This article has been accepted for publication in JNNP following peer review.

The definitive copyedited, typeset version is available online at

Title: Diagnostic performance of machine learning models versus established risk stratification for intracranial aneurysm rupture: a systematic review and bivariate meta-analysis

Authors:

Shaan Patel[1]†, Shiva A Nischal[2]†*, Yi-Hein Chai[3], Angelette Mendonca[4], Kush M. Kale[2], James Castiglione[1], Pious Patel[1], Reid Gooch[1], Stavropoula I Tjoumakaris[1], Pascal Jabbour[1]

1. Department of Neurological Surgery, Thomas Jefferson University Hospital, Philadelphia, PA 19107, USA
2. Department of Physiology, Anatomy & Genetics, Medical Sciences Division, University of Oxford, OX1 3PT, UK
3. University College London Medical School, University College London, WC1E 6DE, UK
4. School of Clinical Medicine, University of Cambridge, CB2 0SP, UK

†**Contributed equally and share first-authorship**: Shaan Patel & Shiva A. Nischal

***Corresponding author:** Shiva A. Nischal (shiva.nischal@dpag.ox.ac.uk)

Abbreviations:

Area Under The Curve (AUC)

Deep Learning (DL)

Decision Tree (DT)

Diagnostic Odds Ratio (dOR)

Ensemble Method (EM)

K-Nearest Neighbour (KNN)

Machine Learning (ML)

Negative Likelihood Ratio (NLR)

Negative Predictive Value (NPV)

Positive Likelihood Ratio (PLR)

Positive Predictive Value (PPV)

Prediction Model Risk Of Bias Assessment Tool (PROBAST)

Regression Model (RM)

SHapley Additive exPlanations (SHAP)

Summary Receiver Operating Curve (SROC)

Support Vector Machine (SVM)

Keywords: Intracranial aneurysm; rupture risk; risk stratification; diagnostic accuracy; machine learning; predictive modelling; meta-analysis

Abstract

Background: Machine learning (ML) models have been proposed to improve discrimination of intracranial aneurysm rupture status beyond established clinical risk stratification tools. However, reported performance is heterogeneous and relative contribution of model architecture and feature dominance remains unclear.

Methods: We performed a PRISMA-DTA systematic review and diagnostic meta-analysis of studies evaluating ML models for intracranial aneurysm rupture discrimination. PubMed, Embase, and CENTRAL were searched to February 2026. Sensitivity and specificity were pooled using a bivariate random-effects model, with SROC curves generated across training, internal testing, and external validation datasets. Models were compared with regression-based approaches and PHASES scores. Subgroup and meta-regression analyses explored associations between algorithm family and feature domain.

Results: Sixty-two retrospective cohorts (29,709 patients, 209 models) met inclusion criteria. In training datasets, pooled sensitivity and specificity for ML were 0.81 (95% CI 0.75–0.85) and 0.83 (0.80–0.86), with AUC 0.878, exceeding PHASES (AUC 0.667). In testing datasets, ML retained higher discrimination (AUC 0.837) than regression models (0.806) and PHASES (0.646). In external validation, sensitivity was preserved (0.82), but specificity declined (0.66). Deep learning demonstrated the highest AUCs (training and testing). Incorporation of haemodynamic or radiomic features improved pooled discrimination relative to morphology alone. Evidence of small-study effects and mostly unclear PROBAST ratings were observed.

Conclusions: ML approaches demonstrate higher pooled discrimination for aneurysm rupture status than conventional risk scores in retrospective datasets but reduced external validation specificity and heterogeneity limit confidence for clinical translation. Prospective, externally-validated, calibrated models are required before integration into routine cerebrovascular risk stratification.

What is already known on this topic: Rupture risk estimation guides management of unruptured intracranial aneurysms. PHASES provides prospective risk stratification using limited clinical and anatomical variables. Machine learning models for rupture discrimination are increasingly reported, but external validation is inconsistent.

What this study adds: Across 62 cohorts (209 models; 29,709 patients), machine learning showed higher pooled discrimination for rupture status than regression models and PHASES in training and test datasets. Specificity declined in external validation, and most studies had unclear risk of bias. Deep learning and models incorporating haemodynamic or radiomic features yielded the highest point estimate performance.

How this study might affect research, practice, or policy: Higher retrospective discrimination does not yet establish reliable prospective risk prediction. External validation, calibration, and decision-threshold analysis should precede clinical adoption. Minimum methodological standards are needed before integration of machine learning into aneurysm risk stratification pathways.

Introduction

Aneurysmal subarachnoid haemorrhage remains among the most devastating forms of stroke, characterised by abrupt onset, high early mortality, and substantial long-term neurological morbidity.[1] Contemporary population-based estimates suggest a global incidence of approximately 6–8 per 100,000 person-years, with geographic heterogeneity and non-trivial case fatality despite advancements in aneurysm occlusion and neurocritical care.[2-4] Rupture prevention in patients with known intracranial aneurysms therefore remains a central objective of modern cerebrovascular practice. Yet most aneurysms are discovered in unruptured states, often incidentally.[5] Management is consequently defined by uncertainty: clinicians must weigh the probabilistic risk of future rupture against procedural risks of microsurgical or endovascular intervention.[6] Both overtreatment and undertreatment carry meaningful harm[7 8]. The challenge is therefore not simply technical, but predictive.

The PHASES score[9], incorporating population, hypertension, age, size, earlier subarachnoid haemorrhage, and site, represents the most widely adopted framework for estimating 5-year rupture risk.[6] Derived from pooled prospective cohorts, PHASES provides pragmatic, population-level risk stratification using readily available clinical and radiographic variables. However, rupture prediction remains imperfect because PHASES incorporates only a small number of variables and does not account for morphological, haemodynamic, or high-dimensional imaging features.[10] Moreover, aneurysm behaviour varies across populations and healthcare systems, raising legitimate transportability concerns for any single global score.

Against this backdrop, machine learning (ML) approaches have gained traction. Unlike conventional regression frameworks, ML systems are designed to learn patterns directly from data, accommodating non-linear relationships, high-dimensional feature spaces, and complex interactions.[11] Deep learning (DL) architectures can derive latent imaging representations without predefined feature engineering. Radiomic pipelines can extract hundreds to thousands of quantitative descriptors from standard angiography. Computational fluid dynamics can characterise local wall shear stress environments.[12 13] In theory, such models offer a means of integrating multiple domains within a single predictive framework.[14]

However, several conceptual tensions remain unresolved. First, many ML studies classify rupture status retrospectively rather than predict prospective rupture risk using pre-rupture data, potentially conflating discrimination with true prediction.[15] Second, reported performance is heterogenous and frequently lacks rigorous external validation.[13] Third, relatively few investigations benchmark ML systems directly against established clinical tools (including PHASES), limiting interpretability in real world decision-making contexts. Finally, improvements in area under the curve (AUC) do not automatically translate into clinically meaningful probability shifts at treatment thresholds.[6 8 16]

Accordingly, the central question is not whether machines can fit data more flexibly, but whether ML-based models demonstrate reproducible, externally valid, and clinically relevant discrimination relative to conventional risk stratification, recognising that these tools address different prediction targets.

Prior reviews report promising discrimination for ML-based rupture classification but emphasise heterogeneity, limited external validation, and unclear risk of bias.[17-19] In this PRISMA-DTA compliant systematic review and bivariate diagnostic meta-analysis, we synthesised evidence from studies evaluating ML models for intracranial aneurysm rupture discrimination to address these limitations. We aimed to quantify pooled diagnostic performance of ML approaches, compare ML-based models with regression-based models and the PHASES score, and explore whether model family and feature domain (morphological, haemodynamic, radiomic) are associated with differences in discrimination. By explicitly benchmarking the "machine" against the established clinical score, this study seeks to clarify the current evidentiary position of data-driven rupture modelling and define the methodological priorities required before translation into routine cerebrovascular practice.

Materials & Methods

This systematic review and diagnostic meta-analysis was conducted in accordance with PRISMA-DTA[20] guidelines and appraised using the AMSTAR-2[21] framework. A protocol was registered prospectively with PROSPERO (CRD420261304470).

Search strategy

We searched PubMed, Embase, and CENTRAL from database inception to February 2026. Search strategies combined controlled vocabulary (MeSH and Emtree terms) with free-text keywords relating to artificial intelligence, ML, DL, intracranial aneurysm, and rupture. Strategies were adapted for each database (Supplementary Table 1). Searches were limited to human studies published in English. Reference lists of included studies and relevant systematic reviews were manually screened to identify additional eligible articles.

Eligibility criteria

Studies were included if they evaluated ML models for predicting or discriminating intracranial aneurysm rupture risk, used real world clinical or imaging datasets, including imaging-derived or clinical variables as model inputs, and reported sufficient diagnostic performance data to permit diagnostic accuracy synthesis (including sensitivity, specificity, AUC, positive predictive value (PPV), negative predictive value (NPV), accuracy, or confusion-matrix elements). Studies were excluded if they used simulated or synthetic datasets only, did not report extractable performance metrics, inappropriately designed (case reports or series with less than 10 patients, conference abstracts, reviews, editorials, or commentaries), or non-English.

Study selection

Two reviewers independently screened titles and abstracts, followed by full-text assessment. Disagreements were resolved by consensus with a third reviewer. Inter-rater reliability was quantified using Cohen's kappa score ($\kappa = 0.91$).

Data extraction

Data was extracted independently using a standardised template. We recorded study characteristics (author, year, country), cohort characteristics (sample size, ruptured and unruptured counts, age, sex, aneurysm size, multiplicity, location, imaging modality), model characteristics (algorithm type, training and test and external validation splits, cross-validation strategy, feature domains, regularisation or tuning methods, model interpretability outputs including SHapley Additive exPlanations (SHAP) feature importance), and performance metrics (true positive, false positive, true negative, false negative, sensitivity, specificity, AUC, accuracy, PPV, NPV, diagnostic odds ratio

(dOR)) across training, internal testing, and external validation datasets. Feature domains were categorised as morphological, haemodynamic, radiomic/DL-derived, and clinical.

Statistical analysis

Analyses were performed using R (v4.3.2)[22]. Statistical significance was defined as $\alpha < 0.05$.

Diagnostic performance was synthesised using a bivariate random-effects (Reitsma) model, pooling sensitivity and specificity. From confusion-matrix data, pooled positive likelihood ratio (PLR), negative likelihood ratio (NLR), and dOR were calculated. Summary receiver operating characteristic (SROC) curves were generated, and AUC was reported as a global discrimination measure. To summarise the joint behaviour of prevalence-dependent metrics, we visualised accuracy against precision (PPV) using Bull's-eye plots, prespecifying a target performance accuracy zone > 0.80 and PPV > 0.80.

Models were grouped at two levels:

(i) Primary comparison: ML, regression models (RMs), PHASES.
(ii) Algorithm family comparison (Supplementary Table 2): DL, support vector machines (SVMs), ensemble methods (EMs), decision trees (DTs), k-nearest neighbours (KNNs), RMs, PHASES.

Analyses were stratified by dataset type (training, test, external validation). Given variability in decision thresholds and rupture prevalence across studies, SROC/AUC was prioritised as a threshold-robust summary measure. Prevalence-dependent metrics were interpreted in the context of included cohorts.

To explore sources of inter-study heterogeneity, we performed meta-regression within the bivariate framework, modelling transformed sensitivity and false-positive rate (FPR; 1 – specificity). DL was used as reference. Prespecified covariates included dataset type, algorithm family, and input feature domains. Regression coefficients (β) are reported with P-values.

Risk of bias and certainty of evidence

Risk of bias and applicability were assessed independently by two reviewers using PROBAST[23] scores across four domains: participants, predictors, outcome, and analysis. Certainty of evidence for primary diagnostic outcome was evaluated using GRADE[24]. Small-study effects were assessed using Deeks' funnel plot asymmetry test. Clinical utility was explored using Fagan nomograms to estimate post-test rupture probabilities following positive and negative results, based on pooled likelihood ratios.

Results

Study selection

Database searches yielded 660 records, of which 62 studies met inclusion criteria (Figure 1). In total, 209 models derived from 62 retrospective cohort studies comprising 29,709 patients with 36,749 aneurysms were included in the diagnostic test accuracy meta-analysis. Study sample sizes ranged from 28 to 3,915 patients.

Study characteristics

Studies originated most frequently from China (n=32), United States (n=12), and South Korea (n=5). Mean proportion of ruptured aneurysms was 41.0%. Mean age was 56.7 years in ruptured cohorts and 58.1 years in unruptured cohorts, and proportion of female patients was 62.0%. Mean aneurysm size was 5.4 mm, and aneurysm multiplicity was present in 29.0% of patients. Imaging modality was reported in 57/62 studies (91.9%); computed tomography angiography was the most common (31 studies; 53.4%), followed by digital subtraction angiography (29 studies; 50.0%) (Supplementary Table 3).

Modelling and validation practices

Only seven studies directly compared ML approaches and PHASES. Cross-validation was used for model optimisation in 44 studies. Median sample sizes were 311.5 for training (60 studies; range 36–3,312), 109.5 for internal testing (50 studies; range 9–1079), and 80.5 for external validation (18 studies; range 28–1,501) (Supplementary Table 4). SHAP-based feature importance analyses were reported in seven studies, and calibration of models were performed in nine studies (Supplementary Table 4).

Primary comparison: machine learning versus regression versus PHASES

Across datasets, pooled discrimination was highest for ML models, intermediate for RMs, and lowest for PHASES (Table 1). The primary pooled analyses included 60 training datasets, 50 test datasets, and 18 external validation datasets (Supplementary Table 5).

In training, ML achieved pooled sensitivity 0.81 (95% CI: 0.75–0.85) and specificity 0.83 (95% CI: 0.80–0.86), with AUC 0.878, dOR 20.30 (95% CI: 11.96–34.56), and accuracy 0.80 (95% CI: 0.79–0.80). RMs performed similarly with sensitivity 0.76 (95% CI: 0.64–0.86), specificity 0.84 (95% CI: 0.78–0.89), AUC 0.865, dOR 17.00 (95% CI: 6.43–45.38), and accuracy 0.81 (95% CI: 0.80–0.82). PHASES showed more modest discrimination with sensitivity 0.74 (95% CI: 0.44–0.91), specificity 0.57 (95% CI: 0.35–0.77), AUC 0.667, dOR 3.82 (95% CI: 0.41–35.58), and

accuracy 0.65 (95% CI: 0.62–0.68). Pooled likelihood ratios were higher for ML (PLR 4.76; NLR 0.24) and RM (PLR 4.78; NLR 0.28) than PHASES (PLR 1.73; NLR 0.45).

In testing, ML retained the highest overall discrimination (AUC 0.837), with sensitivity 0.74 (95% CI: 0.71–0.77) and specificity 0.79 (95% CI: 0.77–0.81). RM performance remained similar (AUC 0.806), with sensitivity 0.71 (95% CI: 0.66–0.76) and specificity 0.81 (95% CI: 0.76–0.85). PHASES performed substantially worse (AUC 0.646), with sensitivity 0.65 (95% CI: 0.32–0.87) and specificity 0.58 (95% CI: 0.38–0.76). Pooled likelihood ratios again favoured ML (PLR 3.51; NLR 0.33) and RM (PLR 3.75; NLR 0.35) over PHASES (PLR 1.54; NLR 0.61).

In external validation, ML and RM showed similar sensitivity (0.82 (95% CI: 0.77–0.87) versus 0.83 (95% CI: 0.70–0.91)) but lower specificity (0.66 (95% CI: 0.63–0.70) versus 0.65 (95% CI: 0.59–0.70)). PHASES remained modest (sensitivity 0.68 (95% CI: 0.45–0.84), specificity 0.58 (95% CI: 0.46–0.68)). The external validation AUC for ML was 0.806, compared with 0.778 for RM and 0.667 for PHASES. Across all datasets combined, pooled AUCs were 0.86 for ML, 0.82 for RM, and 0.66 for PHASES (Figure 2A–D).

Algorithm-family performance (subgroup synthesis)

Algorithm-family analyses are summarised in Table 4. Across datasets, AUC point estimates ranged from 0.764 (KNN; n=2) to 0.899 (DL; n=12) in training, 0.758 (KNN; n=7) to 0.875 (DL; n=12) in testing, and 0.677 (KNN; n=3) to 0.840 (DT; n=1) in external validation. DL showed the strongest discrimination in training and testing (AUC 0.899, 0.875), supported by high dORs (training 29.53, testing 21.14). Notably, DL models were more frequently trained on high-dimensional imaging-derived or radiomic inputs, whereas classical algorithms were typically applied to manually extracted morphological or clinical features. EMs and SVMs demonstrated balanced performance in testing (both AUC 0.846), with broadly similar sensitivity and specificity profiles. DTs exhibited lower test-set performance (AUC 0.760), and the external validation estimate (AUC 0.840) was based on a single model, characterised by high sensitivity (0.98) and poor specificity (0.39). KNN models showed the lowest AUC estimates and widest uncertainty (Figures 2E–H).

Accuracy-precision trade-off (bull's-eye plots)

Bull's-eye plots (Figure 3) summarised the joint behaviour of accuracy and PPV, using a predefined target zone accuracy > 0.80 and PPV > 0.80. In training, only DL (accuracy 0.84; PPV 0.84) and DT (accuracy 0.81; PPV 0.84) met the target, whereas other approaches fell below threshold largely to reduced PPV. In testing, only DL remained within the target zone (accuracy 0.81; PPV 0.83), while PPV fell for most other approaches.

Feature domain analyses: haemodynamic, radiomics, and morphology

Parameter-domain SROC analyses (Figure 4) demonstrated highest discrimination for models incorporating both morphological and haemodynamic inputs (AUC 0.854; sensitivity 0.788; specificity 0.786), followed by radiomic-based models (AUC 0.842; sensitivity 0.737; specificity 0.803). Models using morphology alone performed less well (AUC 0.814; sensitivity 0.776; specificity 0.729). All feature-domain strata showed substantial heterogeneity, reflected by dispersion around SROC curves.

In dataset-stratified analyses, haemodynamic inputs were associated with improved specificity in the test set ($P < 0.05$) but not in external validation ($P = 0.35$). Sensitivity did not differ significantly by haemodynamic inclusion in either test ($P = 0.10$). Accuracy improved in the test set ($P < 0.01$) but not in external validation ($P = 0.78$).

Among seven studies reporting SHAP, size-related (5 of 7 studies) and aspect ratio features (3 of 7 studies) were most frequently ranked as important.

Meta-regression

Meta-regression of transformed sensitivity and transformed FPR used DL as the reference. PHASES showed lower sensitivity (β -0.610; $P < 0.05$) and higher FPR (β +0.914; $P < 0.001$) relative to DL. Compared with DL, EMs showed lower sensitivity (β -0.522; $P < 0.01$).

Regarding inputs, inclusion of morphological parameters was associated with higher sensitivity (β +0.540; $P < 0.01$), whereas inclusion of haemodynamic (β -0.418; $P < 0.001$) and radiomic features (β -0.325; $P < 0.05$) was associated with a lower FPR. Finally, FPR was higher in test (β +0.233; $P < 0.05$) and external validation datasets (β +0.867; $P < 0.001$) relative to training.

Clinical utility (Fagan nomograms)

Using a pre-test probability of 20.0%, Fagan nomograms (Supplementary Figure 1) showed larger post-test probability shifts for ML and RM than PHASES across training, test, and validation datasets.

Publication bias, risk of bias, and certainty of evidence

Deeks' tests suggested small-study effects/publication bias in training ($P < 0.001$) and testing ($P < 0.001$), but not in external validation ($P = 0.48$) (Supplementary Figure 2). Overall, 57/62 studies were judged to have an unclear (moderate) risk of bias, with none rated as serious risk (Supplementary Table 6). Certainty of evidence for primary diagnostic outcomes was moderate, with downgrading driven primarily by risk of bias (Supplementary Table 7).

Discussion

Principal findings of the study

In this PRISMA-DTA systematic review and bivariate meta-analysis of 62 retrospective cohorts (209 models across 29,709 patients), ML models demonstrated higher pooled discrimination for aneurysm rupture status than RMs and PHASES across training and test datasets, with attenuation in external validation driven primarily by reduced specificity. Importantly, this comparison should be interpreted within the caveat that ML models and PHASES address non-equivalent prediction targets: the former predominantly classify retrospective rupture status, whereas PHASES estimates prospective rupture risk from longitudinal natural history cohorts[9]. Algorithm-family syntheses suggested the strongest point-estimate discrimination for DL in training and testing, while performance in external validation was more variable and based on limited model counts. Feature domain analyses indicated incorporating haemodynamic or radiomic information improved discrimination relative to morphology alone, with meta-regression suggesting a trade-off in which morphology was more closely associated with sensitivity, while haemodynamic and radiomic domains were associated with reduced FPRs (higher specificity). Finally, although pooled likelihood ratios produced larger post-test probability shifts for ML and RMs than PHASES, evidence of small-study effects and unclear PROBAST ratings temper confidence in apparent performance gains.

Why machine learning tends to outperform PHASES for "rupture status"

The observed separation between ML-derived models and PHASES should be interpreted in light of differences in prediction target and feature representation. PHASES estimates future rupture risk and was developed from prospective cohorts using a restricted set of clinical and anatomical predictors.[9] By contrast, most ML studies were trained on cross-sectional/retrospective datasets labelled by rupture status. These are not equivalent. Longitudinal natural history cohorts (including ISUIA and UCAS) were explicitly designed to estimate prospective rupture risk rather than classifying rupture status.[6 8 16] A model can achieve strong discrimination by learning features that correlate with rupture having occurred (including post-rupture geometric changes, selection effects in which aneurysms are imaged only after rupture, or treatment/referral biases), without providing stable prospective risk prediction.[10] This distinction likely explains consistent separations between ML/RM and PHASES in pooled AUC, while also explaining why specificity often degrades in external validation: cross-sectional correlates of rupture may not transport well across institutions, imaging protocols, and case-mix.

Mechanistic interpretation: from morphology to wall mechanics and rupture biology

Rupture risk is ultimately determined by balance between local haemodynamic loading and time-dependent wall weakening arising from maladaptive remodelling.[12] In this framework, commonly

used morphological descriptors (size, aspect ratio, irregularity, lobulation) can be viewed as integrative surrogates of cumulative remodelling, reflecting the net effect of growth dynamics, intraluminal flow organisation, and wall response over time.[25] This may help explain why morphological features were frequently prioritised in explainability analyses and why, in meta-regression, inclusion of morphological parameters was associated with higher pooled sensitivity.

By contrast, haemodynamic and radiomic feature domains plausibly contribute to complementary information that is less captured by gross geometry alone. Computational haemodynamic indices are candidate markers of flow environments implicated in endothelial dysfunction and downstream remodelling pathways, whereas radiomic and deep feature embeddings may encode higher-order shape, texture, and intensity patterns not captured by conventional morphometrics.[12 14] In this review, haemodynamic and radiomic inclusion was associated with lower FPR in meta-regression, consistent with a role in refining discrimination along aneurysms with overlapping morphological profiles. However, haemodynamic gains were not consistently retained in external validation, which is compatible with known sensitivity of haemodynamic pipelines to image acquisition, segmentation, and therefore underscores the importance of standardisation and multicentre transportability assessments.[13 15]

Algorithm-family patterns: performance, but uncertain clinical meaning

Beyond feature biology, model architecture itself warrants consideration. DL achieved highest pooled training and testing AUCs, with EMs and SVMs demonstrating broadly comparable performance, whereas DTs and KNNs showed greater variability. However, these gradients should not be interpreted as intrinsic algorithm superiority. Performance is contingent on feature representation: DL models frequently leverage high-dimensional imaging or radiomic embeddings, whereas classical algorithms are often trained on morphological and clinical predictors, creating a structural confound between algorithm type and input feature dimensionality. Apparent gains may therefore reflect input dimensionality rather than model architecture.[14] The independent contributions of algorithm architecture and input feature dimensionality cannot be disentangled within the current study-level meta-analytic framework. In addition, most studies relied on retrospective datasets with internal optimisation and limited external validation. Intra-dataset tuning can inflate discrimination, and external validation stratum in this evidence base is comparatively sparse and occasionally driven by single models. Consequently, current data does not establish consistent algorithm-level superiority across heterogeneous clinical settings. Finally, discrimination does not directly translate to clinical utility. Given rupture prevalence varied across cohorts, prevalence-dependent measures such as PPV and accuracy shift even when sensitivity and specificity remain stable. Algorithm comparison is therefore meaningful only when interpreted alongside validation design, case-mix, and calibration.

Interpretability and biological plausibility

Only a minority of studies reported SHAP-based feature importance analyses. Interpretability is not ancillary in this context, but essential for determining whether models are capturing biologically plausible determinants of rupture versus artefact. Among these studies, size-related and aspect ratio metrics were most frequently ranked as influential predictors. These variables are established morphological correlates of rupture and may function as integrative markers of cumulative remodelling. Their prominence in data-driven models suggests convergence with existing pathophysiological understanding rather than discovery of entirely novel determinants.[25] However, feature ranking alone does not establish validity. A model may prioritise plausible variables yet remain poorly calibrated or threshold-dependent. Interpretability therefore requires pairing with calibration assessment and external validation before clinical translation can be considered.

Clinical translation: beyond discrimination

High AUC values do not solely justify clinical implementation. Clinical decision-making in unruptured aneurysms depends on calibrated absolute risk estimates, clearly defined probability thresholds for intervention, and explicit time horizons. Contemporary guidelines[7] emphasise individualised decision-making that integrates patient age, comorbidity burden, aneurysm size/location, and procedural risk, recognising that preventive intervention carries non-trivial morbidity and may not be justified for lesions with low absolute rupture risk.[6 8 16] Most included studies were designed to discriminate rupture status retrospectively rather than predict prospective rupture, which requires modelling temporal growth, remodelling, and competing risks. Few studies reported calibration metrics or decision-curve analyses, and prospective longitudinal validation was uncommon. The Fagan analyses illustrate potential post-test probability shifts under assumed pre-test probabilities but do not establish clinical benefit. Demonstrating clinical utility will require prospective, externally validated models that provide calibrated risk estimates aligned with treatment thresholds.

Limitations of the evidence base

The literature is dominated by retrospective cohorts with heterogenous inclusion criteria, imaging protocols, and rupture definitions. Many studies relied on internal cross-validation without prespecified external validation, and reporting of calibration, missing data handling, and threshold selection was inconsistent, consistent with predominantly unclear PROBAST ratings. Evidence of small-study effects in training and testing analyses suggests possible inflation of pooled performance estimates. External validation datasets were comparatively limited in number and size, constraining precision and limiting robust algorithm-family comparisons. In addition, variation in prevalence and case-mix across cohorts affects prevalence-dependent metrics, reinforcing the importance of

prioritising externally validated sensitivity and specificity when assessing transportability. This is further compounded by the predominance of studies from Chinese cohorts which may limit generalisability as rupture prevalence, imaging protocols, and treatment thresholds vary across populations and healthcare systems. Finally, the study-level design precludes separation of the independent effects of algorithm architecture and input feature dimensionality, which frequently co-vary across studies.

Future directions

Progress will depend on three methodological advances. First, multicentre datasets with harmonised imaging acquisition, segmentation, and feature extraction pipelines are required to improve transportability. Cross-population validation efforts have demonstrated that statistical learning approaches can be extended across national cohorts, but only when feature harmonisation is addressed.[13] Recent multicentre modelling efforts illustrate both the feasibility and methodological requirements of such approaches.[25] Second, models should be reframed towards prospective rupture prediction anchored to longitudinal imaging and growth trajectories rather than rupture-status classification. Third, integration of morphological evolution, haemodynamic loading, and patient-level factors into mechanistically informed frameworks may improve biological fidelity and interpretability.

Conclusion

In retrospective cohorts, ML approaches demonstrate higher pooled discrimination for aneurysm rupture status than RMs and PHASES. However, reduced specificity in external validation, evidence of small-study effects, and methodological heterogeneity indicate that current performance should be regarded as exploratory rather than practice-changing. Prospective, externally validated, and calibrated models are required before integration into routine cerebrovascular risk stratification. These findings support continued investigation of ML approaches for aneurysm rupture discrimination, with the critical caveat that retrospective classification does not establish prospective predictive validity.

Statements & Declarations

Ethics approval: This study is a meta-analysis of published RCTs. Ethical approval was obtained for each included RCT by their respective Ethics Committees/Institutional Review Boards. As such, a new ethics approval was not required for this study. No direct interactions with participants took place, no new data was collected. Informed consent was obtained from all included patients by their respective authors.

Consent to participate: Not applicable.

Consent to publish: Not applicable.

Competing interests: Dr. Jabbour is a consultant for Medtronic, MicroVention, Balt and Cerus Endovascular. Dr. Tjoumakaris is a consultant for MicroVention. Dr. Gooch is a consultant for Stryker.

Authors' contributions: SP and SAN had full access to all the data in this study and takes responsibility for the integrity of the data and accuracy of the data analysis. SP and SAN contributed equally to this work. Concept and design (SP, SAN, MRG, SIJ, PMJ); search strategy (SP, SAN); extraction of data (SP, SAN, YHC, AM, KMK); analysis and interpretation of data (SP, SAN, YHC, AM, KMK, JG); drafting the paper (SP, SAN, KP, PDP, MRG, SIJ, PMJ); final manuscript review (all authors); guarantor (SAN).

Funding: There was no funding provided for this research.

Availability of data and materials: Not applicable.

References

1. Rinkel GJ, Algra A. Long-term outcomes of patients with aneurysmal subarachnoid haemorrhage. *Lancet Neurol* 2011;10(4):349–56. doi: 10.1016/S1474-4422(11)70017-5
2. de Rooij NK, Linn FH, van der Plas JA, et al. Incidence of subarachnoid haemorrhage: a systematic review with emphasis on region, age, gender and time trends. *J Neurol Neurosurg Psychiatry* 2007;78(12):1365–72. doi: 10.1136/jnnp.2007.117655 [published Online First: 20070430]
3. Etminan N, Chang HS, Hackenberg K, et al. Worldwide Incidence of Aneurysmal Subarachnoid Hemorrhage According to Region, Time Period, Blood Pressure, and Smoking Prevalence in the Population: A Systematic Review and Meta-analysis. *JAMA Neurol* 2019;76(5):588–97. doi: 10.1001/jamaneurol.2019.0006
4. Nieuwkamp DJ, Setz LE, Algra A, et al. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *Lancet Neurol* 2009;8(7):635–42. doi: 10.1016/S1474-4422(09)70126-7 [published Online First: 20090606]
5. Etminan N, Ruigrok YM, Hackenberg KAM, et al. Epidemiology, pathogenesis, and emerging concepts in unruptured intracranial aneurysms. *Lancet Neurol* 2025;24(11):945–57. doi: 10.1016/S1474-4422(25)00264-9
6. Wiebers DO, Whisnant JP, Huston J, 3rd, et al. Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment. *Lancet* 2003;362(9378):103–10. doi: 10.1016/s0140-6736(03)13860-3
7. Thompson BG, Brown RD, Jr., Amin-Hanjani S, et al. Guidelines for the Management of Patients With Unruptured Intracranial Aneurysms: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke* 2015;46(8):2368–400. doi: 10.1161/STR.0000000000000070 [published Online First: 20150618]
8. International Study of Unruptured Intracranial Aneurysms I. Unruptured intracranial aneurysms--risk of rupture and risks of surgical intervention. *N Engl J Med* 1998;339(24):1725–33. doi: 10.1056/NEJM199812103392401
9. Greving JP, Wermer MJ, Brown RD, Jr., 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(1):59–66. doi: 10.1016/S1474-4422(13)70263-1 [published Online First: 20131127]
10. Kleinloog R, de Mul N, Verweij BH, et al. Risk Factors for Intracranial Aneurysm Rupture: A Systematic Review. *Neurosurgery* 2018;82(4):431–40. doi: 10.1093/neuros/nyx238
11. Shi Z, Hu B, Schoepf UJ, et al. Artificial Intelligence in the Management of Intracranial Aneurysms: Current Status and Future Perspectives. *AJNR Am J Neuroradiol* 2020;41(3):373–79. doi: 10.3174/ajnr.A6468 [published Online First: 20200312]

12. Frosen J, Cebral J, Robertson AM, et al. Flow-induced, inflammation-mediated arterial wall remodeling in the formation and progression of intracranial aneurysms. *Neurosurg Focus* 2019;47(1):E21. doi: 10.3171/2019.5.FOCUS19234
13. Detmer FJ, Hadad S, Chung BJ, et al. Extending statistical learning for aneurysm rupture assessment to Finnish and Japanese populations using morphology, hemodynamics, and patient characteristics. *Neurosurg Focus* 2019;47(1):E16. doi: 10.3171/2019.4.FOCUS19145
14. Berg P, Saalfeld S, Voss S, et al. A review on the reliability of hemodynamic modeling in intracranial aneurysms: why computational fluid dynamics alone cannot solve the equation. *Neurosurg Focus* 2019;47(1):E15. doi: 10.3171/2019.4.FOCUS19181
15. Maroufi SF, Pachon-Londono MJ, Ghoche M, et al. Machine Learning-Based Rupture Risk Prediction for Intracranial Aneurysms: A Systematic Review and Meta-Analysis. *Neurosurgery* 2025;97(5):1072–82. doi: 10.1227/neu.0000000000003531 [published Online First: 20250530]
16. Investigators UJ, Morita A, Kirino T, et al. The natural course of unruptured cerebral aneurysms in a Japanese cohort. *N Engl J Med* 2012;366(26):2474–82. doi: 10.1056/NEJMoa1113260
17. Daga K, Agarwal S, Moti Z, et al. Machine Learning Algorithms to Predict the Risk of Rupture of Intracranial Aneurysms: a Systematic Review. *Clin Neuroradiol* 2025;35(1):3–16. doi: 10.1007/s00062-024-01474-4 [published Online First: 20241115]
18. Zhong J, Jiang Y, Huang Q, et al. Diagnostic and predictive value of radiomics-based machine learning for intracranial aneurysm rupture status: a systematic review and meta-analysis. *Neurosurg Rev* 2024;47(1):845. doi: 10.1007/s10143-024-03086-5 [published Online First: 20241112]
19. Shu Z, Chen S, Wang W, et al. Machine Learning Algorithms for Rupture Risk Assessment of Intracranial Aneurysms: A Diagnostic Meta-Analysis. *World Neurosurg* 2022;165:e137–e47. doi: 10.1016/j.wneu.2022.05.117 [published Online First: 20220608]
20. Mcinnes FDM, Moher D, Thombs DB, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies. *JAMA* 2018;319(4):388. doi: 10.1001/jama.2017.19163
21. Shea JB, Reeves CB, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* 2017:j4008. doi: 10.1136/bmj.j4008
22. R: A language and environment for statistical computing [program]. 4.3.2 version, 2023.
23. Wolff FR, Moons GMK, Riley DR, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine* 2019;170(1):51–58. doi: 10.7326/M18-1376

24. Schünemann JH, Oxman DA, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336(7653):1106–10. doi: 10.1136/bmj.39500.677199.ae
25. Fujimura S, Yanagisawa T, Kudo G, et al. Development and Validation of a Prediction Model for Intracranial Aneurysm Rupture Risk. *JAMA Netw Open* 2025;8(12):e2550772. doi: 10.1001/jamanetworkopen.2025.50772 [published Online First: 20251201]

Figure Legends

Figure 1: PRISMA flow diagram of study selection

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram outlining the study selection process. The number of records identified, screened, assessed for eligibility, and included in the final analysis are detailed at each stage.

Figure 2: Predictive performance of machine learning, regression, and PHASES

Summary receiver operating characteristic (SROC) curves comparing the performance of PHASES score with alternative predictive approaches across (A-B) training, (C-D) testing, (E-F) external validation, and (G-H) overall. Panels A, C, E and G compare the performance of PHASES (solid red line) with machine learning (ML)-based (dashed blue line) and regression models (dotted green line). Panels B, D, F, and H further compare PHASES with individual ML model subtypes. Individual study estimates of sensitivity and false positive rate are plotted as symbols, with corresponding SROC curves overlaid. Abbreviations: RM, regression models; DL, deep learning; DT, decision tree; EM, ensemble methods; KNN, K-nearest neighbours; SVM, support vector machine; PH, PHASES score (Population, Hypertension, Age, Size of aneurysm, Earlier subarachnoid haemorrhage, Site of aneurysm).

Figure 3: Bull's-eye plots of model performance

Bull's-eye plots showing the relationship between accuracy and positive predictive value (PPV) for different predictive approaches. Dashed vertical and horizontal lines indicate the predefined performance threshold (accuracy > 0.80 and precision > 0.80), with the shaded area representing the target performance zone. (A) Summary bull's-eye plot showing overall performance across training, validation, and test datasets. (B) Bull's-eye plot for the training dataset. (C) Bull's-eye plot for the independent test dataset. Abbreviations: ML, machine learning; RM, regression models; DL, deep learning; DT, decision tree; EM, ensemble methods; KNN, K-nearest neighbours; SVM, support vector machine; PHASES, PHASES score (Population, Hypertension, Age, Size of aneurysm, Earlier subarachnoid haemorrhage, Site of aneurysm).

Figure 4: Predictive performance of models incorporating different parameters

Summary receiver operating characteristic (SROC) curves comparing models incorporating morphological parameters alone (blue dashed line), combined morphological and haemodynamic parameters (green dashed line), and radiomic parameters (purple dashed line). Individual study estimates of sensitivity and false positive rate are plotted as symbols, with corresponding SROC curves overlaid. Abbreviations: AUC, area under the curve.

Tables

Algorithm	Set	Sensitivity	Specificity	DOR	PLR	NLR	Accuracy	AUC
ML	Train	0.81 (0.75–0.85)	0.83 (0.80–0.86)	20.30 (11.96–34.56)	4.76 (3.70–6.10)	0.24 (0.18–0.31)	0.80 (0.79–0.80)	0.878
RM	Train	0.76 (0.64–0.86)	0.84 (0.78–0.89)	17.00 (6.43–45.38)	4.78 (2.95–7.44)	0.28 (0.16–0.46)	0.81 (0.80–0.82)	0.865
PHASES	Train	0.74 (0.44–0.91)	0.57 (0.35–0.77)	3.82 (0.41–35.58)	1.73 (0.67–3.97)	0.45 (0.11–1.63)	0.65 (0.62–0.68)	0.667
ML	Test	0.74 (0.71–0.77)	0.79 (0.77–0.81)	10.60 (7.89–14.23)	3.51 (3.01–4.10)	0.33 (0.29–0.38)	0.77 (0.76–0.78)	0.837
RM	Test	0.71 (0.66–0.76)	0.81 (0.76–0.85)	10.59 (6.26–18.03)	3.75 (2.77–5.12)	0.35 (0.28–0.44)	0.78 (0.77–0.79)	0.806
PHASES	Test	0.65 (0.32–0.87)	0.58 (0.38–0.76)	2.53 (0.29–22.13)	1.54 (0.52–3.68)	0.61 (0.17–1.80)	0.58 (0.56–0.60)	0.646
ML	Validation	0.82 (0.77–0.87)	0.66 (0.63–0.70)	9.15 (5.63–14.87)	2.44 (2.06–2.86)	0.27 (0.19–0.37)	0.71 (0.70–0.73)	0.806
RM	Validation	0.83 (0.70–0.91)	0.65 (0.59–0.70)	9.12 (3.33–24.84)	2.36 (1.70–3.07)	0.26 (0.12–0.51)	0.71 (0.67–0.74)	0.778
PHASES	Validation	0.68 (0.45–0.84)	0.58 (0.46–0.68)	2.81 (0.70–11.36)	1.59 (0.84–2.66)	0.57 (0.23–1.19)	0.62 (0.54–0.71)	0.667

Table 1: Pooled discrimination metrics.

Summary of pooled performance metrics for machine learning (ML) models, regression models (RM), and PHASES score across training, testing, and validation datasets. Abbreviations: DOR, diagnostic odds ratio; PLR, positive likelihood ratio; NLR, negative likelihood ratio; AUC, area under the curve.

Algorithm (n)	Set	Sensitivity	Specificity	DOR	PLR	NLR	Accuracy	AUC
DL (12)	Train	0.85 (0.77–0.91)	0.84 (0.78–0.89)	29.53 (11.60–75.05)	5.28 (3.47–7.89)	0.18 (0.11–0.30)	0.84 (0.83–0.85)	0.899
DT (3)	Train	0.81 (0.41–0.96)	0.86 (0.67–0.95)	24.90 (1.40–449.23)	5.64 (1.23–18.48)	0.23 (0.04–0.88)	0.81 (0.79–0.82)	0.888
EM (15)	Train	0.78 (0.73–0.82)	0.83 (0.76–0.88)	16.87 (8.48–33.23)	4.52 (3.05–6.74)	0.27 (0.20–0.36)	0.78 (0.77–0.79)	0.878
KNN (2)	Train	0.91 (0.05–0.99)	0.39 (0.00–0.99)	6.49 (0.00–332001.00)	1.49 (0.05–333.00)	0.23 (0.00–946.00)	0.79 (0.77–0.81)	0.764
PHASES (2)	Train	0.74 (0.44–0.91)	0.57 (0.35–0.77)	3.82 (0.41–35.58)	1.73 (0.67–3.97)	0.45 (0.11–1.63)	0.65 (0.62–0.68)	0.667
RM (10)	Train	0.76 (0.64–0.86)	0.84 (0.78–0.89)	17.00 (6.43–45.38)	4.78 (2.95–7.44)	0.28 (0.16–0.46)	0.81 (0.80–0.82)	0.865
SVM (7)	Train	0.74 (0.64–0.82)	0.83 (0.76–0.88)	13.59 (5.51–33.27)	4.31 (2.65–6.87)	0.32 (0.21–0.48)	0.80 (0.79–0.81)	0.879
DL (12)	Test	0.81 (0.73–0.87)	0.83 (0.78–0.88)	21.14 (9.51–47.39)	4.87 (3.27–7.22)	0.23 (0.15–0.34)	0.81 (0.79–0.83)	0.875
DT (6)	Test	0.64 (0.47–0.78)	0.72 (0.58–0.83)	4.70 (1.22–18.11)	2.32 (1.12–4.70)	0.49 (0.26–0.92)	0.71 (0.68–0.75)	0.76
EM (37)	Test	0.73 (0.69–0.76)	0.79 (0.75–0.82)	10.00 (6.80–14.60)	3.44 (2.78–4.25)	0.34 (0.29–0.41)	0.77 (0.76–0.78)	0.846
KNN (7)	Test	0.70 (0.50–0.84)	0.77 (0.49–0.92)	7.57 (0.98–58.55)	2.99 (0.99–10.04)	0.40 (0.17–1.01)	0.76 (0.74–0.78)	0.758
PHASES (5)	Test	0.65 (0.32–0.87)	0.58 (0.38–0.76)	2.53 (0.29–22.13)	1.54 (0.52–3.68)	0.61 (0.17–1.80)	0.58 (0.56–0.60)	0.646
RM (14)	Test	0.71 (0.66–0.76)	0.81 (0.76–0.85)	10.59 (6.26–18.03)	3.75 (2.77–5.12)	0.35 (0.28–0.44)	0.78 (0.77–0.79)	0.806
SVM (19)	Test	0.77 (0.70–0.82)	0.79 (0.73–0.83)	12.02 (6.53–22.09)	3.57 (2.64–4.82)	0.30 (0.22–0.40)	0.77 (0.76–0.78)	0.846
DL (6)	Validation	0.75 (0.59–0.86)	0.80 (0.72–0.86)	11.90 (3.66–38.38)	3.77 (2.10–6.27)	0.32 (0.16–0.58)	0.74 (0.71–0.77)	0.801
DT (1)	Validation	0.98 (NA–NA)	0.39 (NA–NA)	31.20 (NA–NA)	1.60 (NA–NA)	0.05 (NA–NA)	0.83 (NA–NA)	0.84
EM (15)	Validation	0.83 (0.76–0.88)	0.65 (0.60–0.69)	8.72 (4.76–15.87)	2.35 (1.91–2.84)	0.27 (0.18–0.40)	0.71 (0.69–0.73)	0.833
KNN (3)	Validation	0.81 (0.66–0.91)	0.62 (0.55–0.69)	7.17 (2.36–21.97)	2.15 (1.46–2.93)	0.30 (0.13–0.62)	0.63 (0.58–0.69)	0.677
PHASES (3)	Validation	0.68 (0.45–0.84)	0.58 (0.46–0.68)	2.81 (0.70–11.36)	1.59 (0.84–2.66)	0.57 (0.23–1.19)	0.62 (0.54–0.71)	0.667
RM (5)	Validation	0.83 (0.70-0.91)	0.65 (0.59–0.70)	9.12 (3.33–24.84)	2.36 (1.70–3.07)	0.26 (0.12–0.51)	0.71 (0.67–0.74)	0.778
SVM (8)	Validation	0.83 (0.69–0.92)	0.64 (0.56–0.72)	8.96 (2.88–27.77)	2.34 (1.58–3.25)	0.26 (0.12–0.55)	0.71 (0.68–0.74)	0.807

Table 2: Model subtype discrimination metrics.

Summary of performance metrics for machine learning model subtypes, regression models (RM), and PHASES score across training, testing, and validation datasets. Abbreviations: n, number of models; DOR, diagnostic odds ratio; PLR, positive likelihood ratio; NLR, negative likelihood ratio; AUC, area under the curve; DL, deep learning; DT, decision tree; EM, ensemble methods; KNN, k-nearest neighbors; SVM, support vector machine.