

Machine learning-based detection of endoleak following endovascular aneurysm repair: A systematic review and meta-analysis

Callum Ward, BSc, MSc, MBBS,^a Ahmed Boalot, MBBS,^b Danesh Haemmanaath, MBBS,^c Mohamad Hamady, MBBS,^d Ahmed Abdaldayem, MBBS,^e and Bella Hausen, MBBS,^f *Preston, Huddersfield, Glasgow, London, and Manchester, United Kingdom*

ABSTRACT

Objective: To evaluate whether machine learning (ML) models can detect endoleak following endovascular aneurysm repair (EVAR) and quantify their diagnostic accuracy.

Methods: A systematic review and meta-analysis were conducted in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. PubMed, Web of Science, Embase, and Scopus were searched from inception to August 2025 for studies developing or validating ML models for endoleak prediction after EVAR. Eligible studies included adult patients undergoing endovascular aortic repair. Data extraction was performed independently by two reviewers, and risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool. Pooled diagnostic performance metrics were estimated using random-effects meta-analysis, including a bivariate Reitsma model to jointly synthesize sensitivity and specificity.

Results: Thirteen studies comprising 4051 patients, including 1641 with endoleaks, were included. Reported ML approaches encompassed logistic regression, random forests, support vector machines, and deep learning models. The pooled area under the receiver operating characteristic curve was 0.89 (95% confidence interval [CI], 0.83-0.94). Pooled sensitivity was 0.86 (95% CI, 0.75-0.92), specificity 0.87 (95% CI, 0.83-0.91), and accuracy 0.85 (95% CI, 0.80-0.90). Substantial heterogeneity was observed across all pooled estimates ($I^2 > 85\%$). All included studies were judged to be at high overall risk of bias, primarily because of retrospective single-center designs, lack of external validation, and inconsistent handling of missing data.

Conclusions: ML-based models demonstrate high average diagnostic performance for predicting endoleak after EVAR; however, substantial heterogeneity and pervasive risk of bias limit their current clinical applicability. Future research should prioritize large, multicenter prospective validation studies, methodological standardization, and rigorous assessment of clinical impact before routine implementation. (JVS-Vascular Insights 2026;4:100508.)

Keywords: Abdominal aortic aneurysm; Detection model; Endoleak; Endovascular surgery; Machine learning; Vascular surgery

From the University of Lancashire Medical School, University of Lancashire, Preston^a; the Calderdale and Huddersfield NHS Foundation Trust, Huddersfield^b; the University of Glasgow, University Ave, Glasgow^c; the Queen Elizabeth the Queen Mother Wing (QEQM), St Mary's Campus, London^d; the Manchester Royal Infirmary, Oxford Rd, Manchester^e; and the Lancashire Teaching Hospitals, Royal Preston Hospital, Preston.^f

C.W. and A.B. to be considered co-first authors.

Additional material for this article may be found online at www.jvsvi.org.

Correspondence: Callum Ward, BSc, MSc, MBBS, School of Medicine, University of Lancashire, Preston PR1 7BH, United Kingdom (e-mail: Cward8@lancashire.ac.uk).

The editors and reviewers of this article have no relevant financial relationships to disclose per the Journal Policy that requires reviewers to decline review of any manuscript for which they may have a conflict of interest.

2949-9127

© 2026 THE AUTHOR(S). Published by ELSEVIER INC. on behalf of the Society for Vascular Surgery. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.1016/j.jvsvi.2026.100508>

Although several machine learning (ML) models have been developed for predicting endoleak risk,¹ no comprehensive synthesis currently exists to evaluate these models. It remains unclear which algorithms have been used,² which clinical and imaging variables are most incorporated, how models perform based on key metrics, such as accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, and specificity,³ and whether these tools have been externally validated or adopted in clinical workflows. This knowledge gap limits both implementation and future research directions.

Despite advances in stent graft design, endoleaks remain a common complication that may necessitate reintervention or even contribute to rupture.⁴ Computed tomography angiography (CTA), interpreted by experienced vascular radiologists, remains the current gold standard for post-EVAR endoleak detection. Contrast-

enhanced ultrasound may provide complementary diagnostic information in selected patients but has not replaced CTA in routine surveillance.⁵ Ultrasound is safer but limited by operator dependency and reduced sensitivity.⁶ Current risk stratification relies on static clinical and anatomical parameters, failing to capture complex interrelationships between variables.⁷

ML techniques may address these limitations by synthesizing diverse data sets to generate more accurate, individualized detections.⁸ By identifying patients at increased risk for specific endoleak types, ML could support earlier imaging, better resource allocation, and timelier reintervention.⁹

This review aims to systematically identify and summarize ML approaches used to predict endoleak risk following endovascular aneurysm repair (EVAR).¹⁰ It will evaluate model performance using metrics such as accuracy, AUC, sensitivity, and specificity, and determine which clinical, procedural, and imaging features are most frequently incorporated into detective modeling. In addition, it will critically appraise the methodological quality of existing studies to identify current limitations and inform future research.¹¹

The purpose of this study is to evaluate whether ML models can detect endoleak following EVAR and to quantify their diagnostic accuracy.

METHODS

Search strategy and study selection. A literature search was conducted in alignment with Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. To ensure a high-quality systematic review, we also utilized the AMSTAR-2 checklist as a guide. Four databases were extensively searched utilizing search strings related to “Machine Learning,” “Endoleak,” and “Abdominal Aortic Aneurysm.” Databases searched include PubMed, Web of Science, Embase via Ovid, and Scopus from inception to August 2025. No restrictions were applied to either language or date of publication. Types of studies included were randomized controlled trials, nonrandomized trials, and observational studies (prospective and retrospective). The protocol was registered in PROSPERO (CRD420251054492). The review included studies on adult patients undergoing endovascular aortic repair, including infrarenal, juxtarenal, and thoracoabdominal aneurysms. Studies including children or animal models will be excluded, as well as in vitro models, conference papers, and preprints. Duplicates were removed, and screening of articles was carried out by two independent authors (C.W. and D.H.), and any disagreements were discussed with a senior reviewer (A.B.). The screening process is shown in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart, which can be found in [Fig 1](#).

ARTICLE HIGHLIGHTS

- **Type of Research:** Systematic review and meta-analysis
- **Key Findings:** Across 13 studies involving 4051 patients, machine learning models demonstrated strong diagnostic performance for predicting endoleak after endovascular aneurysm repair, with a pooled area under the receiver operating characteristic curve of 0.89, sensitivity of 0.86, and specificity of 0.87. However, all studies showed a high risk of bias and substantial heterogeneity.
- **Take Home Message:** Machine learning models show promise for improving post-endovascular aneurysm repair surveillance but require multi-center prospective validation before routine clinical implementation.

Eligibility criteria. This systematic review included studies involving adult patients undergoing EVAR for abdominal aortic aneurysm (AAA), in which ML-based models were used to detect post-EVAR endoleak occurrence. Comparisons included traditional risk stratification methods or the absence of predictive modeling, with outcomes focused on the detection of any endoleak subtype. Eligible studies were peer reviewed, published in English, and employed any form of ML or artificial intelligence approach, including supervised, unsupervised, and deep learning (DL) models. Studies not utilizing ML techniques, as well as animal studies, in vitro or purely theoretical models without clinical application, case reports, editorials, and narrative reviews, were excluded.

Ethics and institutional review board statement. As this study was a systematic review and meta-analysis of previously published data, institutional review board approval and informed consent were not required.

Data extraction. Data extraction was performed using a standardized Microsoft Excel spreadsheet following study selection via Rayyan, an online tool for systematic review screening. Two reviewers independently screened all titles, abstracts, and full-text articles according to predefined eligibility criteria. Any discrepancies in inclusion of decisions were resolved through discussion with a third reviewer. For each included study, the following data were extracted: first author, journal of publication, study design, year of publication, database utilized, model development stage, type of ML algorithm, study aim, total number of patients, and number of patients with endoleak. Diagnostic performance metrics, including sensitivity, specificity, accuracy, F1 score, positive predictive value, negative predictive value, and AUC, are recorded. All extracted

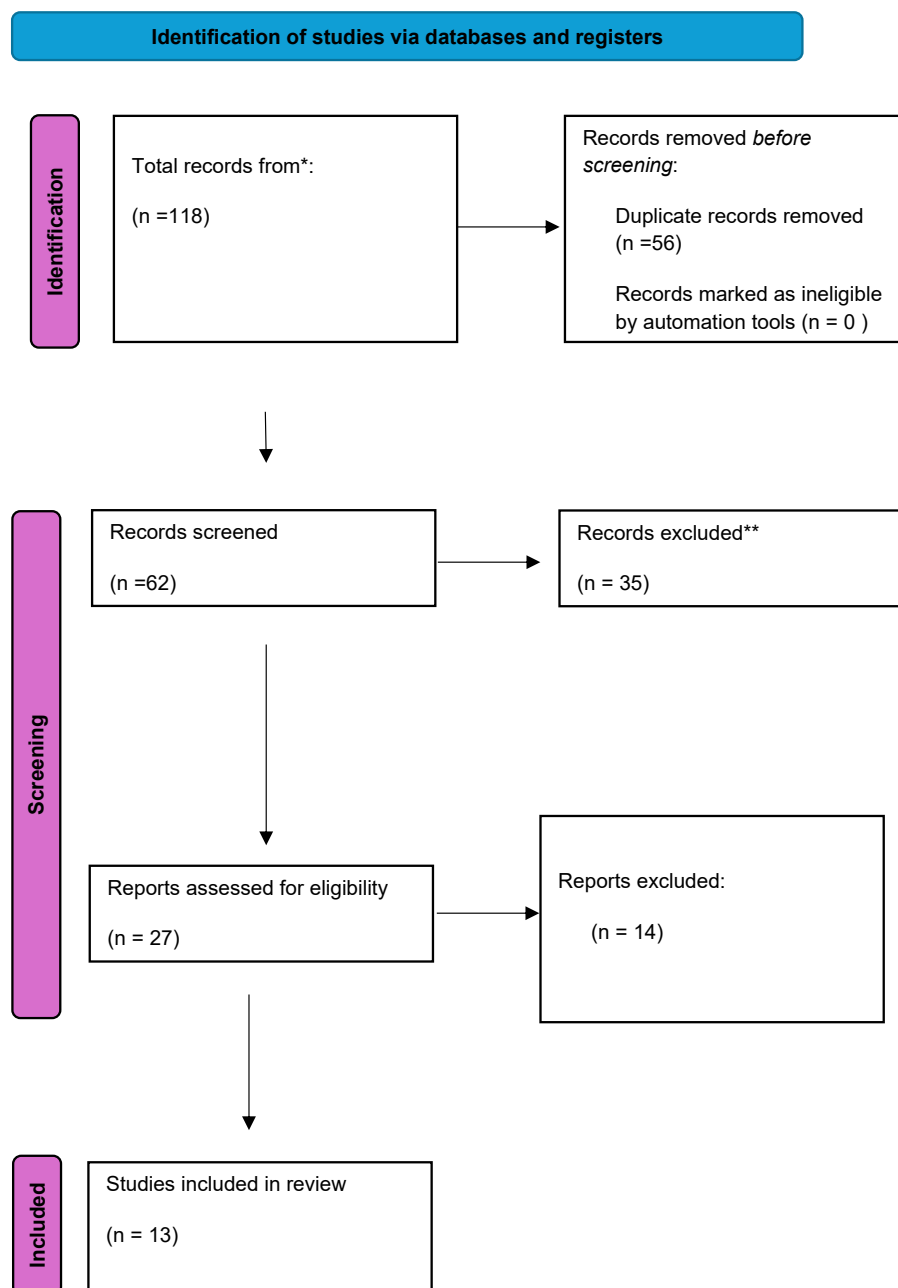


Fig 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram. PRISMA flow diagram illustrating study identification, screening, eligibility assessment, and final inclusion of studies in the systematic review and meta-analysis.

data were cross-checked by a second reviewer for completeness and consistency.

Quality assessment. Quality assessment was completed by two reviewers independently. The Prediction Model Risk of Bias Assessment Tool (PROBAST) was used to assess the risk of bias and applicability of detection model/algorithm studies.

Statistical analysis. All statistical analyses were performed in accordance with established methodological guidance for diagnostic test accuracy meta-analyses.

Descriptive statistics were summarized as frequencies and proportions for categorical variables.

For quantitative synthesis, pooled estimates of AUC were calculated using a random-effects model (restricted maximum likelihood estimation) with inverse-variance weighting. Standard errors (SEs) were derived from reported confidence intervals (CIs) where available or approximated using binomial variance methods when necessary.

Sensitivity, specificity, and accuracy were pooled using random-effects proportional meta-analysis. For sensitivity,

true positives were divided by total endoleak-positive cases; for specificity, true negatives were divided by total endoleak-negative cases. Accuracy was defined as:

$$(True\ positives + True\ negatives) / Total\ sample\ size$$

To jointly synthesize sensitivity and specificity while accounting for their inherent correlation and between-study heterogeneity, a bivariate random-effects meta-analysis based on the Reitsma model was performed. Logit-transformed sensitivity and false positive rate ($1 - specificity$) were modeled as correlated outcomes. A hierarchical summary receiver operating characteristic (HSROC) curve was constructed, including 95% confidence and prediction regions.

Between-study heterogeneity was quantified using Cochran's Q statistic, τ^2 (between-study variance), and I^2 statistics. Prediction intervals (PIs) were calculated to estimate the range of effects expected in future comparable studies.

Small-study effects and potential publication bias were assessed using funnel plot visual inspection, Egger's regression test, Begg's rank correlation test, and trim-and-fill analysis.

All meta-analyses were conducted in R (RStudio, version 4.0.1) using the mada, meta, and metafor packages. Descriptive analyses and data management were performed using Microsoft Excel (Microsoft Corp). Statistical significance was defined as $P < .05$.

The 13 included studies were published between 2019 and 2025 and conducted across diverse geographical settings, including Asia (China, Japan), Europe (UK, Netherlands, and Greece), and North America (USA). Study designs were predominantly single-center retrospective analyses, with sample sizes ranging from 56 to 493 patients. The proportion of patients with endoleaks varied across cohorts, from 19 cases (24%) in Egypt to 99 cases (30%) in USA, with the largest data set originating from China ($n = 493$, 80 endoleaks).

All 13 studies utilized CTA as the primary imaging modality, whereas ultrasound-based imaging was not used for ML model development in the included studies. All included studies used preoperative, early postoperative, or follow-up imaging data to train or validate detection models, depending on study design. Several studies integrated clinical or procedural variables,¹²⁻¹⁴ whereas others incorporated radiomics features¹⁵ to enhance detection. Most models were trained to detect any type of endoleak, though a subset specifically targeted type II endoleaks.¹⁶

Earlier studies primarily employed traditional supervised learning algorithms including logistic regression, support vector machines, and random forests, whereas more recent investigations increasingly utilized DL architectures, such as convolutional neural networks and RetinaNet. In addition to algorithmic differences, studies varied substantially in predictor selection, with models incorporating combinations of CTA-derived imaging

features, radiomic parameters, anatomical measurements, procedural characteristics, and clinical variables.

Across studies, the discriminative performance of ML/DL models was consistently strong, with reported AUC values ranging from 0.85 to 0.94 and accuracy between 0.81 and 0.89. Sensitivity and specificity were generally above 0.80, although one study¹⁷ reported a lower AUC of 0.71 despite maintaining high sensitivity (0.89) and specificity (0.91).

RESULTS

A total of 13 studies were included within this systematic review and meta-analysis, published between 2015 and 2025 (present). Sample sizes ranged from small, single-center cohorts (<100 patients) to larger data sets (>300 patients). A total of 1641 patients were included. All studies investigated ML- or artificial intelligence-based predictive models developed for endoleaks in patients following AAA repair using an endovascular approach. The studies employed a variety of ML algorithms, including logistic regression, support vector machines, random forests, and DL-based approaches (a table of baseline characteristics can be found within the tables and figures section, under the [Table](#)).

Risk of bias. The risk of bias was assessed across all included studies using the PROBAST framework. This can be viewed within [Supplementary Table 1](#) (online only), evaluating four domains: participants, predictors, outcome, and analysis. Across the data set, all 13 studies were rated as high risk of bias overall, primarily because of lack of external validation, single-center retrospective designs, and limited reporting of calibration and missing data handling.

In the predictors and outcome domain, all studies used standard and prespecified definitions for endoleak, with outcome assessors blinded to predictor information where applicable. These domains were consistently rated as low risk.

The participants and analysis domain showed a high risk. A small subset of studies did not explicitly report handling of missing data or did not perform external validation, resulting in some concerns.

Study characteristics and analysis. The mean or median age of included populations ranged from the mid-60s to early 70s, with a predominance of male patients, reflecting the known epidemiology of AAAs. Most studies included patients treated for infrarenal AAA, although a small number incorporated juxtarenal or thoracoabdominal aneurysms (a table of median/mean ages of patients, as well as the number of endoleak patients per study, can be found respectively within [Supplementary Tables II and III](#), online only).

Endoleak outcome detection also varied: some studies modeled any endoleak, whereas others specifically

Table. Baseline characteristics in included studies

First author (year)	Country/setting	Sample size (n)	EL patients (n)	ML algorithm(s)	Features used	EL detection	Key metrics reported
Wang et al (2022) ¹⁵	China (single center)	493	80 (17 Ia, 18 Ib, 3 III, and 42 II)	Logistic regression, RF, and SVM	Radiomics + CTA	Type I, II, and III	AUC, 0.93; accuracy, 0.86; and sensitivity, 0.89
Wang et al (2025) ¹⁶	Belgium (single center)	94 (75 train, 15 test)	90	CNN (DL)	CTA imaging	Type II	AUC, 0.917, accuracy, 0.84
Hahn et al (2020) ¹	USA	334	99	RetinaNet CNN (DL)	CTA	Any EL	AUC, 0.94; accuracy, 0.89; sensitivity, 0.90; and specificity, 0.88
Coatsaliou et al (2024) ¹⁷	UK	56	22 (39.3%)	CNN (AI software)	CTA (PRAEVAorta)	Any EL	AUC, 0.71; sensitivity, 0.89; and specificity, 0.91
Hu et al (2024) ¹⁸	China	216	64	RF, SVM, and logistic regression	CTA + clinical	Any EL	AUC, 0.89; sensitivity, 0.86; and specificity, 0.84
Van Veldhuizen et al (2024) ¹⁹	Netherlands	147	147	SVM, RF	CTA morphology	Type II	AUC, 0.88; accuracy, 0.83
Hahn et al (2019) ²⁰	USA	760	334	RF, DT, and SVM	Clinical + CTA	Any EL	AUC, 0.87, accuracy, 0.82, and sensitivity, 0.84
Charalambous et al (2022) ²¹	Greece	944	15	CNN	CTA	Any EL	AUC, 0.92; accuracy, 0.88
Peng et al (2024) ²²	China	381	381	LR, RF	CTA + clinical	Any EL	AUC, 0.86; sensitivity, 0.82; and specificity, 0.85
Yang et al (2024) ²³	China	245	167	DL (CNN)	CTA	Type II	AUC, 0.90; accuracy, 0.85
Talebi et al (2020) ²⁴	USA	50	20	RF, DL	CTA + clinical	Any EL	AUC, 0.91; accuracy, 0.87; and sensitivity, 0.85
Masuda et al (2022) ²⁵	Japan	111	111	DL (CNN)	CTA	Any EL	AUC, 0.92; accuracy, 0.88
Smorenburg et al (2025) ²⁶	Netherlands	220	111	LR, RF	CTA + procedural	Any EL	AUC, 0.85; accuracy, 0.81

AI, Artificial intelligence; AUC, area under the receiver operating characteristic curve; CNN, convolutional neural network; CTA, computed tomography angiography; DL, deep learning; DT, decision tree; EL, endoleak; LR, logistic regression; ML, machine learning; RF, random forest; SVM, support vector machine.

Baseline characteristics of included studies evaluating machine learning models for EL detection following endovascular aneurysm repair. Data include study setting, sample size, number of EL cases, machine learning algorithms used, feature types incorporated (clinical, radiomic, imaging, and procedural), targeted EL subtype, and reported diagnostic performance metrics.

focused on type I or type II endoleaks because of their differing clinical significance.

Overall, the methodological quality of the included studies was variable, with many at risk of bias because of small sample sizes, a lack of external validation, and inconsistent reporting of model development. Nonetheless, all studies demonstrated the feasibility of using ML

to detect endoleak risk, with promising diagnostic accuracy.

We pooled AUC (primary; sensitivity/specificity/accuracy secondary) using random-effects (restricted maximum likelihood estimation, inverse variance). SEs came from reported CIs or were approximated as $\sqrt{[p(1-p)/n]}$; heterogeneity was quantified with Cochran's

Q , τ^2 , and I^2 . Small-study effects were examined by funnel plot, Egger/Begg test, and trim-and-fill method. All analyses were done in RStudio. Some studies contributed only AUC data and were therefore included in the meta-analysis but not in baseline characteristic comparisons.

Within Fig 2, visual inspection of the funnel plot of AUC vs SE suggested mild asymmetry, with a relative paucity of small, imprecise studies reporting lower AUCs. Egger's regression indicated evidence of small-study effects ($t = -2.47$; $df = 11$; $P = .031$), whereas Begg's rank correlation was not significant (Kendall's $\tau = -0.35$; $P = .098$). The trim-and-fill method did not impute any missing studies on the right (0 added), and the pooled AUC was unchanged (0.889; 95% CI, 0.842-0.936). Given the very high between-study heterogeneity ($I^2 = 95\%$) and that AUC is a bounded metric, these signals should be interpreted cautiously, as asymmetry may reflect heterogeneity rather than publication bias per se.

Within Fig 3, nine studies contributed accuracy data. The pooled accuracy was 0.85 (95% CI, 0.80-0.90), with a 95% PI of 0.67 to 0.94, indicating that performance in a new study could plausibly fall anywhere within this range. Between-study heterogeneity was considerable ($I^2 = 85.8\%$; Q -test $P < .0001$; $\tau^2 = 0.1772$), consistent with meaningful variability in case-mix, model types, and validation strategies across studies.

Within Fig 4, 13 studies contributed to AUC estimates. The pooled AUC was 0.89 (95% CI, 0.83-0.94). The 95% PI was 0.69 to 1.00 (upper bound truncated at 1.00 for display; the analytical PI slightly exceeded 1.0 because of the SE approximation used for some studies). Between-study heterogeneity was considerable ($I^2 = 98.1\%$; $\tau^2 = 0.0078$; $P < .0001$), indicating marked dispersion in discriminative performance across data sets and modeling approaches.

Within Fig 5, 11 studies contributed sensitivity data. The pooled sensitivity was 0.86 (95% CI, 0.75-0.92), with a 95% PI of 0.50 to 0.97. Between-study heterogeneity was considerable ($I^2 = 92.7\%$; $\tau^2 = 0.5739$; $P < .0001$), indicating wide dispersion of sensitivity estimates across data sets and modeling approaches.

Within Fig 6, eight studies ($n = 943$). The pooled specificity was 0.87 (95% CI, 0.83-0.91), with a 95% PI of 0.75 to 0.94. Heterogeneity was moderate ($I^2 = 51.1\%$; $\tau^2 = 0.0960$; $Q P = .0458$), indicating some between-study variation but broadly consistent ability of models to correctly classify nonendoleak cases.

Within Supplementary Fig (online only), an HSROC curve derived from the bivariate random-effects (Reitsma) model. The solid ellipse represents the 95% confidence region around the pooled estimate; the dashed ellipse represents the 95% prediction region.

Across the eight studies reporting both sensitivity and specificity, the bivariate random-effects (Reitsma) model estimated a pooled sensitivity of 0.88 (95% CI, 0.80-0.93)

and a specificity of 0.87 (95% CI, 0.83-0.91). The HSROC (Supplementary Fig, online only) sits toward the upper left of the plot, with a tight 95% confidence region around the summary point and a wider 95% prediction region, indicating high average performance with moderate between-study variability, suggesting performance may differ across settings and underscoring the need for external validation.

DISCUSSION

Our systematic review and meta-analysis provide a quantitative synthesis of ML-based models developed to predict endoleak following EVAR. Across 13 studies, pooled diagnostic performance was high, with an AUC of 0.89 and balanced sensitivity and specificity. These findings indicate that ML approaches can discriminate between scans or patients with and without endoleak within the data sets on which they were developed. However, this performance should be interpreted as an average benchmark across heterogeneous studies rather than evidence of ready clinical utility. Importantly, the principal contribution of this meta-analysis is not simply confirmation that ML models perform well, but the provision of pooled diagnostic accuracy estimates across heterogeneous studies together with a formal assessment of methodological quality and between-study heterogeneity. These findings demonstrate that although reported diagnostic performance is consistently encouraging, the overall certainty of evidence remains limited by a high risk of bias and poor external validation.²⁷

Importantly, most included models were trained to detect the presence of endoleak on postprocedural imaging rather than to prospectively predict future endoleak risk. As such, claims regarding improved surveillance strategies or long-term outcomes must be interpreted cautiously. None of the included studies evaluated downstream clinical outcomes, tested surveillance pathways, or assessed whether ML-guided follow-up. Similarly, very few studies investigated discordant cases in which ML detected endoleaks that were not identified by radiologists. Consequently, it remains unknown whether these additional detections represent clinically meaningful findings requiring intervention or incidental abnormalities without therapeutic consequence, reintervention rates, radiation exposure, or costs. In addition, false-negative detections are of particular clinical concern, as missed high-risk endoleaks (eg, type I or III) may lead to catastrophic consequences. Furthermore, the available evidence does not support equivalent diagnostic performance across all endoleak subtypes. Most studies evaluated detection of any endoleak, whereas others focused specifically on type II endoleaks, limiting meaningful comparison between subtypes.²⁸⁻³⁰

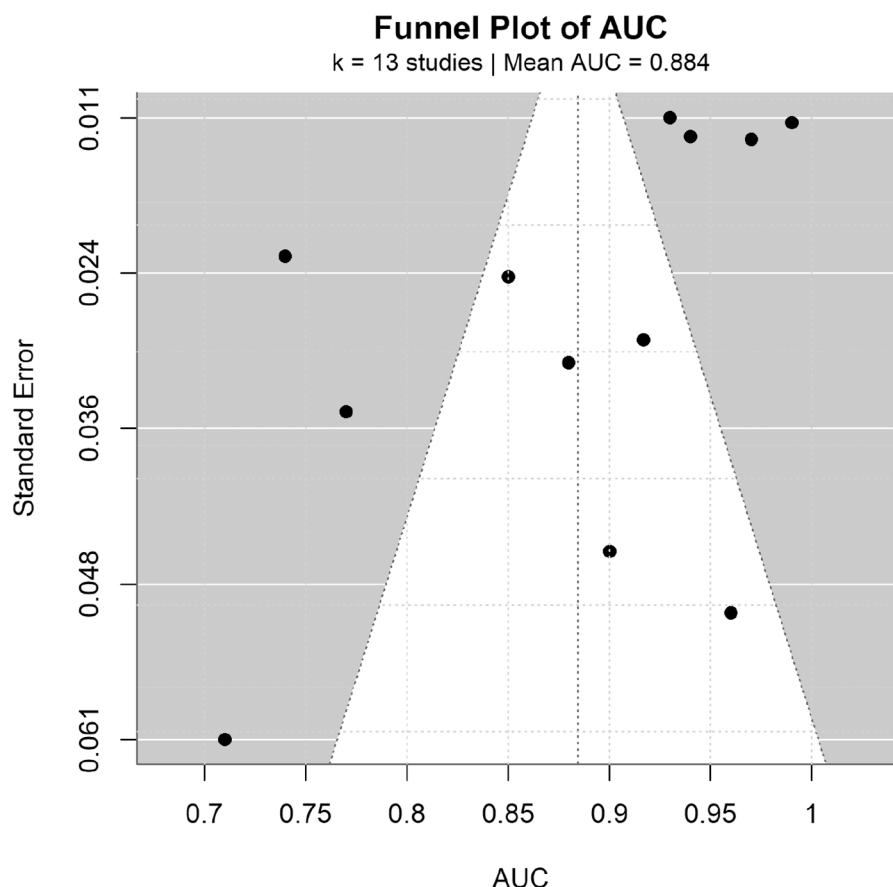


Fig 2. Funnel plot of area under the receiver operating characteristic curve (AUC). Funnel plot of AUC vs standard error to assess small-study effects. The vertical line represents the pooled AUC estimate. Asymmetry may reflect publication bias or between-study heterogeneity.

The apparently strong performance of many models likely reflects several methodological factors. Most studies used retrospective, single-center data sets with limited case-mix variation and high-quality CTA, where endoleak features may be visually conspicuous. In such settings, ML models may primarily reduce inter-reader variability or automate recognition of obvious radiological patterns rather than uncover novel prognostic signals. Few studies directly compared ML models with expert vascular radiologists using the same imaging data sets, preventing any robust conclusions regarding superiority or noninferiority. Consequently, current evidence supports ML as a potential adjunct to radiologist interpretation rather than a replacement for expert clinical assessment.²³

A key finding of this meta-analysis is the substantial heterogeneity across studies, reflected in very high I^2 values and wide detection intervals. The included models differed fundamentally in algorithm type, feature sets, imaging inputs, outcome definitions, and validation strategies. Consequently, pooled AUC and summary estimates largely reflect methodological variability rather than reproducible clinical performance. There is no guarantee that any individual model, or a

newly developed model, would achieve the meta-analytic AUC observed here. The pooled estimates should therefore be interpreted as descriptive averages rather than targets for implementation.

Future research should prioritize the development of single, well-designed ML models using large, multicenter data sets with robust external validation, transparent reporting, and prospective evaluation. Emerging directions include longitudinal models incorporating serial imaging, multi-modal approaches combining clinical, procedural, and imaging data, and assessment of clinical impact through decision-analytic or implementation studies. To date, ML-based endoleak detection remains primarily a research tool rather than a part of routine clinical practice. No included study evaluated widespread implementation within established surveillance pathways.^{18,31,32}

One of the identified strengths of this systematic review is its comprehensive methodology. We performed an extensive literature search across multiple databases (PubMed, Web of Science, Embase, and Scopus) from inception to April 2025 to reduce the likelihood of missing relevant studies as much as possible. Study screening, data extraction, and risk-of-bias assessment were conducted independently by two reviewers, which

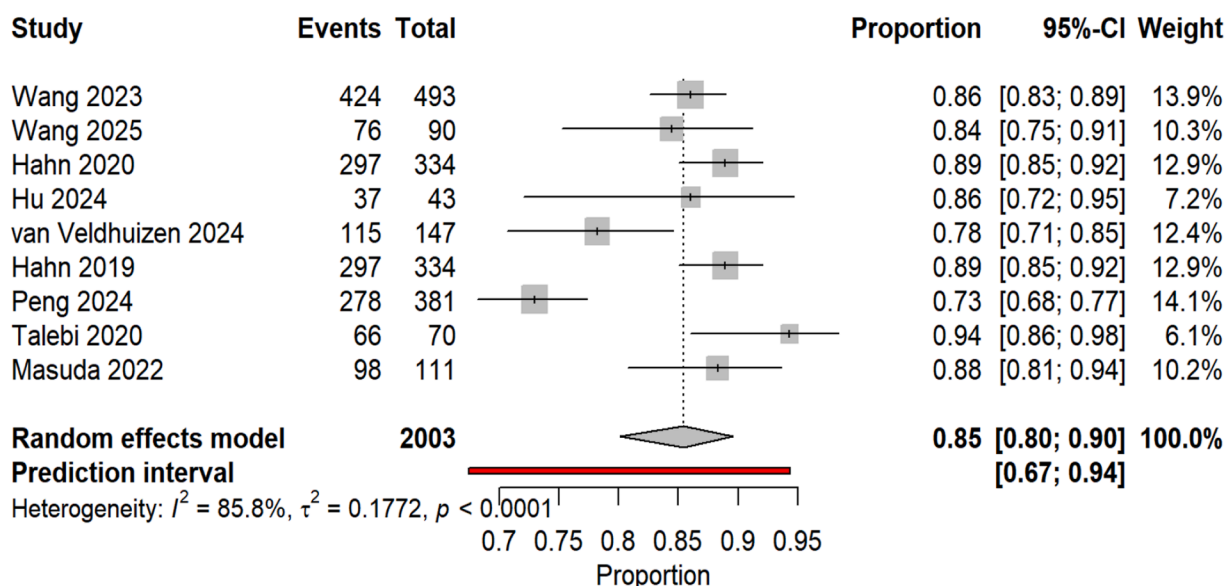


Fig 3. Table of diagnostic accuracy. Forest plot of diagnostic accuracy across included studies. *Squares* represent individual study estimates weighted by inverse variance; *horizontal lines* represent 95% confidence intervals (CIs). The *diamond* represents the pooled random-effects estimate with 95% CI and prediction interval (PI).

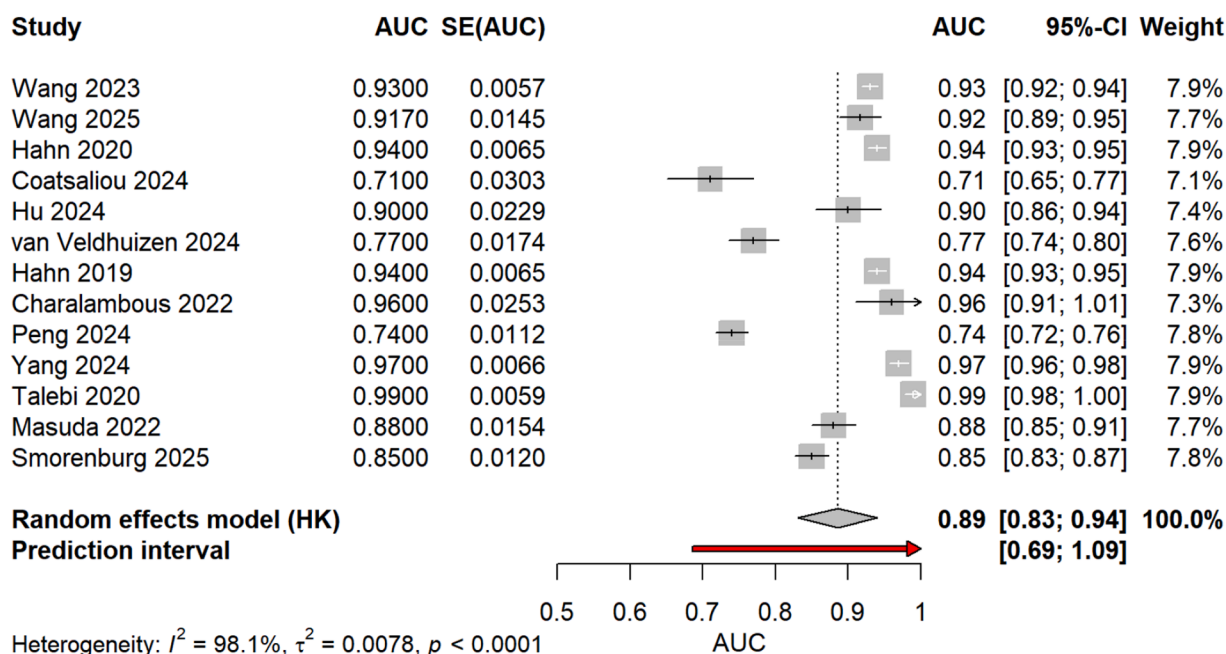


Fig 4. Table of area under the receiver operating characteristic curve (AUC). Forest plot of AUC for machine learning (ML) models detecting endoleak. The pooled estimate was calculated using a random-effects model (restricted maximum likelihood estimation). SE, Standard error.

minimizes subjectivity and enhances reliability. As per our knowledge, this systematic review is the first meta-analysis that specifically looks at ML-based detection models for endoleaks following EVAR and is clinically relevant at the time of writing because of the assimilation of the current evidence landscape.

However, several limitations should be acknowledged. First, all included studies were not considered to be of

very high quality based on the high overall risk of bias using the PROBAST. Second, significant variability ($I^2 > 85\%$) among studies in areas such as algorithm selection, imaging techniques, outcome definitions, and validation approaches limited the ability to interpret combined estimates. Third, inconsistent management of missing data and an absence of calibration reporting further reduce confidence in the models' reproducibility. Finally, while

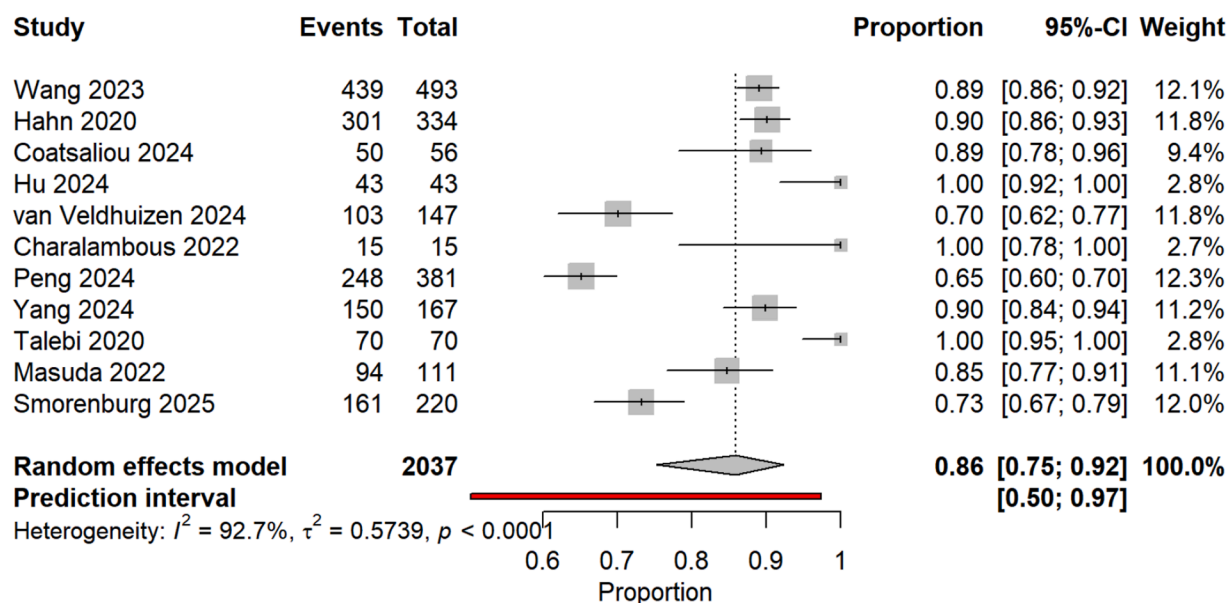


Fig 5. Table of sensitivity of machine learning (ML) models. Forest plot of pooled sensitivity estimates across included studies using random-effects meta-analysis. Considerable heterogeneity was observed. *CI*, Confidence interval.

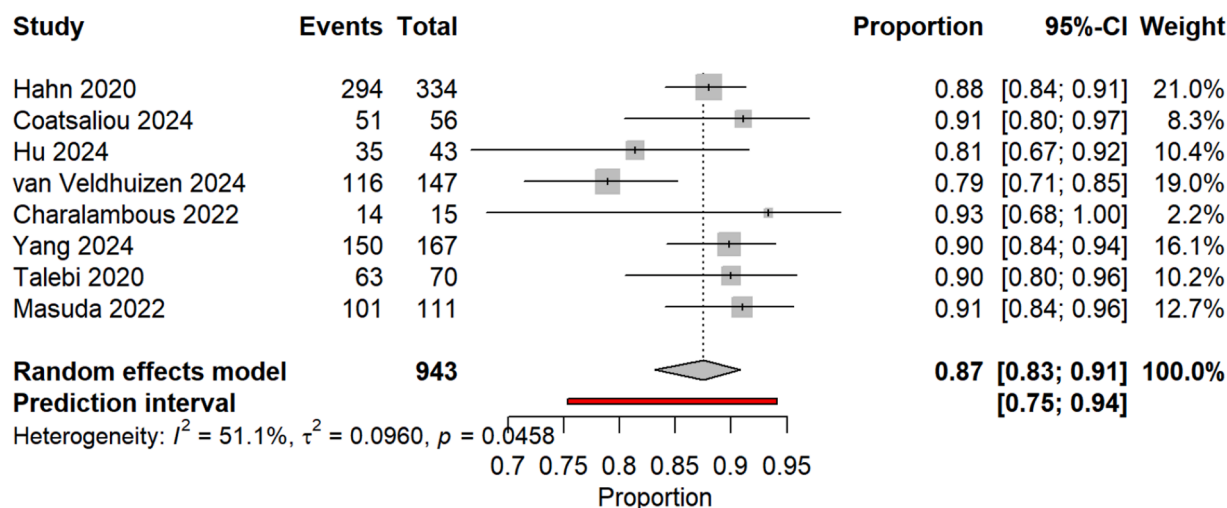


Fig 6. Table of specificity of machine learning (ML) models. Forest plot of pooled specificity estimates across studies using random-effects meta-analysis. Moderate heterogeneity was identified. *CI*, Confidence interval.

publication bias could not be formally evaluated because of the small number of studies, the potential for selective reporting of high-performing models remains.

CONCLUSIONS

In conclusion, this meta-analysis not only confirms the high diagnostic utility of ML models for predicting endoleak following EVAR but also highlights the substantial variability in their performance. This is the first quantitative meta-analysis focused specifically on ML-based detection of endoleak after EVAR. Although pooled diagnostic performance was encouraging, the available evidence does not currently support routine clinical implementation of

ML-based endoleak detection as a standalone diagnostic tool. Instead, ML should presently be regarded as an investigational adjunct to conventional radiological assessment until prospective multicenter validation and clinical impact studies have been completed.

AUTHOR CONTRIBUTIONS

Conception and design: CW, AA, BH
Analysis and interpretation: CW, AB, MH, AA, BH
Data collection: CW, AB, DH
Writing the article: CW, AB, DH
Critical revision of the article: CW, AB, MH, AA, BH
Final approval of the article: CW, AB, DH, MH, AA, BH

Statistical analysis: CW, AB
Obtained funding: Not applicable
Overall responsibility: CW

FUNDING

None.

DATA AVAILABILITY

All data generated or analyzed during this study are included in this published article.

DISCLOSURES

None.

REFERENCES

- Hahn S, Perry M, Morris CS, Wshah S, Bertges DJ. Machine deep learning accurately detects endoleak after endovascular abdominal aortic aneurysm repair. *JVS Vasc Sci*. 2020;1:5-12. <https://doi.org/10.1016/j.jvssci.2019.12.003>
- Dossabhoy SS, Ho VT, Ross EG, Rodriguez F, Arya S. Artificial intelligence in clinical workflow processes in vascular surgery and beyond. *Semin Vasc Surg*. 2023;36:401-412. <https://doi.org/10.1053/j.semvasc-surg.2023.07.002>
- Nishibe T, Iwasa T, Fukuda S, et al. Machine learning and abdominal aortic aneurysm: a new paradigm in prediction and prognosis after endovascular aneurysm repair. *Ann Vasc Dis*. 2026;19:2-6. <https://doi.org/10.3400/avd.ra.25-00120>
- Aras T, Tayeh M, Aswad A, Sharkawy M, Majd P. Exploring type IIIB endoleaks: a literature review to identify possible physical mechanisms and implications. *J Clin Med*. 2024;13:4293. <https://doi.org/10.3390/jcm13154293>
- Wang Y, Zhang Y, Li M, Liu Z, Ren J. Feasibility of contrast-enhanced ultrasound in the detection and classification of endoleaks after endovascular aneurysm repair. *Quant Imaging Med Surg*. 2024;14:6556-6565. <https://doi.org/10.21037/qims-24-758>
- Ceriani E, Schiavon R, La Cava L, Ruscitti C, Cogliati C. Point of care ultrasound: focus on evidence for a critical appraisal. *Eur J Intern Med*. 2025;141:106376. <https://doi.org/10.1016/j.ejim.2025.06.005>
- Patel S, Ormesher DC, Ray Smith S, et al. A risk-adjusted and anatomically stratified cohort comparison study of open surgery, endovascular techniques and medical management for juxtarenal aortic aneurysms—the UK COMPLEX Aneurysm Study (UK-COMPASS): a study protocol. *BMJ Open*. 2021;11:e054493. <https://doi.org/10.1136/bmjopen-2021-054493>
- Huang Y, Guo J, Chen W-H, et al. A scoping review of fair machine learning techniques when using real-world data. *J Biomed Inform*. 2024;151:104622. <https://doi.org/10.1016/j.jbi.2024.104622>
- Lareyre F, Raffort J, Kakkos SK, et al. Imaging analysis using Artificial Intelligence to predict outcomes after endovascular aortic aneurysm repair: protocol for a retrospective cohort study. *BMJ Open*. 2025;15:e098724. <https://doi.org/10.1136/bmjopen-2024-098724>
- Baek S, Arzani A. Current state-of-the-art and utilities of machine learning for detection, monitoring, growth prediction, rupture risk assessment, and post-surgical management of abdominal aortic aneurysms. *Appl Eng Sci*. 2022;10:100097. <https://doi.org/10.1016/j.apples.2022.100097>
- Calster BV, Collins GS, Vickers AJ, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *Lancet Digit Health*. 2025;7:100916. <https://doi.org/10.1016/j.landig.2025.100916>
- Saqib NU. Managing type II endoleaks—endovascular today. *Endovascular Today*. 2025. <https://evtoday.com/articles/2013-feb/managing-type-ii-endoleaks>
- Shaw PM, Loree J, Gibbons RC. *Abdominal Aortic Aneurysm*. StatPearls Publishing; 2022. <https://pubmed.ncbi.nlm.nih.gov/29262134/>
- Song P, He Y, Adeloey D, et al, Group (CHERG), on behalf of the G. H. E. R. The global and regional prevalence of abdominal aortic aneurysms: a systematic review and modelling analysis. *Ann Surg*. 2022;277:912-919. <https://doi.org/10.1097/SLA.0000000000000576>
- Wang Y, Zhou M, Ding Y, et al. A radiomics model for predicting the outcome of endovascular abdominal aortic aneurysm repair based on machine learning. *Vascular*. 2022;31:654-663. <https://doi.org/10.1177/17085381221091061>
- Wang Y, Zhou M, Ding Y, et al. A deep learning model for predicting the outcome of persistent type 2 endoleaks after endovascular abdominal aortic aneurysm repair. *Acta Chir Belg*. 2025;125:139-147. <https://doi.org/10.1080/00015458.2022.2129282>
- Coatsaliou Q, Lareyre F, Raffort J, et al. Use of artificial intelligence with deep learning approaches for the follow-up of infrarenal endovascular aortic repair. *J Endovasc Ther*. 2024. <https://doi.org/10.1177/15266028241252097>
- Hu G, Ding N, Wang Z, Jin Z. Unenhanced computed tomography radiomics help detect endoleaks after endovascular repair of abdominal aortic aneurysm. *Eur Radiol*. 2024;34:1647-1658. <https://doi.org/10.1007/s00330-023-10000-y>
- van Veldhuizen WA, de Vries J-PPM, Tuinstra A, et al. Machine learning based prediction of post-operative infrarenal endograft apposition for abdominal aortic aneurysms. *Eur J Vasc Endovasc Surg*. 2024;68:568-576. <https://doi.org/10.1016/j.ejvs.2024.07.003>
- Hahn S, Morris CS, Bertges DJ, Safwan W. Deep learning for recognition of endoleak after endovascular abdominal aortic aneurysm repair. 2022 IEEE 19th Int Symp Biomed Imaging (ISBI). 2019;63:759-763. <https://doi.org/10.1109/isbi.2019.875918>
- Charalambous S, Klontzas ME, Kontopodis N, et al. Radiomics and machine learning to predict asymptomatic type 2 endoleaks after endovascular aneurysm repair: a proof of concept. *Acta Radiol*. 2022;63:1293-1299. <https://doi.org/10.1177/02841851211032443>
- Peng C, He W, Wang S, et al. Predict the endoleak risk after EVAR based on multi-dimensional anatomical features before AAA surgery. 2024;435-440. <https://doi.org/10.1109/cis-ram61939.2024.10673221>
- Yang Q, Hu J, Luo Y, et al. Detection of endoleak after endovascular aortic repair through deep learning based on non-contrast CT. *Cardiovasc Intervent Radiol*. 2024;47:1267-1275. <https://doi.org/10.1007/s00270-024-03805-x>
- Talebi S, Madani MH, Madani A, et al. Machine learning for endoleak detection after endovascular aortic repair. *Sci Rep*. 2020;10:18343. <https://doi.org/10.1038/s41598-020-74936-7>
- Masuda T, Baba Y, Nakaura T, et al. Applying patient characteristics, stent-graft selection, and pre-operative computed tomographic angiography data to a machine learning algorithm: Is endoleak prediction possible? *Radiography*. 2022;28:906-911. <https://doi.org/10.1016/j.radi.2022.06.004>
- Smorenburg S, Hoksbergen J, Yeung KK, Wolterink JM. Multitask deep learning for automated detection of endoleak at digital subtraction angiography during endovascular aneurysm repair. *Radiol Artif Intell*. 2025;7:e240392. <https://doi.org/10.1148/ryai.240392>
- Stonko DP, Hicks CW. Mature AI/ML-enabled medical tools impacting vascular surgical care: a scoping review of late-stage, FDA approved/cleared technologies relevant to vascular surgeons. *Semin Vasc Surg*. 2023;36:460.
- Wazzan M, Abduljabbar A, Ajlan A, et al. Reference normal diameters of the abdominal aorta and common iliac arteries in the Saudi population. *Cureus*. 2022;14:e30695. <https://doi.org/10.7759/cureus.30695>
- White S, Stavropoulos S. Management of endoleaks following endovascular aneurysm repair. *Semin Intervent Radiol*. 2009;26:033-038. <https://doi.org/10.1055/s-0029-1208381>
- Yanamaladoddi VR, Sarvepalli SS, Vemula SL, et al. The challenge of endoleaks in endovascular aneurysm repair (EVAR): a review of their types and management. *Cureus*. 2023;15:e39775. <https://doi.org/10.7759/cureus.39775>
- Yang X, Chen Y-X, Zhang B, et al. Contrast-enhanced ultrasound in predict endoleaks with failed computed tomography angiography diagnosis after endovascular abdominal aortic aneurysm repair. *Chin Med J*. 2015;128:2491-2497. <https://doi.org/10.4103/0366-6999.164935>
- Zamin SA, Devalla R, Mitchell ME. Contained giant abdominal aortic aneurysm rupture secondary to type I and III endoleaks requiring open repair with endograft explantation. *Am Surg*. 2023;89:3623-3625. <https://doi.org/10.1177/00031348231167410>

Submitted Jun 9, 2026; accepted Jul 13, 2026.

Additional material for this article may be found online at <https://www.jvsvi.org>.