Supplementary Materials

AI-Driven CT Analysis for Aortic Disease: A Task-Based Systematic Review of Segmentation, Measurement, and Detection Applications

Tables S1–S5 are provided in the accompanying spreadsheet file (Table S1, search strategies; Table S2, reports excluded after full-text assessment; Tables S3–S5, data extraction for segmentation, measurement and detection studies). This file contains Tables S6–S8 (quality appraisal), Table S9 (subgroup summaries of performance statistics) and Table S10 (decision rules for the CLAIM-informed limitation tiers). Study labels are those used in Tables 1–3 of the main text; numbers in square brackets refer to the reference list of the main text.

Table S6. CLAIM-informed appraisal of segmentation studies (n = 42).

Study	Main limitations	Reference and data	Validation	Reporting	Limitation tier
Aortic dissection (n = 14)					
Abaid 2024 [17]	Single-source dataset; no reader agreement; no external test	+	+	−	Moderate
Cao 2019 [18]	Internal 30-patient test from one multicentre registry; exclusions for slice thickness and artefacts	−	+	−	Moderate
Chen 2021 [19]	Single-expert annotation; cross-validation only; centre allocation across folds not reported	+	+	−	Moderate
Feng 2023 [20]	Restricted post-operative population; no external test	+	+	−	Moderate
Guo 2024 [21]	24 patients; cross-validation only	−	+	−	Moderate
Hahn (LD) 2020 [22]	Only 8 test patients; internal test only	−	+	−	Moderate
Imran 2024 [23]	Small sample; cross-validation only	−	+	−	Moderate
Imran 2026 [24]	Challenge benchmark from a single data source	+	+	−	Moderate
Jung 2024 [25]	Sequential two-reader reference without agreement statistics; thick slices (2.5–5 mm)	−	−	+	Lower
Kesavuori 2023 [26]	Single-reader reference, partly CNN-assisted	+	+	−	Moderate
Lyu 2021 [27]	Unclear annotator qualifications; no reader variability; cross-validation only	+	+	−	Moderate
Wobben 2021 [28]	Small cohort; single-reader reference	+	+	−	Moderate
Yao 2021 [29]	Single-centre dataset; thrombus DSC inflated by thrombus-free scans	−	+	+	Moderate
Yu 2021 [30]	Thrombosed false lumen and difficult cases excluded	−	++	−	Higher
Abdominal aortic aneurysm (n = 21)					
Abdolmanafi 2022 [31]	Small, poorly characterised external cohort	−	−	+	Lower
Brutti 2022 [32]	14-scan internal test; single centre	−	+	−	Moderate
Caradu 2021 [33]	Internal validation only; proprietary software	−	+	+	Moderate
Chandrashekar 2023 [34]	Reference masks generated by the authors' own deep learning pipeline; tube-current exclusion threshold derived from test results	++	++	−	Higher
Derycke 2026 [35]	Multi-institutional but not institution-independent test; inconsistently described reference standard	+	+	+	Moderate
Fantazzini 2020 [36]	10-scan internal test; no multi-reader review	+	+	−	Moderate
Hendricks 2026 [37]	Infrarenal pre-EVAR population only; limits	−	−	+	Lower

Study	Main limitations	Reference and data	Validation	Reporting	Limitation tier
	of agreement shown graphically only				
Hwang 2022 [38]	Cross-validation only	−	+	−	Moderate
Jung 2022 [39]	Cross-validation only	−	+	−	Moderate
Lopez-Linares 2018 [40]	13 patients; single-surgeon reference; post-processing parameters set empirically within the same cross-validation	+	++	−	Higher
Lopez-Linares 2019 [41]	Reference generated by a prior model and then manually corrected; 16-scan test	++	+	−	Higher
Lyu 2023 [42]	10-scan internal test; single-operator reference	+	+	−	Moderate
Lyu 2024 [43]	30-scan test; reference may include the aneurysm wall	+	+	−	Moderate
Ma 2023 [44]	30 patients; cross-validation only; evaluation on fixed regions of interest	−	+	−	Moderate
Ma 2025 [45]	Weak labels for training; 30-patient cohort; cross-validation only	+	+	−	Moderate
Mu 2023 [46]	Small sample; no external test	−	+	−	Moderate
Roby 2025a [47]	Contradictory reference standard (manual versus in-house tool); training data source not stated	++	+	−	Higher
Roby 2025b [48]	In-house software output as reference; patient-level separation not stated	++	+	−	Higher
Salvi 2021 [49]	Threshold selected visually; very small sample	−	++	+	Higher
Siriapisith 2022 [50]	Reader agreement assessed on 40 single images only; partial volumes (64 slices); no external test	+	+	−	Moderate
Wang 2022 [51]	Single scanner; no independent reading	+	+	+	Moderate
Thoracic aortic aneurysm or dilatation (n = 2)					
Aghayeva 2025 [52]	Slice-level evaluation on classifier-selected slices	−	++	−	Higher
Comelli 2021 [53]	Single-operator semi-automatic reference; cross-validation only	+	+	−	Moderate
Mixed or unselected cohorts (n = 5)					
Chen 2022 [54]	Patient-level split not stated; test images counted after augmentation	−	++	−	Higher
El Rahal 2026 [55]	Qualitative reference; small subgroups	++	−	−	Higher
Liu 2026 [56]	Mostly healthy public data; slice-level split of a four-case private dataset	+	++	−	Higher
Sieren 2022 [57]	Single centre and scanner; internal test only	−	+	−	Moderate
Vagenas 2025 [58]	Public data from healthy subjects; pathological test set of 10 patients	+	+	−	Moderate

CLAIM was used descriptively across three areas (reference standard and data; validation and generalisability; reporting and reproducibility) and not as a risk-of-bias instrument. Limitation tiers are relative to this body of literature.

Totals: Lower, n = 3; Moderate, n = 28; Higher, n = 11. Area ratings: −, no concern; +, concern; ++, critical concern; definitions and tier rule in Table S10. Population restrictions are applicability notes and do not contribute to the tier.

Abbreviations: CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CNN, convolutional neural network; DSC, Dice similarity coefficient; EVAR, endovascular aneurysm repair.

Table S7. CLAIM-informed appraisal of measurement studies (n = 38).

Study	Main limitations	Reference and data	Validation	Reporting	Limitation tier
Aortic dissection (n = 4)					
Bäumler 2026 [59]	Small external test (30 CTAs); landmarks only, no diameters	−	−	−	Lower
Krebs 2024 [60]	No reference for volumes; zone-level DSC only; 4-fold cross-validation	++	+	−	Higher
Nakano 2026 [61]	20-patient test; single-reader reference	+	+	−	Moderate
Pepe 2023 [62]	12-scan test; landmark placed by the operator, with AI predicting the plane only	−	+	−	Moderate
Abdominal aortic aneurysm (n = 12)					
Broda 2025 [63]	No agreement statistic between AI and readers	−	+	++	Higher
Caradu 2022 [64]	Single setting; uncertain reference standard	+	−	−	Moderate
Coatsaliou 2024b [65]	AI output on the arterial phase used as reference; no human measurement	++	+	−	Higher
Hatzl 2024 [66]	Vendor training data not reported	−	−	+	Lower
Kim (T) 2024 [67]	Internal test; readers not stated; semi-automatic measurement	+	+	+	Moderate
Lareyre 2022 [68]	Uncertain variability of reference measurements	+	+	−	Moderate
Pouncey 2026 [69]	Two-centre sample of 50 CTAs; inconsistently reported number of operators	−	−	+	Lower
Robbi 2025 [70]	External test limited to 20 public CTAs	−	−	−	Lower
Sharma 2025 [71]	Reproducibility only; AI and reference compared by means only	++	−	++	Higher
van Tongeren 2025 [72]	Incompletely detailed reference-reader procedures	+	+	+	Moderate
Wegner 2025 [73]	Selected complex-repair population; single centre	−	+	−	Moderate
Weiss 2026 [74]	Primary-reviewer reference with documented inter-reviewer agreement (mean DSC 0.95); infrarenal EVAR population only	−	−	−	Lower
Thoracic aortic aneurysm or dilatation (n = 8)					
Fink 2024 [75]	Vendor prototype; origin of training data not reported	−	+	+	Moderate
Graby 2023 [76]	Clinical reports as reference; inconsistently reported size of the repeat cohort	+	−	+	Moderate
Katakol 2023 [77]	Scan-level split of serial CTAs; possible leakage	−	++	−	Higher
Livolsi 2026 [78]	Cross-validation only; readers not stated	+	+	−	Moderate
Lo Piccolo 2023 [79]	No agreement statistic; visual review of AI-flagged cases only	++	−	−	Higher
Marway 2025 [80]	Internal test only; readers not stated	+	+	−	Moderate
Pradella 2021 [81]	Clinical-report reference; research-reader subset of 21 cases	−	−	+	Lower
Rueckel 2021 [82]	18 patients; narrow workflow setting	−	+	−	Moderate
Mixed or unselected cohorts (n = 14)					
Adam 2021 [83]	Incompletely reported reference standard; internal test only	+	+	−	Moderate
Artzner 2022 [84]	Single centre; vendor training data not reported	−	+	−	Moderate

Study	Main limitations	Reference and data	Validation	Reporting	Limitation tier
Bratt 2021 [85]	Reproducibility only; AI and manual measurements compared by means only; manual plane selection	++	+	−	Higher
Gatinot 2026 [86]	32 patients; vendor training data not reported	−	+	−	Moderate
Hamelink 2023 [87]	Purposive sampling by age and diameter; second reader on a subsample only	−	−	+	Lower
Heo 2025 [88]	Trauma cohort without aortic disease; cross-validation only	+	+	−	Moderate
Hong 2026 [89]	Lumen reference versus outer-wall output; student annotators	+	−	−	Moderate
Koo 2025 [90]	Healthy adults only; AI-corrected reference	++	+	−	Higher
Park 2026 [91]	Model development not disclosed; ICC includes AI as a rater	−	+	+	Moderate
Paulsen 2026 [92]	Single-reader clinical measurements as reference; reader variability not assessed	+	−	−	Moderate
Postiglione 2024 [93]	Only part of the test set from a centre outside training; each reference segment measured by a single expert, inter-expert variability not assessed	+	−	−	Moderate
Saitta 2022 [94]	9 aneurysms in the test set; single-operator reference	+	+	−	Moderate
Tahoces 2021 [95]	Single centre; limited landmark agreement	+	+	−	Moderate
Yang 2023 [96]	Agreement shown only graphically; aortic disease excluded	+	+	++	Higher

CLAIM was used descriptively across three areas (reference standard and data; validation and generalisability; reporting and reproducibility) and not as a risk-of-bias instrument. Limitation tiers are relative to this body of literature.

Abbreviations: AI, artificial intelligence; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CTA, computed tomography angiography; DSC, Dice similarity coefficient; EVAR, endovascular aneurysm repair; ICC, intraclass correlation coefficient. Area ratings: −, no concern; +, concern; ++, critical concern; definitions and tier rule in Table S10. Population restrictions are applicability notes and do not contribute to the tier.

Totals: Lower, n = 7; Moderate, n = 22; Higher, n = 9.

Table S8. QUADAS-2/QUADAS-AI appraisal of detection studies (n = 38).

Study	Main concern	Overall risk of bias
Aortic dissection or acute aortic syndrome (n = 22)		
Chang 2025 [97]	Stratified, enriched sampling (predefined lesion locations and comorbidity quotas; 100/260 positive); severe artefacts excluded	High
Cheng 2024 [98]	Balanced AD/non-AD sampling; unclear centre separation	High
Dong 2024 [99]	Matched controls; small external test	High
Gao 2026 [100]	Protocol paper; no diagnostic accuracy reported	Not formally appraised
Harris 2019 [101]	Report-based reference without blinded re-reading	High
Hata 2021 [102]	1:1 case–control design; single-reader reference	High
Hu 2025 [103]	Low risk of bias; few events in the prospective pilot	Low
Huang 2022 [104]	Low risk of bias; applicability concern (single-centre temporal test)	Low
Kim (TB) 2024 [105]	Image-level split; possible leakage	High
Laletin 2024 [106]	Independence of test data from vendor training data not stated (overlapping periods)	Unclear
Liu 2024 [107]	Restrictive exclusions; single centre	High
Norajitra 2025 [108]	Enriched cohorts; external cases and controls drawn from different datasets	High

Study	Main concern	Overall risk of bias
Raj 2024 [109]	Case–control sampling; external cases and controls drawn from different public datasets	High
Tang 2026 [110]	1:1 matched controls; model selection partly based on external validation	High
Wada 2023 [111]	Enriched sample; record-based reference; single centre	High
Wang 2025 [112]	Low risk of bias; applicability concern (CTA-referred population)	Low
Wu 2026 [113]	Component detection; patient-level triage on an enriched external test set (71 TBAD, 20 controls)	High
Xiong 2022 [114]	Selected cohorts; incomplete labelling	High
Yang 2026 [115]	Reference process not described; no external test	High
Yi 2022 [116]	Propensity-matched controls (case–control design)	High
Zhang 2026 [117]	Non-consecutive AD/non-AD sampling; age and vendor imbalance between groups	High
Zhuang 2025 [118]	Index-test input differs by class (manual masks for all controls); cross-validation only	High
Abdominal aortic aneurysm (n = 8)		
Camara 2022 [119]	Propensity-matched case–control design; image-level split	High
Chatterjee 2024 [120]	Report-based reference with likely under-reporting; false positives not re-read	High
Chen 2025 [121]	Partial verification; false negatives not estimated	High
Golla 2021 [122]	Positives and negatives in the verification set drawn from different sources; specificity not reported	High
Mohammadi 2019 [123]	Patch-level metrics; sampling not described	High
Rengarajan 2020 [124]	Case–control groups (elective versus urgent repair); feature preselection possibly outside LOOCV	High
Spinella 2023 [125]	Small cohort with case–control-like sampling	High
Tomihama 2023 [126]	Case–control design; likely image-level leakage	High
Endoleak after EVAR (n = 4)		
Coatsaliou 2024a [127]	Cohort enriched for endoleak (acknowledged by the authors); single-reader reference; unit of analysis not stated	High
Hahn (S) 2020 [128]	Case–control sampling (all endoleaks plus random controls); report-based reference	High
Talebi 2020 [129]	Case–control sampling; single-reader reference	High
Yang 2024 [130]	Case–control groups; manual sac segmentation within the index test; cross-validation only	High
Thoracic aortic aneurysm or dilatation (n = 2)		
Macruz 2022 [131]	High prevalence by design; failed cases handled inconsistently across analyses	High
Pradella 2022 [132]	Only AI-positive cases re-read; sensitivity not estimable	High
Mixed conditions (n = 2)		
Baltaci 2025 [133]	Image-level split; possible leakage	High
Fujiwara 2026 [134]	Incorporation of radiology reports into the reference standard	High

Overall risk of bias was judged across the QUADAS-2 domains, supplemented by QUADAS-AI items (data partitioning, leakage and external testing): Low if all domains were at low risk; High if at least one domain was at high risk; Unclear otherwise. Applicability concerns are reported but do not contribute to the overall judgement. Not formally appraised: the endpoint was localisation only, or no diagnostic accuracy was reported.

Totals: Low, n = 3; Unclear, n = 1; High, n = 33; Not formally appraised, n = 1.

Abbreviations: AD, aortic dissection; AI, artificial intelligence; CTA, computed tomography angiography; EVAR, endovascular aneurysm repair; LOOCV, leave-one-out cross-validation; QUADAS, Quality Assessment of Diagnostic Accuracy Studies; TBAD, type B aortic dissection.

Table S9. Subgroup summaries of the headline performance statistics by task.

Values are unweighted medians [range] across the studies reporting the statistic, followed by the number of studies; subgroups are defined as in Section 2.7 of the main text. Study sets are those of the summary lines under Tables 1–3. External test as defined in Section 2.5 and Tables 1–3 (for measurement: external or partially external test stated by the authors, including vendor-independent evaluation of commercial software; "not stated" grouped with "no").

Statistic	All studies	Sample size, smaller	Sample size, larger	External test reported	No external test (or not stated)	Additional subgroup A	Additional subgroup B
Segmentation – DSC by compartment (sample size ≤200 vs >200 patients or scans; additional subgroups not applicable)							
Outer wall	0.96 [0.89–0.99], 11	0.95 [0.89–0.97], 7	0.96 [0.94–0.99], 4	0.97 [0.97–0.97], 2	0.95 [0.89–0.99], 9	—	—
Lumen (non-dissected)	0.92 [0.86–0.97], 8	0.91 [0.86–0.93], 7	0.97, 1	0.86, 1	0.92 [0.89–0.97], 7	—	—
True lumen	0.92 [0.87–0.96], 6	0.92 [0.87–0.96], 4	0.93 [0.92–0.93], 2	0.92, 1	0.93 [0.87–0.96], 5	—	—
False lumen	0.90 [0.88–0.95], 5	0.92 [0.89–0.95], 4	0.88, 1	0.88, 1	0.92 [0.89–0.95], 4	—	—
Intraluminal thrombus	0.86 [0.80–0.94], 8	0.84 [0.80–0.89], 7	0.94, 1	—, 0	0.86 [0.80–0.94], 8	—	—
False-lumen thrombus	0.63 [0.52–0.90], 4	0.63 [0.52–0.90], 3	0.63, 1	0.63, 1	0.63 [0.52–0.90], 3	—	—
Measurement – maximum diameter (sample size ≤200 vs >200 patients; A = model developed in the study; B = commercial or open-source software applied)							
Absolute bias (mm)	0.26 [0–1.3], 8	0.37 [0.04–1.3], 3	0.23 [0–0.92], 5	0.19 [0–0.37], 4	0.58 [0.04–1.3], 4	0.30 [0.04–1.3], 6	0.14 [0–0.28], 2
Half-width of 95% LoA (mm)	5.2 [2.4–8.5], 7	4.3 [2.4–8.5], 3	5.7 [4.7–7.8], 4	4.7 [2.4–5.2], 3	7.0 [4.3–8.5], 4	5.7 [2.4–8.5], 6	4.7, 1
Mean absolute error (mm)	1.5 [0.6–2.5], 8	1.6 [0.6–2.5], 4	1.5 [1.2–2.1], 4	1.3 [0.6–1.6], 3	2.1 [0.8–2.5], 5	1.4 [0.8–2.1], 4	1.9 [0.6–2.5], 4
Detection – aortic dissection or acute aortic syndrome, patient or examination level (sample size ≤500 vs >500 patients or scans; A = non-contrast CT; B = contrast-enhanced CT)							
Internal AUC	0.96 [0.80–0.995], 14	0.94 [0.80–0.995], 9	0.96 [0.94–0.98], 5	0.95 [0.80–0.99], 11	0.96 [0.94–0.995], 3	0.94 [0.80–0.98], 8	0.98 [0.89–0.995], 6
External AUC	0.97 [0.85–0.995], 9	0.97 [0.85–0.995], 6	0.96 [0.94–0.97], 3	0.97 [0.85–0.995], 9	n/a	0.94 [0.85–0.97], 5	0.98 [0.96–0.995], 4
External sensitivity	0.93 [0.88–1.00], 12	0.95 [0.89–1.00], 8	0.91 [0.88–0.98], 4	0.93 [0.88–1.00], 12	n/a	0.90 [0.88–0.98], 6	0.96 [0.88–1.00], 6
External specificity	0.92 [0.55–1.00], 12	0.88 [0.55–1.00], 8	0.94 [0.91–0.96], 4	0.92 [0.55–1.00], 12	n/a	0.85 [0.55–0.94], 6	0.94 [0.87–1.00], 6

Abbreviations: AUC, area under the ROC curve; DSC, Dice similarity coefficient; LoA, limits of agreement; n/a, not applicable.

Table S10. Decision rules for the CLAIM-informed limitation tiers (segmentation and measurement studies).

Area (CLAIM 2024 items)	No concern (–)	Concern (+)	Critical concern (++)
Reference standard and data (CLAIM items on data source, reference standard, annotators)	Human reference established by ≥2 readers (including sequential verification by a second reader) or inter-reader variability documented; data source described	Single reader or operator without documented variability; reader qualifications not reported; reference partly software-assisted but human-verified; clinical report as reference; weak or indirect labels	Reference generated or corrected by the model under evaluation or by the authors' own software; no reference segmentation or measurement, with agreement assessed by reader scoring or visual review only
Validation and generalisability (items on data partitioning, internal and external testing)	External or independent test reported (other institution, public dataset, or independent cohort), partitioned at patient level	Internal hold-out or cross-validation only; small test set; vendor training data not reported; single data source	Test data not independent of training (image- or scan-level partitioning of multi-image data; thresholds or parameters selected on the

Area (CLAIM 2024 items)	No concern (–)	Concern (+)	Critical concern (++)
			test data); evaluation restricted to selected slices or cases; difficult cases excluded
Reporting and reproducibility (items on performance metrics and their reporting)	Agreement or overlap statistic reported with its uncertainty or distribution	Incompletely reported methods (e.g., number of readers or operators, cohort size inconsistently stated, agreement statistics incomplete, second reader on a subsample only)	Performance shown only graphically or as group means; no agreement statistic
Tier rule			
Higher: any area with a critical concern (++). Lower: no critical concern, validation area with no concern (–) and reference-standard area with no concern (–). Moderate: all other combinations. Restrictions of the study population (e.g., healthy subjects only, a single operative setting, purposive sampling) are recorded as applicability notes in Tables S6 and S7 and do not contribute to the tier, in analogy with the separate handling of applicability in QUADAS-2.			