* 2021;63:1985-1994
23. 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
24. Terasaki Y, Yokota H, Tashiro K, Maejima T, Takeuchi T, Kurosawa R, et al. Multidimensional deep learning reduces false-positives in the automated detection of cerebral aneurysms on time-of-flight magnetic resonance angiography: a multi-center study. *Front Neurol* 2022;12:742126
25. Wei X, Jiang J, Cao W, Yu H, Deng H, Chen J, et al. Artificial intelligence assistance improves the accuracy and efficiency of intracranial aneurysm detection with CT angiography. *Eur J Radiol* 2022;149:110169
26. Lehnen NC, Haase R, Schmeel FC, Vatter H, Dorn F, Radbruch A, et al. Automated detection of cerebral aneurysms on TOF-MRA using a deep learning approach: an external validation study. *AJNR Am J Neuroradiol* 2022;43:1700-1705
27. You W, Sun Y, Feng J, Wang Z, Li L, Chen X, et al. Protocol and preliminary results of the establishment of intracranial aneurysm database for artificial intelligence application based on

CTA images. *Front Neurol* 2022;13:932933

28. Heit JJ, Honce JM, Yedavalli VS, Baccin CE, Tatit RT, Copeland K, et al. RAPID Aneurysm: artificial intelligence for unruptured cerebral aneurysm detection on CT angiography. *J Stroke Cerebrovasc Dis* 2022;31:106690
29. Ham S, Seo J, Yun J, Bae YJ, Kim T, Sunwoo L, et al. Automated detection of intracranial aneurysms using skeleton-based 3D patches, semantic segmentation, and auxiliary classification for overcoming data imbalance in brain TOF-MRA. *Sci Rep* 2023;13:12018
30. Tajima T, Akai H, Yasaka K, Kunimatsu A, Yoshioka N, Akahane M, et al. Comparison of 1.5 T and 3 T magnetic resonance angiography for detecting cerebral aneurysms using deep learning-based computer-assisted detection software. *Neuroradiology* 2023;65:1473-1482
31. Liu X, Mao J, Sun N, Yu X, Chai L, Tian Y, et al. Deep learning for detection of intracranial aneurysms from computed tomography angiography images. *J Digit Imaging* 2023;36:114-123
32. 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
33. Ishihara M, Shiiba M, Maruno H, Kato M, Ohmoto-Sekine Y, Antoine C, et al. Detection of intracranial aneurysms using deep learning-based CAD system: usefulness of the scores of CNN's final layer for distinguishing between aneurysm and infundibular dilatation. *Jpn J Radiol* 2023;41:131-141
34. Wang J, Sun J, Xu J, Lu S, Wang H, Huang C, et al. Detection of intracranial aneurysms using multiphase CT angiography with a deep learning model. *Acad Radiol* 2023;30:2477-2486
35. Zhou Y, Yang Y, Fang T, Jia S, Nie S, Ye X. Joint two-stage convolutional neural networks for intracranial aneurysms detection on 3D TOF-MRA. *Phys Med Biol* 2023;68:185001
36. Colasurdo M, Shalev D, Robledo A, Vasandani V, Luna ZA, Rao AS, et al. Validation of an automated machine learning algorithm for the detection and analysis of cerebral aneurysms. *J Neurosurg* 2023;139:1002-1009
37. Hu B, Shi Z, Lu L, Miao Z, Wang H, Zhou Z, et al.; China Aneurysm AI Project Group. A deep-learning model for intracranial aneurysm detection on CT angiography images in China: a stepwise, multicentre, early-stage clinical validation study. *Lancet Digit Health* 2024;6:e261-e271
38. Adamchic I, Kantelhardt SR, Wagner HJ, Burbelko M. Artificial intelligence can help detecting incidental intracranial aneurysm on routine brain MRI using TOF MRA data sets and improve the time required for analysis of these images. *Neuroradiology* 2024;66:2195-2204
39. Bizjak Ž, Choi JH, Park W, Pernuš F, Špiclin Ž. Deep geometric learning for intracranial aneurysm detection: towards expert rater performance. *J Neurointerv Surg* 2024;16:1157-1162
40. Li Y, Zhang H, Sun Y, Fan Q, Wang L, Ji C, et al. Deep learning-based platform performs high detection sensitivity of intracranial aneurysms in 3D brain TOF-MRA: an external clinical validation study. *Int J Med Inform* 2024;188:105487
41. You W, Feng J, Lu J, Chen T, Liu X, Wu Z, et al. Diagnosis of intracranial aneurysms by computed tomography angiography using deep learning-based detection and segmentation. *J Neurointerv Surg* 2024;17:e132-e138
42. Wei J, Song X, Wei X, Yang Z, Dai L, Wang M, et al. Knowledge-augmented deep learning for segmenting and detecting cerebral aneurysms with CT angiography: a multicenter study. *Radiology* 2024;312:e233197
43. De Toledo OF, Gutierrez-Aguirre SF, Lara-Velazquez M, Qureshi AI, Camp W, Erazu F, et al. Use of artificial intelligence software to detect intracranial aneurysms: a comprehensive stroke center experience. *World Neurosurg* 2024;188:e59-e63
44. Goertz L, Jünger ST, Reinecke D, von Spreckelsen N, Shahzad R, Thiele F, et al. Deep learning-assistance significantly increases the detection sensitivity of neurosurgery residents for intracranial aneurysms in subarachnoid hemorrhage. *J Clin Neurosci* 2025;132:110971
45. Zhuo L, Zhang Y, Song Z, Mo Z, Xing L, Zhu F, et al. Enhancing radiologists' performance in detecting cerebral aneurysms using a deep learning model: a multicenter study. *Acad Radiol* 2025;32:1611-1620
46. Hu B, He H, Shi Z, Wang L, Liu Q, Sun Z, et al. Evaluating a clinically available artificial intelligence model for intracranial aneurysm detection: a multi-reader study and algorithmic audit. *Neuroradiology* 2025;67:855-864
47. Schmidt CC, Stahl R, Mueller F, Fischer TD, Forbrig R, Brem C, et al. Evaluation of AI-powered routine screening of clinically acquired cMRIs for incidental intracranial aneurysms. *Diagnostics (Basel)* 2025;15:254
48. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 2020;14:49-58
49. Kleppe A, Skrede OJ, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer* 2021;21:199-211
50. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800-809
51. Park SH, Han K, Jang HY, Park JE, Lee JG, Kim DW, et al. Methods

for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* 2023;306:20-31

52. Nam HH, Jang DK, Cho BR. Complications and risk factors after digital subtraction angiography: 1-year single-center study. *J Cerebrovasc Endovasc Neurosurg* 2022;24:335-340
53. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell* 2019;1:20-23
54. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean J Radiol* 2021;22:442-453
55. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al.; CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell* 2024;6:e240300
56. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* 2020;3:326-331
57. Liu H, Ding N, Li X, Chen Y, Sun H, Huang Y, et al. Artificial intelligence and radiologist burnout. *JAMA Netw Open* 2024;7: e2448714
58. Wekenborg MK, Gilbert S, Kather JN. Examining human-AI interaction in real-world healthcare beyond the laboratory. *NPJ Digit Med* 2025;8:169
59. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health* 2022;4:e384-e397

Supplementary Material 1. **Detailed scoring criteria for methodological assessment**

For Questions 1 to 4, a score of 0 indicates that a specific exclusion criterion was applied, while a score of 1 indicates that it was not. For Question 5, no scoring was assigned; instead, the concurrent findings excluded in each study were listed.

For Question 6, prospective studies were assigned a score of 1, and retrospective studies a score of 0. For Question 7, studies that performed external validation received a score of 1, whereas those that did not were assigned a score of 0. Question 8 evaluated the size of the test set. Studies that included more than 100 cases in either an independent test set or an external validation set were marked as a score of 1, whereas those assessing diagnostic performance in fewer than 100 cases received a score of 0. Questions 9 and 10 were applicable only to studies that performed external validation. For Question 9, a score of 1 was assigned if the scanners used in the external validation set differed from those used during model training, or if at least 3 different scanners were employed. For Question 10, a score of 1 indicated that the study used a diagnostic cohort for validation, while a score of 0 indicated that it did not.

For Questions 11 to 13, studies were scored 1 if the corresponding metric was reported and 0 if it was not. For Question 13, a score of 0.5 was given when the metric was not explicitly stated but could be inferred from the manuscript.