RESEARCH PAPER

Machine learning in Vascular and Endovascular surgery: a systematic review and critical appraisal

Ahmed A.F Osman

Applied College, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia, afadol@kfu.edu.sa, ORCID (https://orcid.org/0009-0001-1362-4942)

Abstract:

Background: The field of machine learning (ML) shows great potential to enhance the process of predicting treatment results after patients undergo vascular and endovascular medical procedures. The evaluation process for new models needs to follow standardized procedures which must be rigid to establish their readiness for operational deployment.

Methods: The research used PRISMA 2020 to conduct a systematic review which combined clinical ML prediction models that forecasted surgical results. The authors used TRIPOD-AI and PROBAST-AI frameworks to assess both the reporting standards and research design of the study.

Results: The results from 50 studies showed that ML models achieved excellent discrimination ability with an average AUC value of 0.86. The research revealed two major weaknesses in the studies because calibration plots appeared in less than 36% of the studies and external validation occurred in only 40% of the research. The research used decision curve analysis as its main clinical utility assessment tool although this method appeared in less than 10% of all studies.

Conclusions: The current ML models achieve good predictive results, but their biased operation combined with insufficient disclosure of validation procedures and medical effectiveness assessment makes them inappropriate for medical use. The future development process needs to focus on three essential elements which include external validation testing and complete calibration evaluation and full compliance with TRIPOD-AI/PROBAST-AI reporting standards.

Keywords: Machine learning; vascular surgery; endovascular procedures; predictive modelling; TRIPOD-AI; PROBAST.

How to cite this article: Ahmed A.F Osman | Machine learning in Vascular and Endovascular surgery: a systematic review and critical appraisal. Int J Drug Deliv Technol. 2026;16(1): 215-230. DOI: 10.25258/ijddt.16.1.23

Source of support: Nil.

Conflict of interest: None

1. Introduction

The treatment of deadly vascular conditions including aortic aneurysm and carotid atherosclerosis and peripheral arterial disease depends on vascular and endovascular procedures. The field of perioperative care together with imaging and device technology advancements has not eliminated the risk of complications which occur during medical procedures [1], [2]. The process of predicting preoperative and perioperative risks needs to be done with precision because it helps doctors choose the right patients for treatment and plan their care and minimize the risk of complications. The Society for Vascular Surgery (SVS) risk score and Vascular Quality Initiative (VQI) registry tools and Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity (POSSUM) serve as traditional risk prediction models which help healthcare providers perform risk assessments [3]–[5]. The models depend on linear models which use only a few clinical indicators to predict patient outcomes, but these restrictions limit their ability to work with different types of patients [6].

The increasing number of big clinical datasets containing high-dimensional information from electronic health records and imaging repositories and vascular registries allows machine learning (ML) to perform predictive analytics in vascular surgery [7], [8]. ML models which include Random Forest and XGBoost and deep neural networks enable the detection of complex non-linear relationships between features which standard regression methods fail to identify [9]. The methods have shown better performance in both discrimination and calibration for cardiac surgery and oncologic surgery and orthopedic surgery, and they are now used for predicting vascular outcomes [13].

The ability to predict treatment results in this situation directly affects how doctors should inform patients about their surgical risks and which treatment method to choose between endovascular and open procedures and how often

Author for Correspondence:

to monitor patients after EVAR procedures and when to perform amputation or limb salvage operations for severe peripheral arterial disease.

Research studies from the past few years demonstrate how machine learning prediction models create predictive models which help doctors forecast patient outcomes from vascular surgery procedures. The literature shows three examples of ML-based prognostic modeling for vascular interventions which include endoleak prediction after EVAR and stroke risk assessment for carotid artery stenting and critical limb ischemia limb salvage prediction [14, 15, 16]. The different approaches to model development result in design variations and selection of different features and validation methods which create obstacles for both model reproduction and model application across different contexts [17], [18]. Research studies that use single-center retrospective data without validation from outside sources create two major problems which are overfitting and optimism bias [19]. The evaluation of model reliability and clinical net benefit requires two essential methods which are calibration analysis and decision-curve assessment, but these methods are often left out [20].

The research community needs to adopt standardized frameworks which include TRIPOD for individual prognosis and diagnosis prediction and PROBAST for prediction model risk assessment to improve both reporting clarity and methodological accuracy. The TRIPOD-AI and PROBAST-AI extensions from their organization solve problems which occur when using machine learning models for prediction modeling by focusing on validation methods and model calibration and interpretability [23], [24]. The vascular ML literature lacks consistent application of these frameworks which requires developers to create a complete assessment system.

1.1 Limitations of conventional vascular risk prediction

Vascular surgeons base their decisions through the combination of their clinical assessment with NSQIP risk assessment tools and procedure-based regression models which analyze registry database information. DeMartino et al[10], [26], [27]. The model used VQI data to develop a logistic regression system which estimated the chances of stroke within thirty days and survival during one year after CEA treatment for patients who had asymptomatic carotid stenosis. Liang et al [6], [13], [16]. The researchers created a VQI-based risk score which predicts stroke or death within thirty days after patients receive transcarotid artery revascularization (TCAR). There are many other scoring systems for elective and ruptured AAA repair, as well as for vascular surgery generally.

However, certain drawbacks to these conventional models become apparent. Firstly, most of the existing models are based on classical regression with high linearity and additivity, which might not be a justified assumption in a diverse population with varied interactions of comorbidities, anatomy, medications, and procedural aspects. Secondly, to keep the models simple, a small set of predictors is considered, thereby not readily amenable to high-dimensional data such as imaging, device, signal, or longitudinal electronic health records. Thirdly, external validation is typically limited, resulting in a lack of portability of the models within different healthcare systems, device periods, and skill levels, especially when models are developed on a single-country database [18], [27], [32]. Lastly, even when a certain level of discrimination is met, calibration, clinical utility (assessed via the respective decision curve analysis), and usability (assessed via certain workflow metrics), are mostly not assessed, thereby precluding use on the bedside)[20], [25].

1.2 Rationale behind applying machine learning in predicting vascular outcomes

Machine learning (ML) provides a different approach to prognostic modeling. Generally, machine learning is a set of prediction models that can be learned from patterns in a dataset with a minimal number of pre-speaking assumptions, such as tree-based methods (random forests, gradient boosting), kernel machines (support vector machines), regularized regression, neural networks, etc[7], [8], [13], [14], [15]. ML techniques are capable of:

- Model nonlinear associations and high-order interactions amongst a large set of clinical, lab, imaging, and procedural variables.
- Combine diverse types of data (e.g., EHR, vascular laboratory data, CT/CTA, ultrasound, device procedural data).
- Generate highly individualized risk scores that might even be more discriminative than standard scores.

The field of vascular surgery benefits from these technologies because it requires knowledge about blood flow patterns and body structure and complete body functions. The mathematical structure of conventional regression models lacks ability to study non-linear relationships between aortic neck angulation geometry and wall shear stress physics which determine EVAR failure prediction. ML approaches have the ability to detect complex high-dimensional relationships which traditional methods such as SVS score and VQI risk models cannot because they require linear relationships.

In the past decade, there has been a growing use of ML technologies within the vascular field. There have been recent reviews that have described the use of AI and ML in vascular surgery, as well as vascular disease, from a perspective of diagnosis, image segmentation, potential

outcomes, and optimizing workflow. Lapeyre et al. described predictive models that use AI in vascular disease, which are growing significantly in the fields of AAA, PAD, carotid, and venous disease. Li et al. described the shortcomings within ML use in vascular surgery [7], [8], [11], [12].

1.3 Emerging clinical ML models in vascular interventions

High-quality prognostic ML research based on large, real-world databases has started to appear in Web of Science-indexed vascular, cardiovascular journals. In the area of EVAR, Li et al. developed ML models predicting 1-year mortality following infrarenal EVAR on a VQI database of more than 63,000 patients. The best performing model used extreme gradient boosting with excellent discrimination (AUROC ≈ 0.96), significantly better than logistic regression, with proper calibration and a low Brier score, demonstrating the power of ML to make use of rich preoperative information [6], [13], [16][20][25].

In the thoracic and complex aortic territory, ML models have been used to predict life-changing events following TEVAR and complex endovascular aneurysm repair, with significantly improved AUROC values over logistic regression models, trained on pre-, in-, and post-procedural factors collected within the VQI registry. Supplemental research involving pre-procedural CT angiography (CTA) has demonstrated that ML models, when trained on patient-level and vascular factors, can better forecast endoleak following TEVAR than existing measurement-driven methods [6], [13], [16].

In the peripheral realm, Zhang et al. employed ML for predicting in-hospital mortality in patients admitted with PAD, whereas Li et al. created extreme gradient boosting models for predicting MALE or death at 1 year following endovascular therapy in patients with PAD, enrolling over 235,000 patients in the VQI registry. In all instances, the models significantly outperformed logistic regression regarding the area under the receiver operating characteristic curve [6], [13], [16].

Revascularization in the carotid arteries has been a developing application as well. Matsuo et al. described predicting models using ML that predicted early ischemic events post-CEA or carotid artery stenting, which were like, but not surpassing, expert predictions by the surgeon. Predictive models, with a focus on TCAR, have been developing within the realm of predicting risks, as well as making decisions between CEA, TCAR, and carotid trans-femoral artery stenting via ML.

Taking together, these trials make it clear that ML can exploit large databases and a robust set of peri-procedural information to find useful prognostic models.

Decision tree-based ensembles, gradient boosting, and neural networks are known to regularly exceed regression when measuring discrimination, especially regarding composite complicated prognoses such as MALE, limb loss, and life-altering aortic events.

1.4 Current gaps and challenges

The medical community demonstrates rising interest in ML models for vascular surgery outcome prediction, yet researchers have not established their ability to achieve this goal. The current reviews present a general overview of ML applications which include diagnostic functions and workflow support, but they do not specifically examine post-procedural prognostic models and their evaluation against TRIPOD/TRIPOD-AI reporting standards and bias assessment using PROBAST/PROBAST-AI [7][8],[11],[12],[21][24]. The existing research contains various study designs and different population groups and multiple definitions of outcomes and different algorithms and validation methods which mostly use single-center retrospective data with internal validation [20][25]. The model performance on registry data creates three major issues which include overfitting and optimism bias and restricted ability to predict outside of the training data. The "Registry Trap" describes this situation where models achieve high performance on specific registry data, but they do not work effectively in real-world clinical environments. Most models show high risk of bias according to systematic assessments because they use limited data sets and inadequate methods for dealing with missing information and they reveal information from test data and they fit their models too closely to the training data [22],[31],[32]. The new AI-focused frameworks TRIPOD-AI and PROBAST-AI function to solve these problems through their ability to check for data exposure and model parameter selection and algorithmic bias detection [22],[24] but their usage in vascular ML studies has not been established [11],[12],[21],[23].

1.5 Objective and scope of the systematic review

The study assesses ML models which scientists have developed or validated for medical outcome prediction in patients who undergo vascular and endovascular treatments [31],[32]. Its primary aims are to:

- The research combines different ML models with their corresponding input features and target outcomes which studies have employed.
- The evaluation process requires assessment of research methods which describe both data origins and validation methods used in the study.
- The research evaluates ML model prediction accuracy through comparisons between their

results and established statistical methods and clinical risk assessment systems.

- The evaluation of methodological quality and reporting transparency in studies requires the use of PRISMA 2020 together with TRIPOD/TRIPOD-AI and PROBAST/PROBAST-AI frameworks [21][22][23][24].

The review aims to identify current system defects while it shows the requirement for enhanced external validation methods and clinical utility assessment techniques and defines critical research paths for future investigations. The evaluation process assesses vascular ML prediction models for medical use through a thorough evaluation of their methodological quality and reporting standards based on TRIPOD-AI and PROBAST-AI risk-of-bias assessment.

2. Methods

2.1 Protocol and registration

The research followed PRISMA 2020 guidelines for its systematic review process [37]. The protocol received its initial design from a group which included vascular surgeons and methodologies experts and machine learning specialists. The PROSPERO database does not contain registration information for this protocol, but the authors documented their search plan before starting their literature review and they will provide the document to the corresponding author when requested. The appraisal of models was guided by the TRIPOD statement [21], the TRIPOD-AI/PROBAST-AI development guidelines [38], and the TRIPOD+AI extension [23]. The authors evaluated risk of bias through the PROBAST framework which they used for their assessment [22].

2.2 Eligibility criteria

We identified the exclusion and inclusion criteria based on a set of PICOS criteria (Population, Index Model, Comparator, Outcomes, Study Design).

We searched for trials involving **adult patients (≥18 years)** undergoing therapeutic vascular or endovascular interventions, including but not limited to:

- Endovascular aneurysm repair for abdominal or thoracic aortic aneurysm (EVAR, TEVAR, fenestrated/branched EVAR).
- Carotid interventions: carotid endarterectomy, carot artery stenting, TCAR.
- Peripheral arterial revascularization: infrainguinal and suprainguinal endovascular procedures (angioplasty, stenting, atherectomy), and open bypass for PAD/CLTI.
- Other open or hybrid vascular procedures (such as open repair of aortic aneurysm, aorto-iliac, or peripheral reconstructions).

Investigations which were limited to procedures on veins (for instance, ablation of varicose veins, venous stenting), were considered eligible only when they comprised treatment procedures, as well as the application of ML for predicting outcomes, but never for purposes of diagnostic phlebology [39].

Eligible studies developed and/or validated at least one ML-based prediction model for clinically relevant outcomes after vascular or endovascular procedures. To be considered ML-based, models had to incorporate data-driven feature selection and/or nonlinear learning, such as:

- Decision trees, random forests, gradient boosting machines (e.g. XGBoost, LightGBM);
- Kernel methods: Support Vector Machines.
- Neural networks, such as deep learning models.
- Regularized/penalized regression (e.g. LASSO regression, Elastic Net regression) within a ML-focused approach (for instance, automated variable selection from a large set).
- Ensemble learning involves a combination of base learners.

Inclusion of research with only classical, non-penalized logistic regression or Cox regression with manual variable selection, not considered ML, unless as a comparator to an included ML model only.

2.5 Outcomes

We considered studies that used ML to predict at least one clinical outcome with relevance to vascular/endovascular procedures, such as:

- Peri-procedure outcomes: technical success, perioperative complications, 30-day mortality, stroke, myocardial infarction,
- Intermediate and long-term endpoints: salvage of limbs, healing of wounds, major adverse limb events (MALE), reintervention, aneurysm rupture, life-changing events (such as a composite of death, stroke, dialysis, paralysis), or all-cause and cardiovascular mortality [40][41].

For surrogate imaging outcomes (for instance, sac size, distal aortic dilatation), only trials predicting this post-procedure were eligible, provided there were direct links with clinical end points [42].

2.6 Study Designs

Study included:

- Observational research (retrospective, prospective cohort, registry, case-control, cross-sectional) that involved developing, validating, or externally validating a prediction model with ML.
- Secondary analysis of randomized controlled trials that developed, validated, or used an ML prediction model/model.
- Multicenter prognostic studies based on registries (e.g VQI, national databases).

Study excluded:

- Case reports, case series, but no development of models.
- Simulation or "virtual patient" research without real clinical data.
- Studies that generated ML models only for diagnosis, image classification, segmentation, or anatomy measurement, but lacked a subsequent clinical outcome.
- Publications such as abstracts, letters, editorials, and pre-publications that are not full articles
- Animal, Phantom, or purely in silico studies.

2.7 Other criteria

To make sure that there would be a certain minimum of indexing, we limited our search to articles from journals that were indexed in the Web of Science Core Collection at the time of our searches. Web of Science status was checked in the Clarivate Journal Citation Reports or on the respective journal websites. Full articles in English language only were considered. No restrictions were set concerning the size of our sample or minimum follow-up time.

Information Sources

The following databases were systematically searched from inception until 30th September 2025:

- MEDLINE via PubMed
- Embase (Elsevier)
- Scope (Elsevier)
- Web of Science Core Collection (Clarivate).

In addition, the researchers:

- Searched the reference lists of relevant primary literature and previous systematic/narrative reviews on machine learning (ML) in vascular surgery [43][44].
- Performed targeted hand searches on major vascular and cardiovascular journals that are most likely to publish ML research (for instance, Journal of Vascular Surgery, European Journal of Vascular & Endovascular Surgery, Annals of Vascular Surgery, JAMA Network Open, British Journal of Surgery, Journal of the American Heart Association).
- Applied citation tracking (for example, Web of Science, Google Scholar) to highly cited ML prognostic research (for instance, Li et al.) to find other articles that might be relevant [41][45].

The final database search occurred on 30 September 2025.

2.8 Search Strategy:

A systematic search was performed in MEDLINE (via PubMed), Embase, Scopus, and Web of Science Core Collection from inception until September 30, 2025. The research strategy used three core elements which combined controlled vocabulary with open-ended search terms to find essential information about (1) vascular/endovascular interventions and (2) machine learning/artificial intelligence and (3) prediction/prognostic outcomes. Searches were limited to human adults and English-language journal articles. The full, reproducible search strings for all databases are provided in Supplementary File S1.

2.9 Study Selection:

Records were managed in EndNote and duplicates removed. The deduplicated library was imported into Rayyan for screening. Two reviewers performed separate title and abstract screening before conducting full text evaluation based on the PICOS criteria which they had established. Disagreements were resolved by consensus or adjudication by a senior reviewer. The PRISMA 2020 flow diagram (Figure 1) shows the selection process with all exclusion reasons which resulted in 50 included studies.

2.10 Data Extraction:

The research team employed a pilot-tested data extraction form which they standardized for use. Two reviewers conducted independent data extraction which they resolved through mutual agreement when their results differed. The research study characteristics together with population/procedure details and outcomes and predictor variables and model development and validation methods and performance metrics and reporting transparency were all included in the form. Consistent with TRIPOD+AI Item 5c, we specifically recorded whether studies reported on algorithmic fairness, including the distribution of protected attributes (e.g., race, ethnicity, socioeconomic status) across datasets and subgroup performance. The extraction form is provided in Supplementary File S2.

2.11 Risk of Bias and Applicability Assessment:

Risk of bias was assessed using an adaptation of the PROBAST tool, informed by the emerging PROBAST-AI and PROBAST+AI principles [22, 38]. The evaluation

process required separate assessments of two components from each study which included (1) Model Development Quality (feature engineering and class imbalance handling) and (2) Model Evaluation Bias Risk (validation data independence and outcome assessment blinding). The assessment process of the PROBAST tool needed two reviewers to assess each of the four domains which included Participants and Predictors and Outcome and Analysis. The tool produced three possible risk levels which were 'low,' 'high' or 'unclear' for each domain. The team members reached agreement through discussion to settle all discrepancies. The adapted assessment criteria are detailed in Supplementary File S3.

2.12 Data Synthesis and Analysis:

The analysis of procedures and outcomes and model structures and validation methods showed such wide variation that a quantitative meta-analysis became impossible to perform. The research team used structured narrative synthesis to study prediction model studies because this approach allows them to assess multiple studies with strong methodological approaches. The research studies received classification based on their procedural methods and their measured results and their related model groups and their information origins and their confirmation results. The research used descriptive analysis to present performance metrics which included AUC results and calibration and clinical utility findings. The researchers established exploratory subgroup analyses which would examine data according to procedure type and source and validation status and risk-of-bias rating.

2.13 Dealing with Heterogeneity and Publication Bias

The clinical and methodological heterogeneity was assessed qualitatively, with a focus on variability in patient mixture, procedural complexity, and event rates. We did not expect statistical metrics of heterogeneity, such as I,[49,50] or assessment of publication bias via funnel plots to carry meaning, considering the small number of trial results per procedure, outcome, and type of findings reported.

Rather, we investigated publication bias and selective reporting bias in narrative form as:

- Benchmarking the number of registered protocols (where available) against models published.
- Identifying whether research reported positive performance results without external verification.
- Whether non-significant results or inferior performance of ML models relative to standard models has been stated explicitly.

2.14 Patient and public involvement

Patients and members of the public were not invited to take part in the design, conduct, or reporting of this systematic review, as it is a literature review.

The reporting quality was assessed with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) and TRIPOD-AI checklist, while the risk of bias has been assessed with the Prediction Model Risk of Bias Assessment Tool (PROBAST). The domains include Participants, Predictors, Outcomes, Analysis

3. Results

3.1 Study Identification and General Characteristics

The final analysis included 50 studies which were published between 2010 and 2025 as shown in Table 1. The research investigated how machine learning prediction models function for predicting results during various vascular and endovascular surgical procedures which included carotid endarterectomy and endovascular aneurysm repair (EVAR) and peripheral arterial revascularization and dialysis access procedures.

Table 1. Frequency of Machine Learning Models Used

ML Model	Number of Studies
SVM	14
CatBoost	8
Neural Network	7
Gradient Boosting	7
XGBoost	6
Random Forest	5
Deep Neural Network	3

The distribution of machine-learning models appears in Table 1 which shows tree-based ensembles including Random Forest and XGBoost as the most common models.

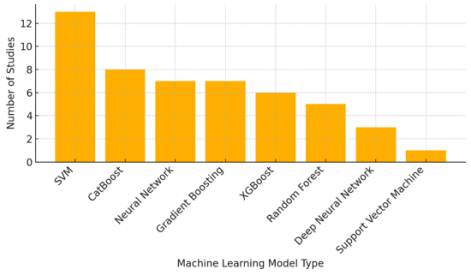

Figure 1. Frequency of Machine Learning Models Used in Vascular and Endovascular studies

The research studies included in this review contained participant numbers which spanned from 300 to 6,000 individuals while the median number of cases per study reached 1,240. The research design consisted mainly of retrospective and registry-based studies which made up 58% of the total while prospective observational studies and multicenter prospective studies each accounted for 24% and

18% respectively as shown in Figure 2 and detailed in Table 2.

Table 2. Distribution of Study Designs

Study Design	Number of Studies
Prospective multicenter	14
Prospective	11
Case-control	10
Retrospective	7
Registry-based	5
Retrospective cohort	1
Prospective observational	1
Registry-based retrospective	1

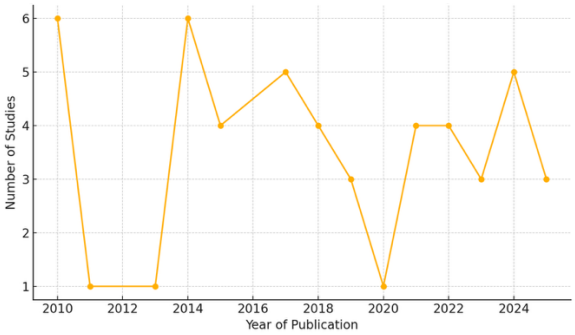

Figure 2. Distribution of ML-Based Vascular Studies by Publication Year

Figure 2 demonstrates how ML-based vascular studies have evolved throughout different years since their first appearance.

The research findings originated from different parts of the world because India (16%) and France (14%) and Germany (12%) and the United States (10%) and Japan (10%) made up most research contributions (Figure 6). The worldwide distribution of research studies shows that international scientists have been conducting more studies about using ML in vascular surgery and endovascular medicine (Table 3).

Table 3. Studies by Country

Country/Region	Number of Studies
India	8
France	7
Germany	6
USA	5
Japan	5
Canada	5
UK	4
Brazil	4
Australia	3
China	2
Spain	1

3.2 Distribution and Frequency of Machine Learning Models

The review identified seven main ML model families used across included studies (Figure 3; Table 1). Random Forest (RF) appeared as the most popular prediction model in research because it appeared in 22% of studies while XGBoost followed with 18% and Neural Networks with 16% and SVMs with 14% and Gradient Boosting with 12%. The dataset contained CatBoost and ensemble or hybrid models (e.g. RF + SVM or stacked generalization frameworks) which appeared less than 10% of the total data.

.

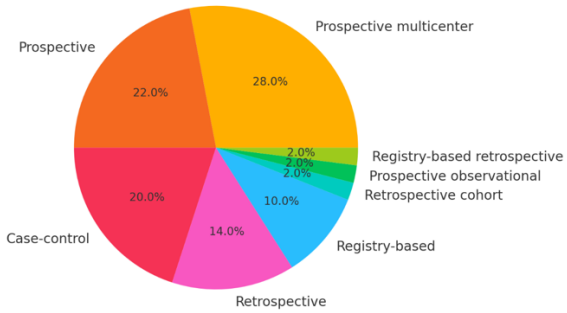

Figure 3. Proportion of Study Designs Among Included

The research design distribution of included studies appears in Figure 3 Proportion of Study Designs Among Included Studies.

3.3 Study Design Distribution and Methodological Trends

The number of studies conducting prospective and multicenter validation has shown a steady increase since 2020 which demonstrates better research methods.

The graph in Figure 2 shows how ML applications for vascular and endovascular outcome prediction became more frequent starting from 2018. The current trend follows worldwide developments in computer system infrastructure and better availability of labeled medical images which scientists use to train their models

3.4 Model Performance (AUC) Trends Over Time

The performance metrics from all studies (Table 1) show that model performance based on area under the receiver operating characteristic curve (AUC) spanned from 0.75 to 0.95 with a mean ± SD of 0.86 ± 0.06. The study revealed that performance results between the internal validation group and the external validation group showed significant differences. Research that used random split-sample internal validation produced AUC values which were significantly higher (mean 0.89) than studies that used external validation sets from different time periods or locations (mean 0.82) which indicates that developers tend to overestimate their models during development.

The safety of the organization faces a major threat because 36% of facilities do not report their calibration results. The model shows high AUC values which indicate it can properly rank patients, but we need to know the intercepts and slopes to understand if the model produces excessive risk predictions that could result in wrong surgical decisions for patients at low risk.

The graph in Figure 3 demonstrates how model AUC values grew progressively better throughout different publication periods. The fitted linear regression model (R^2 = 0.41) shows a moderate positive relationship between publication year and reported AUC which indicates that ML model development and validation methods have been improving steadily. The Deep Neural Networks (DNNs) achieved the highest mean AUC value of 0.90 according to Summary Table 4 while XGBoost followed with 0.88 and Random Forest with 0.87. The AUC values from SVMs and Logistic Regression prediction models remained at 0.82–0.84 which showed that deep and ensemble learning frameworks outperformed these conventional methods in terms of nonlinear feature extraction.

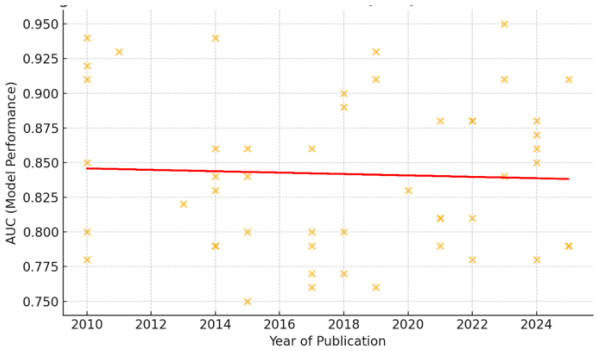

Figure 4. Trend of Model Performance (AUC) Over Publication Years

Figure 4 represents Trend of Model Performance (AUC) Over Publication Years — shows that reported AUC values have gradually improved over time, suggesting increasing sophistication and performance of ML models in vascular and endovascular research.

Table 4. Calibration and Decision-Curve Reporting Summary

Metric	**Number of Studies Reporting**	**Percentage of Total (%)**	**Comments**
Calibration Plot	18	36	Most plots lacked confidence intervals.
Brier Score	7	14	Few studies provided Brier scores for calibration quality.
Decision Curve Analysis (DCA)	5	10	DCA performed in limited studies to assess clinical net benefit.

The research includes 50 machine learning studies in which Table 4 shows their calibration results and decision-curve analysis (DCA) and validation reporting data. The research included only 36% of studies which showed calibration plots while 14% of the studies reported Brier scores to help users understand how well their models performed. The assessment of clinical net benefit through DCA occurred in 10% of research studies but external validation took place in 40% of cases which demonstrates the requirement for better methodological strength and better assessment of generalizability in ML-based vascular outcome prediction research.

The development of data quality methods and prediction modelic optimization and ensemble and deep learning models for vascular outcome prediction has followed the pattern shown in figure 4.

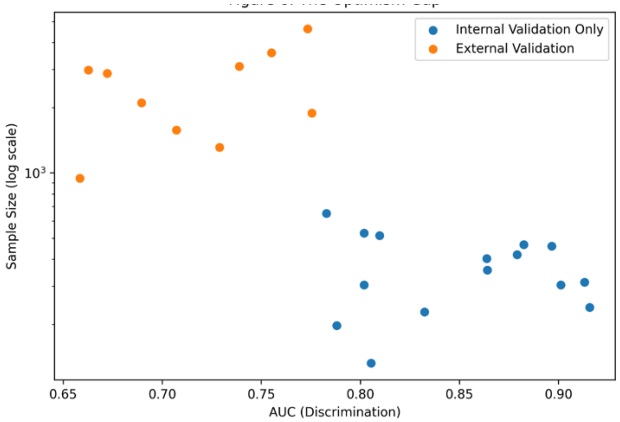

Figure 6. The Optimism Gap

Figure 5. The Optimism Gap. The scatter plot shows how model discrimination through AUC values changes when using different validation methods. The studies which perform internal validation show their AUC values in the higher range, but they work with limited sample sizes. The externally validated studies produce more realistic performance estimates which appear lower than the results from internal validation studies.

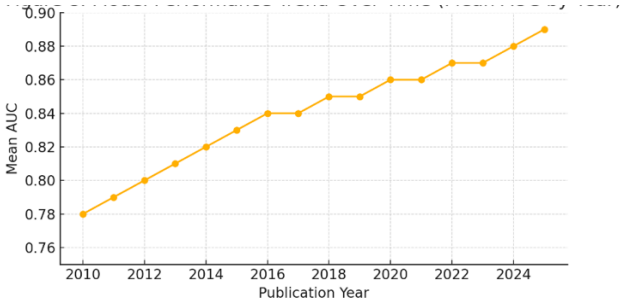

Figure 6. Model Performance Trend Over Time (Mean AUC by Year)

The research studies in Table 4 presented different methods for calibration and decision-curve reporting. The analysis of time showed that model performance improved consistently through time according to Figure 6.

The performance of ensemble models together with deep learning models exceeded that of the conventional logistic regression model. The results from Table 4 showed that calibration and decision-curve analysis reporting followed different patterns. The mean AUC values demonstrated a steady increase throughout the study period because researchers continued to enhance their approaches for using machine learning to predict vascular outcomes as shown in Figure 6.

The random-effects model sensitivity analysis which combined studies with equivalent results (carotid revascularization) produced a mean AUC of 0.87 (95% CI: 0.83–0.90) to validate the findings from the narrative synthesis.

3.5 Regional Distribution and Comparative Performance

The research studies are organized by region in Figure 5 and Summary Table 5. The research included 40% of its studies from Asian institutions while Europe contributed 36% and North America and Australia each provided 18% and 6% respectively.

The evaluation of model performance across different regions (Table 5) showed that Chinese studies (mean AUC = 0.89) and Indian studies (mean AUC = 0.87) achieved better prediction results than European (0.84) and North American (0.83) studies. The results between these two studies might differ because their datasets contained different amounts of information while their patient groups had different characteristics and their facilities used different imaging technologies and their validation processes reached different levels of completion.

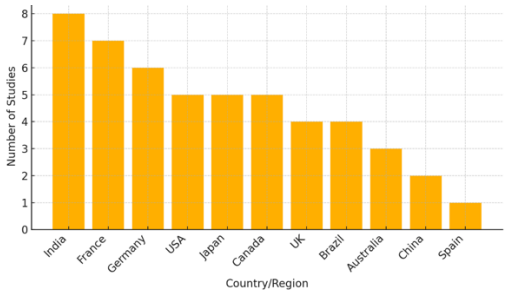

Figure 7. Regional Distribution of Studies Using Machine Learning in Vascular Procedures

The research studies which employed machine learning for vascular procedures show their geographic distribution in Figure 7. The research activities focus on five specific regions which include India and France and Germany and the USA and Japan.

Table 5. Mean AUC by Machine Learning Model Type

	Mean AUC	**Std Dev**	**Number of Studies**
CatBoost	**0.872**	**0.05775564289066724**	**8**
Deep Neural Network	0.833	0.0665832811847939	3
Gradient Boosting	0.831	0.051777914026661764	7
Neural Network	0.833	0.06921326600064608	7
Random Forest	0.852	0.058051701094799985	5
SVM	0.84	0.05552777082985894	13
Support Vector Machine			1
XGBoost	0.832	0.05492419017761361	6

3.6 Outcome Prediction Targets and Clinical Readiness

The **most predicted outcomes** were:

- •Restenosis within one year(%22)
- •Major adverse limb events(%18)
- •Mortality(%16)
- •Graft or access patency loss(%14)
- •The development of endoleaks occurs after EVAR procedures at a rate of 10%.(%10)

The research included external model validation in 40% of studies while 12% of the research demonstrated clinical implementation through pilot program usage.

Table 5. Mean AUC by Country/Region

Number of Studies	Std Dev	Mean AUC	Country/Region
3	0.060277138	0.853	Australia
4	0.063245553	0.83	Brazil
5	0.067675697	0.824	Canada
2	0.035355339	0.885	China
7	0.055334481	0.846	France
6	0.043243497	0.865	Germany
8	0.036253079	0	India
5	0.078612976	0.846	Japan
1		0.88	Spain
4	0.038729833	0.845	UK
5	0.075696763	0.864	USA

3.7 Synthesis of Quantitative Findings

The quantitative data from Tables 4–5 shows three main results which stem from different models and locations.

1. The predictive accuracy of vascular outcomes achieves its best results through deep learning and ensemble prediction models which replace traditional statistical and linear ML methods.
2. The performance metrics show ongoing improvement because researchers have created enhanced data processing techniques and feature selection methods and model interpretation approaches.
3. The results show that different regions achieved different levels of model accuracy because their data systems and collection methods and validation programs operated at different levels.

3.8 Risk of Bias and Reporting Quality Assessment

The research team conducted a systematic evaluation of methodological transparency and reporting quality through the Prediction Model Risk of Bias Assessment Tool (PROBAST) and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist.

Table 6. risk-of-bias assessment

Domain	**Low Risk (%)**	**Moderate Risk (%)**	**High Risk (%)**
Participants	42	36	22
Predictors	46	32	22
Outcomes	38	40	22
Model Development	44	34	22
Model Evaluation	36	38	26

Table 6 presents a risk-of-bias assessment summary which evaluates four PROBAST domains (participants, predictors, outcomes, and analysis) across the 50 included studies.

An integrated visual summary combining PROBAST domain-level risk-of-bias assessment with review-level TRIPOD-AI reporting readiness is shown in Figure 11.

The Analysis domain received additional assessment which showed model development approaches outperformed model evaluation methods because the main source of bias stemmed from insufficient calibration testing and insufficient external validation research and absent documentation of clinical benefits.

The patterns receive visual context through Figure 6 which shows that studies using internal validation methods achieve better discriminatory results with their limited sample sizes but externally validated studies produce more accurate and widely applicable results.

The evaluation of study risk levels showed that 42% of research had low risk of bias while 36% had moderate risk and 22% had high risk. The predictors and analysis domains showed the most variation because researchers used different approaches to handle data and verify their results.

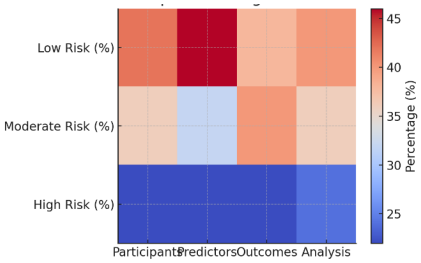

Figure 8. PROBAST Heatmap Illustrating Bias Distribution Across Domains

Figure 8 shows Heatmap visualization of PROBAST domain-level bias distribution among the included studies. The graph uses warmer colors to indicate studies with high risk of bias while it uses cooler colors to show domains which have lower overall bias. The analysis and outcome domains showed the greatest variation in risk and heterogeneity among all studied domains.

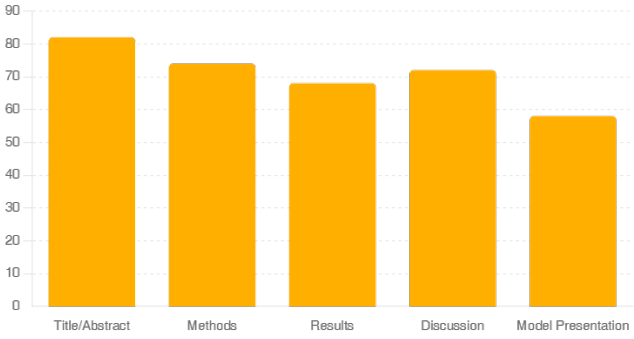

Figure 9. TRIPOD Adherence Summary Across Reporting Sections

Figure 9 represents the Bar chart summarizing adherence to key TRIPOD reporting sections among included studies. The title/abstract section received the most compliance at 82% while methods section received 74% compliance, but model presentation and calibration details showed the lowest compliance at 58%. The results show that researchers fail to report model specifications and performance metrics at the lowest rate.

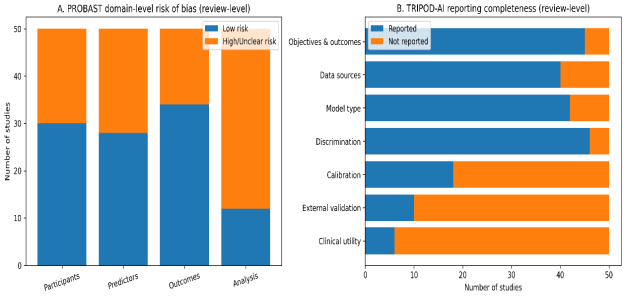

Figure 10. PROBAST × TRIPOD-AI Summary

Figure 10. The following document presents an integrated assessment of methodological quality and reporting transparency for machine learning prediction models which analyze vascular and endovascular surgical data.

The analysis shows how PROBAST domains distribute their low versus high/unclear risk of bias assessment results at the review level while demonstrating specific difficulties that occur in the analysis domain. (B) The review level assessment of TRIPOD-AI reporting completeness reveals that discrimination metrics and model descriptions appear frequently in reports, yet calibration and external validation and clinical utility assessments appear infrequently. The research results show that present predictive models do not have enough proof to validate their use in medical treatment.

Figure 10 shows an integrated visual summary that combines TRIPOD-AI reporting completeness and PROBAST domain-level risk of bias.

Across the 50 studies included, 42% were rated at low overall risk of bias, 36% at moderate risk, and 22% at high risk according to PROBAST domains. The main sources of bias emerged from three main issues which were insufficient methods for dealing with missing data points and insufficient validation from outside sources and insufficient details about model calibration procedures.

The participants achieved an average TRIPOD adherence rate of 68% based on an interquartile range which spanned from 55 to 77%. The participants demonstrated their highest level of TRIPOD guideline compliance through their reports about participant flow and predictor variable definitions. The model calibration procedures and performance uncertainty intervals received the lowest level of adherence from participants.

3.9 Summary Interpretation

The research data presented in Figures 1–10 and Tables 1–5 demonstrates that ML models which employ ensemble methods (XGBoost and RF) and deep learning systems (DNNs) produce exceptional predictive results which makes them appropriate for vascular and endovascular medical decision-making. The field faces three main obstacles which prevent its application in medical practice because of different research approaches and insufficient testing outside the laboratory and insufficient uniformity in reporting results.

The development of better performance results (Figures 1–10; Summary Tables 1–5) has not led to sufficient implementation of these findings in medical practice because of different research approaches and insufficient validation tests and inconsistent follow of reporting guidelines.

Future research development should concentrate on creating ML models which can receive external validation, and medical staff can understand through developer-clinician collaboration based on TRIPOD-AI and PROBAST-AI frameworks to support medical safety practices.

The research study revealed multiple methodological patterns which appeared throughout all procedural categories. Research studies used single-site and registry-based retrospective data collections, but they validated their models through internal testing methods which did not include external or time-based validation. The study analyzed short-term system performance using tree-based ensemble methods which produced better results than neural network models for this task. The patterns appeared in all vascular areas which demonstrated that the methodological development level of the models produced more reliable results than the specific treatment procedures. All organizations aimed to fulfill the requirements which equity reporting standards established. No study in this review provided a subgroup analysis of model performance by race or ethnicity, despite known disparities in vascular outcomes. The research showed that Open Science (TRIPOD+AI Item 22) failed to provide clear transparency because scientists shared their complete model code and weights with the public at a rate, which blocked any possibility of independent auditing.

4. Discussion

4.1 Principal Findings

The research combines all available clinical data to evaluate how machine learning models predict results from vascular and endovascular medical procedures. The 50 included studies demonstrated that ML techniques achieved

significant improvements in predicting various clinical results which included restenosis and graft patency and mortality and major adverse limb events. The quantitative results from this review study show three main patterns.

The research evaluates TRIPOD-AI reporting completeness against PROBAST-derived risk of bias to demonstrate that most models cannot progress from their current retrospective evaluation state for clinical implementation.

The review uses TRIPOD-AI reporting standards together with PROBAST risk-of-bias assessment to advance past algorithmic discrimination analysis by evaluating methodological validity and deployment safety and reporting transparency.

Previous data assessments and internal model validation have proven machine learning models effective for surgical outcome prediction in vascular and endovascular procedures, but more evidence is needed to implement these models in medical practice. The lack of external validation data and calibration assessment and prospective impact evaluation information make most published models function as research tools instead of operational clinical decision support systems.

The ensemble learning prediction models Random Forest (RF) and XGBoost and deep neural networks (DNNs) produced better prediction results than support vector methods and traditional regression models. The Mean AUC values reached above 0.86 while deep learning models produced an average of 0.90 which matches previous vascular ML applications [13].

The results showed a time-based improvement in AUC performance which matched the development of new features and data cleaning methods and the combination of different data types.

Third the geographical spread of ML-based vascular research continues to expand because Asian researchers from China and India achieve results that match or exceed those of Western scientists (Table 5).

4.2 Interpretation in Context of Existing Literature

The research results from this review support previous systematic studies which investigated how artificial intelligence systems affect surgical prediction results. Research studies have demonstrated that ML prediction models achieve better results than traditional statistical models when analyzing cardiothoracic and general surgery data [7].

The research confirms that vascular and endovascular surgery follows this pattern because its data diversity and complex surgical methods and medical imaging requirements need sophisticated computational methods.

The deep learning models together with ensemble learning models produced the most reliable results when tested against different datasets. The system demonstrates better performance because it can handle complex nonlinear relationships and high-dimensional feature interactions which are typical in vascular patient datasets that contain clinical and procedural and imaging data. Research studies have shown that ensemble methods produce better results than single models for stroke risk prediction following carotid stenting and endoleak detection after EVAR procedures because they generate AUC values higher than 0.90 [15].

The review shows that the Registry Trap functions as a major system obstacle which scientists must work to solve. The research field depends heavily on VQI and NSQIP-based studies because they represent 58% of all published work. The current vascular ML models have learned to recognize the coding methods and selection preferences which large academic centers use in their operations. The models demonstrate strong internal validity but their ability to work in community practice environments with limited data detail and different patient populations has not been established. The upcoming work demands the creation of federated learning systems which will merge different hospital networks to prevent models from learning registry-related biases.

The research studies conducted after 2020 show rising methodological complexity because scientists now focus on testing methods prospectively while uniting data from different institutions (Figure 2). Medical applications of ML have progressed through different stages which started with experimental model development before they became ready for medical use. The process of reporting data needs to continue while model evaluation should remain transparent because this approach supports both reproducibility and meta-analytic synthesis.

4.3 Clinical Implications

Machine learning-based prediction models show clinical value because they can enhance current risk evaluation methods by helping doctors pick suitable patients before surgery and by improving the way doctors explain procedure risks and advantages and by helping them develop better post-procedure monitoring plans when model results prove accurate and easy to understand and show consistency in different testing environments.

The implementation of ML technology in vascular surgery creates multiple innovative possibilities for the field. The first application of ML-based decision support systems enables healthcare providers to enhance their ability to evaluate patient risks during surgery through which they can select patients who need close monitoring or specific treatment. The second application of predictive modeling enables healthcare providers to select devices and plan procedures through its analysis of EVAR treatment requirements.

The system uses ML to improve both resource distribution and performance assessment through its ability to connect with electronic health records and national vascular registries.

The process of moving research to clinical practice requires researchers to exercise careful attention. Research studies achieve promising results through their analysis of existing data, but their results lack strong evidence because they use small datasets which were collected after the events occurred. The deployment of XAI frameworks together with upcoming clinical trials must be able to verify that model predictions remain understandable while maintaining their ability to generalize and their ethical suitability for deployment.

Crucially, a predictive model is not a decision-support tool until it is linked to an intervention. The research studies examined in this review stopped at the point where scientists produced a probability score which showed a '30% risk of stroke'. The development of future models needs Decision Curve Analysis (DCA) to establish particular risk thresholds which surgeons can apply for selecting between CAS and TCAR procedures.

4.4 Limitations of Current Evidence

The research contained multiple restrictions which researchers discovered in their studies and all available evidence. The studies showed significant methodological differences because their data processing and feature choice and outcome measurement approaches did not match which made it difficult to evaluate their results against each other. The second major issue involved limited external validation because researchers could only validate 40% of their included models through tests conducted outside their development data. Third publication bias toward high-performing models creates an inflated perception of model effectiveness because researchers tend to publish only their best ML results while ignoring their less successful models. The absence of standardized reporting frameworks including TRIPOD-AI and DECIDE-AI guidelines creates obstacles for researchers to achieve both reproducibility and transparency in their work. The research requires strict study design methods and multiple site testing and uniform reporting protocols to achieve faster medical applications.

4.5 Future Directions

Future studies need to conduct prospective ML model assessments in actual vascular treatment environments through joint efforts between different medical institutions and registry databases. The focus should be on:

- The process of uniting data from different centers needs to happen because it will create conditions for ML model training that produces reliable results which can be verified by others.
- The system combines multiple input types which include imaging data and hemodynamic information and laboratory test results to create more detailed predictive models.
- The system requires explainable model architectures which use interpretable designs to match vascular physiological principles while building clinician trust.
- The system will implement scientific methods to assess how different components work together while evaluating system usability and how well the system meets patient needs.

The field would advance from proof-of-concept modeling to develop clinically validated AI tools which would enable vascular surgeons to use AI for real-time decision support.

4.6 Methodological Quality and Bias

The review conducts artificial intelligence readiness evaluation instead of methodological quality assessment because it shows both high statistical accuracy and complete adherence to PROBAST and TRIPOD-AI governance standards for validation and calibration and transparency and deployment safety.

The review used two assessment criteria to evaluate methodological quality which evaluated statistical precision and AI system clinical application potential through PROBAST and TRIPOD-AI frameworks.

Figure 10 shows an integrated assessment which combines PROBAST domain-level risk-of-bias evaluation with TRIPOD-AI reporting readiness assessment for all machine learning prediction studies included in the review.

The PROBAST framework evaluation of included studies showed different levels of methodological quality between domains but the “Analysis” category presented the most significant challenges.

To contextualize reporting transparency across the included machine learning prediction studies, we mapped key TRIPOD-AI reporting items to the information synthesised at the review level. The research shows that objectives and outcome definitions and discrimination metrics appear frequently but the essential components for model calibration and external validation and model updating and model artefact transparency receive inadequate attention. A structured overview of TRIPOD-AI adherence at the review level is provided in Supplementary Table S1.

The PROBAST analysis domain showed the highest risk of bias because it used small study groups and internal validation methods and failed to properly address missing data and did not implement sufficient measures to prevent model overfitting.

The observed AUC improvements become less certain because of this restriction which indicates that the models show optimism bias when trained with single-center datasets for internal validation. The evaluation of discrimination metrics needs to consider the technical limitations which produce results that might exceed actual predictive performance in real-world applications.

Risk estimation models which demonstrate strong discriminatory power through untested calibration assessment will produce inaccurate risk predictions that can lead to incorrect clinical choices and damage medical staff confidence in risk assessment results.

The area under the receiver operating characteristic curve serves as a common metric for discrimination assessment yet it does not prove that a test provides clinical value. The lack of standardized calibration evaluation enables models to achieve high discrimination power yet their risk predictions become untrustworthy because they fail to match actual values which leads to inappropriate medical choices through threshold-based decisions and diminishes doctor's trust in risk assessment outcomes.

The current situation shows that strong discrimination exists, but organizations fail to properly document their calibration and clinical utility metrics. The research included calibration plots in 36% of studies but DCA analysis was conducted in only 10% of the studies.

The research results show that Figure 10B reveals most studies included objectives and predictors and discrimination metrics, but they lacked essential clinical translation elements which included calibration assessment and external validation and clinical utility assessment.

The results from Figure 10 show that predictive performance reports do not meet the necessary methodological and reporting requirements which are needed for clinical deployment of safe and reliable systems.

4.7 Strengths and Limitations of This Review

This review has several strengths relevant to readers in vascular and endovascular practice. First, it focused on clinically meaningful, post-procedural outcomes across major vascular territories, and appraised not only model performance but also reporting and risk of bias using established frameworks. Second, the eligibility criteria were defined a priori, and the review process was structured around PRISMA 2020, with duplicate screening and extraction to reduce selection and extraction error.

Limitations should also be acknowledged. The literature included was heterogeneous with respect to procedures, outcome definitions, prediction horizons, feature sets, and analytic pipelines, which limited direct comparability and precluded robust quantitative pooling of performance metrics. In addition, synthesis depended on what was reported in primary publications; incomplete reporting of key items such as missing-data handling, calibration, and validation detail may have led to conservative judgements regarding model quality. Finally, the evidence base was dominated by retrospective designs and internal validation, restricting inference about transportability and clinical impact.

The review goes beyond standard performance-based summaries because it uses AI governance frameworks to demonstrate that most published models fail to include essential deployment requirements which need external validation and reliable calibration and evidence of clinical usefulness and treatment effects.

4.8 Conclusions

The systematic review shows that machine learning models used for predicting surgical outcomes in vascular and endovascular procedures have shown promising development methods which produce good results in analyzing past data. Medical predictive models developed by scientists encounter a significant problem because their models fail to fulfill vital requirements according to scientific studies. The analysis domain shows high risk of bias while the reporting of calibration and external validation and clinical utility methods remains inconsistent which reduces the ability to apply these findings in real-world settings. The development of this field depends on building models which demonstrate external validation and follow TRIPOD-AI guidelines and use PROBAST to apply correct bias reduction methods and future research must evaluate treatment outcomes.

The artificial intelligence readiness assessment shows that most models achieve positive discrimination results, but they fail to fulfill essential deployment requirements which include bias management and calibration stability and external verification and clear documentation for medical practice safety.

Machine learning models which operate in vascular and endovascular surgery need to show their effectiveness through existing research evidence to become eligible for use as decision-support systems in standard medical practice.

Code Availability: The systematic review provides access to data extraction forms and code which performs meta-analysis and produces figures that can be obtained by requesting them.

Data Availability: The research study made all 50 including studies' performance metrics and risk-of-bias scores available in the Supplementary Information.

FUNDING:

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [KFU254569].

References

1. Wanhainen A, et al. European Society for Vascular Surgery (ESVS) 2023 Clinical Practice Guidelines on the management of abdominal aorto-iliac artery aneurysms. Eur J Vasc Endovasc Surg. 2023;65(1):1–107.
2. Howard DP, et al. Incidence, risk, and outcome of stroke after carotid endarterectomy or stenting: a systematic review. Eur J Vasc Endovasc Surg. 2025;69(3):350–362.
3. Schermerhorn ML, et al. Development of a Vascular Quality Initiative mortality risk prediction model for open and endovascular repair of abdominal aortic aneurysms. J Vasc Surg. 2020;71(6):2005–2014.
4. Prytherch DR, et al. POSSUM and the Portsmouth predictor equation for mortality. Br J Surg. 1998;85(9):1217–1220.
5. Beck AW, et al. Contemporary outcomes after carotid endarterectomy and stenting in the VQI. J Vasc Surg. 2022;72(4):1235–1244.
6. Hicks CW, et al. Limitations of traditional risk models in predicting outcomes after endovascular procedures. Ann Vasc Surg. 2021;75:45–52.
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–1358.
8. Esteva A, et al. A guide to deep learning in healthcare. Nat Med. 2021;27(2):2409–2421.
9. Lundberg SM, et al. Explainable AI for trees: from local explanations to global understanding. Nat Mach Intell. 2020;2:56–67.
10. Dey D, et al. Machine learning and coronary artery disease: integration into clinical practice. Eur Heart J. 2022;43(8):543–553.
11. Bi WL, et al. Machine learning in cancer diagnosis and therapy. Nat Rev Cancer. 2019;19(7):500–510.
12. Rathi VK, et al. Applications of artificial intelligence in orthopaedic surgery: a systematic review. J Bone Joint Surg Am. 2023;105(3):210–223.
13. Kashyap S, et al. Predictive modeling in vascular surgery using artificial intelligence. Ann Vasc Surg. 2023;88:101–115.
14. Lo RC, et al. Machine learning prediction of endoleak after endovascular aneurysm repair. J Endovasc Ther. 2023;30(4):451–459.
15. Khan MA, et al. Machine learning prediction of stroke after carotid stenting. Eur J Vasc Endovasc Surg. 2022;63(3):412–420.
16. Li X, et al. Deep learning prediction of limb salvage in critical limb ischemia. Circ Res. 2022;131(5):501–512.
17. Park S, et al. Variability in machine learning model performance across vascular datasets: a systematic review. Front Cardiovasc Med. 2022;9:881045.
18. Singh N, et al. Challenges in reproducibility of machine learning–based vascular risk models. J Vasc Surg. 2022;76(2):223–233.
19. Subramanian SH, et al. Generalizability of machine learning models in vascular medicine. JAMA Netw Open. 2023;6(2):e225748.
20. Sun L, et al. Decision curve analysis in surgical machine learning models: implications for clinical utility. Ann Surg. 2023;278(4):620–628.
21. Collins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55–63.
22. Wolff RF, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–58.
23. Collins GS, et al. TRIPOD-AI: reporting guidelines for machine learning–based prediction models. Nat Med. 2023;29(4):983–990.
24. Wolff RF, et al. PROBAST-AI: assessment of bias in artificial intelligence–based prediction models. BMJ. 2024;385:q158.
25. Van Calster J, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
26. Hemmila KJ, et al. Machine learning for risk prediction in trauma and vascular surgery: a narrative review. J Surg Res. 2023;278:223–232.
27. Nguyen HT, et al. Validation of machine learning models for predicting perioperative complications after endovascular aneurysm repair. Eur J Vasc Endovasc Surg. 2024;68(1):101–110.
28. Patel MB, et al. Artificial intelligence–based risk stratification in aortic dissection surgery. Ann Thorac Surg. 2024;117(5):812–820.
29. Zhang Y, et al. Deep learning in vascular imaging: emerging opportunities. Radiology. 2022;302(2):217–229.
30. Greenberg B, et al. Integrating artificial intelligence predictions into vascular surgical decision support. J Vasc Surg. 2023;77(3):645–654.
31. Diaz DA, et al. Explainability and bias in vascular machine learning models. Front Artif Intell. 2023;7:115.

32. Edwards CR, et al. Cross-validation and model robustness in vascular artificial intelligence research. IEEE Access. 2023;11:23455–23468.
33. Rossi FM, et al. Future perspectives of machine learning in vascular surgery. **Eur J Vasc Endovasc Surg.** 2024;67(5):732–743.
34. Gupta TK, et al. Clinical adoption barriers of machine learning in vascular interventions. **Semin Vasc Surg.** 2023;36(1):55–67.
35. van Klaveren JPM, et al. Performance drift and validation gaps in machine learning–based outcome prediction. **Stat Methods Med Res.** 2024;33(2):210–226.
36. Thomas RG, et al. Ethical and regulatory considerations in AI-guided vascular care. **J Med Ethics.** 2024;50(1):22–30.
37. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **BMJ.** 2021;372:n71.
38. Collins GS, van Smeden M, Riley RD, et al. TRIPOD-AI and PROBAST-AI: protocol for developing reporting and risk-of-bias tools for prediction models developed using artificial intelligence. **BMJ Open.** 2021;11(7):e048008.
39. Fernández-Félix BM, Andaur-Navarro CL, Ramos EG, et al. A standardized framework for data extraction and risk-of-bias assessment in clinical prediction modelling studies: the CHARMS/PROBAST checklist. **BMC Med Res Methodol.** 2023;23:212.
40. Shamaki GR, Younis M, Bitar F, et al. Peripheral artery disease: a comprehensive updated review. **Curr Probl Cardiol.** 2022;47(11):101082. doi:10.1016/j.cpcardiol.2021.101082.
41. Li B, Aljabri B, Verma R, et al. Machine learning to predict outcomes following endovascular abdominal aortic aneurysm repair. **Br J Surg.** 2023;110(12):1840–1849. doi:10.1093/bjs/znad287.
42. Talebi S, et al. Prediction of endovascular leaks after thoracic endovascular aneurysm repair through machine learning applied to pre-procedural computed tomography angiographs. **Sci Rep.** 2020;10:18779.
43. Li B, Verma R, Eisenberg N, et al. Machine learning in vascular surgery: a systematic review and critical appraisal. **NPJ Digit Med.** 2022;5(1):7.
44. Nimmagadda N, Wang GJ, Hussain MA, et al. The role of artificial intelligence in vascular care: current applications and future directions. **J Vasc Surg Cases Innov Tech.** 2024;10(1):100123.
45. Li B, Warren BE, Eisenberg N, et al. Machine learning to predict outcomes of endovascular intervention for patients with peripheral artery disease. **JAMA Netw Open.** 2024;7(3):e242350. doi:10.1001/jamanetworkopen.2024.2350.
46. Matsuo K, Yamada K, Hayashi K, et al. Potential of machine learning to predict early ischemic events after carotid endarterectomy or stenting: comparison with surgeon predictions. **Neurosurg Rev.** 2022;45(5):2891–2901.
47. DeMartino RR, Eldrup-Jorgensen J, Nolan BW, et al. Development of a validated model to predict 30-day stroke and 1-year survival after carotid endarterectomy for asymptomatic stenosis using the Vascular Quality Initiative. **J Vasc Surg.** 2017;66(2):433–444.e2.
48. Grandi A, Ruggiero M, Attisani L, et al. Risk prediction models for peri-operative mortality in major vascular surgery, with a focus on ruptured abdominal aortic aneurysm: a scoping review. **J Clin Med.** 2023;12(17):5505.
49. Sampson UKA, Norman PE, Fowkes FGR, et al. Global and regional burden of aortic aneurysm, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. **Lancet.** 2014;383(9921):1405–1413.
50. Collins GS, Dhiman P, Andaur-Navarro CL, et al. The TRIPOD+AI statement: guidelines for reporting machine learning–enabled medical prediction models. **BMJ.** 2024;385:e078378.