GMR
The Original

OPTIMIZATION OF CARDIOVASCULAR DIAGNOSIS THROUGH ARTIFICIAL INTELLIGENCE APPLIED TO THORACIC RADIOLOGICAL IMAGES: A SYSTEMATIC REVIEW OF DIAGNOSTIC PERFORMANCE, GENERALISABILITY AND TRANSLATIONAL READINESS

Denisse Andrea Puentes Aguila[1], Ana Maria Ramírez Isaza[2], Roxana Rivera Valencia[3,4], Valeria Guadalupe de la Cajiga Torres[5]

[1]Universidad Viña del Mar (UVM), Chile, Email: denisse.puentes@uvm.cl, ORCID: https://orcid.org/0000-0002-4559-9702
[2]Universidad CES, Medellín, Colombia, Email: anramirezi@uces.edu.co, ORCID: https://orcid.org/0009-0005-4164-5301
[3]Universidad de Carabobo, Valencia, Venezuela
[4]RedSalud Alameda, RedSalud CChC, Santiago, Chile, Email: roxana.rivera.ext@redsalud.cl
[5]Universidad Autónoma del Estado de Hidalgo, Hidalgo, México, Email: valeriacajiga@gmail.com, ORCID: https://orcid.org/0009-0005-0223-0955

ABSTRACT

Background. Cardiovascular disease remains the leading cause of death worldwide, yet risk stratification depends on inputs that are frequently missing in routine care. Chest radiographs and chest computed tomography (CT) are acquired in vast numbers for non-cardiac indications and contain cardiovascular information that is not systematically extracted. Artificial intelligence (AI) has been proposed as a means of recovering that information.

Objective. To systematically appraise the diagnostic and prognostic performance, the generalisability, and the translational readiness of AI models that infer cardiovascular status from thoracic radiological images.

Methods. Following PRISMA 2020, eight pre-specified queries were executed on 4 September 2026 across three federated evidence-discovery platforms (Consensus, Elicit, scite), which jointly index MEDLINE/PubMed, Scopus, Semantic Scholar and arXiv. Eligible records were primary studies (2019–2026) developing or validating an AI model whose index test was a chest radiograph or chest CT and whose target condition was cardiovascular. Risk of bias was appraised with PROBAST+AI and QUADAS-2 signalling domains. Because of substantial clinical and methodological heterogeneity, synthesis was structured and narrative; a pre-specified exploratory analysis compared paired internal and external discrimination. Studies were mapped onto a two-axis Translational Readiness Matrix combining an adapted imaging-efficacy hierarchy with a five-level generalisability ladder.

Results. Of 130 records retrieved and 116 screened, 44 studies were included (chest radiography n = 28; chest CT n = 16). Discrimination clustered by task. Automated coronary artery calcium (CAC) quantification on non-gated chest CT was the most mature application, with intraclass correlations against gated reference standards of 0.90–0.99 and risk-category agreement (κ) of 0.65–0.95. Radiograph-based inference was consistently more modest: area under the receiver operating characteristic curve (AUC) 0.73–0.90 for any CAC, 0.79–0.92 for reduced left ventricular ejection fraction and left ventricular structural abnormality, and 0.83–0.92 for moderate-to-severe valvular lesions. In six studies reporting paired estimates, discrimination fell on external validation by a median of 0.045 AUC points (range −0.130 to +0.033). Only 11 of 44 studies (25%) reported performance disaggregated by sex, age or ethnicity. Thirty-seven studies (84%) reached only technical or diagnostic-accuracy efficacy; a single study demonstrated therapeutic impact, and none demonstrated patient-outcome or economic benefit.

Conclusions. AI can convert routine thoracic imaging into an opportunistic cardiovascular risk signal, and for CAC quantification on chest CT the technical case is essentially settled. The field's binding constraint is no longer accuracy but evidence architecture: external, demographically disaggregated and prospectively collected validation, and outcome-anchored endpoints. We propose the Translational Readiness Matrix as a reporting and commissioning instrument to make that gap explicit.

KEYWORDS: artificial intelligence; deep learning; chest radiography; computed tomography; cardiovascular diseases; coronary artery calcium; opportunistic screening; external validation; algorithmic fairness; systematic review.

Registration: This review was not prospectively registered; the protocol is available from the corresponding author. Registration on PROSPERO is recommended before any update.

1. INTRODUCTION

Cardiovascular disease is the single largest contributor to global mortality, and the greater part of that burden is attributable to conditions that are, in principle, detectable before they become symptomatic. Contemporary primary-prevention guidelines therefore rest on quantitative risk estimation. In practice this architecture leaks at the first step: the pooled cohort equations and their successors require lipid values, blood pressure and smoking history that are simply absent from the record for a large share of the population, and clinicians cannot compute a risk score they have no inputs for [5]. The consequence is a systematic under-identification of people who would benefit from lipid-lowering therapy, concentrated precisely among those with the least continuous access to preventive care.

Thoracic imaging occupies an unusual position in this landscape. Chest radiography is the most frequently performed radiological examination in the world; roughly 110 million chest radiographs are acquired annually in the United States alone, interpreted by a radiology workforce that has not expanded proportionally [3]. Approximately 19 million non-electrocardiogram-gated chest CT examinations are performed each year in the same setting, overwhelmingly for pulmonary, oncological or trauma indications [2]. Every one of these studies contains the cardiac silhouette, the thoracic aorta, the pulmonary vasculature and, in the case of CT, directly visualised coronary calcium. The cardiovascular content of these images is acquired, stored, and then largely discarded — not because it is absent, but because extracting it reliably by eye is slow, subjective and, for some features, beyond the resolving power of human perception [1].

The cardiothoracic ratio illustrates the ceiling of conventional reading. It is the only quantitative cardiovascular marker in routine radiographic practice, it does not capture great-vessel change, and its normal range is poorly specified by age and sex [18]. Radiologists' assessments of cardiac size and pulmonary venous congestion show meaningful intra- and inter-observer variability even among subspecialists [19]. Against that baseline, the relevant question is not whether an algorithm can replace a radiologist, but whether it can render legible a signal that no reader was ever expected to extract.

Since 2019 a substantial literature has answered that question in the affirmative across a widening range of targets: coronary artery calcium and ten-year event risk from a single frontal radiograph [4–8]; significant coronary stenosis and long-term mortality in patients referred for angina [9,10]; biological age as a cardiovascular prognostic marker [11–14]; left ventricular ejection fraction, chamber geometry and valvular lesions previously considered the exclusive province of echocardiography [15–18]; pulmonary venous and arterial hypertension [19,20,26–28]; and fully automated Agatston-equivalent scoring on chest CT acquired for entirely unrelated reasons [29–44]. Several of these models are open-source, and at least one has been deployed across a national health system [35].

This accumulation has not been matched by a comparable accumulation of evidence that patients are better off. Systematic evaluations of radiological AI as a whole find that the literature is dominated by retrospective, single-centre, internally validated work, with an average performance decrement of roughly six percentage points when models are tested outside their development environment [60]. Of 903 AI-enabled devices cleared by the United States Food and Drug Administration by August 2024, only 8.1% of the supporting clinical studies were prospective and 2.4% randomised, and fewer than a third reported sex-specific results [62]. Across 173 CE-marked radiological AI products, the proportion of evidence addressing clinical decision-making, patient outcome or economic impact has remained static at roughly a quarter, while vendor independence and prospective design have if anything declined [63]. Meanwhile, chest radiograph classifiers have been repeatedly shown to underdiagnose historically under-served populations, with the largest gaps at demographic intersections [54,57].

Existing reviews of this domain have been valuable but narrative, and have generally reported performance without interrogating the conditions under which it was obtained [1,66]. What is missing is a synthesis that treats generalisability and evidence level as primary outcomes rather than as caveats. This review therefore pursues three objectives. First, to characterise systematically the diagnostic and prognostic performance of AI models that infer cardiovascular status from chest radiographs and chest CT. Second, to quantify, where paired estimates permit, the discrimination penalty incurred on external validation. Third, to propose and apply an explicit two-axis instrument — the Translational Readiness Matrix — that positions each study simultaneously on an efficacy hierarchy and a generalisability ladder, making visible the distance between what has been demonstrated and what would be required for clinical commissioning.

2. METHODS

2.1 Protocol and reporting

The review was conducted and is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [48]. The protocol was specified in advance of screening but was not registered on PROSPERO; this is acknowledged as a limitation (Section 5). No ethical approval was required, as the review used only published aggregate data.

2.2 Eligibility criteria

Eligibility was framed using a PICOTS structure (Table 1). In brief, studies were eligible if they were primary research reports that developed, validated or clinically evaluated an artificial-intelligence model whose index test was a thoracic radiological image — a chest radiograph in any projection, or a chest computed tomography

acquisition, whether gated or non-gated — and whose target condition was cardiovascular. Cardiovascular targets included subclinical atherosclerosis and coronary artery calcium, obstructive coronary disease, cardiac chamber size and systolic function, valvular heart disease, pulmonary arterial and venous hypertension, congenital heart disease, and cardiovascular events or mortality. Studies were required to report at least one extractable quantitative performance estimate. Narrative reviews, editorials, correspondence and protocols without results were excluded, as were studies in which the model input was not a thoracic radiological image (for example echocardiography, electrocardiography alone, digital auscultation, tabular clinical data, or non-thoracic CT) and studies whose target condition was non-cardiovascular. Conference abstracts were retained when they reported sufficient methodological and quantitative detail, and are flagged as such throughout. Publication window was 1 January 2019 to 4 September 2026; language was restricted to English.

2.3 Information sources and search strategy

Searches were executed on 4 September 2026 across three federated evidence-discovery platforms — Consensus, Elicit and scite — which jointly index MEDLINE/PubMed, Scopus, Semantic Scholar and arXiv and together cover in excess of 200 million records. Eight pre-specified queries were used, structured around four concept blocks: the technology (artificial intelligence, deep learning, convolutional neural network, foundation model), the modality (chest radiograph, chest X-ray, thoracic computed tomography), the target (cardiovascular disease, coronary artery calcium, left ventricular dysfunction, valvular heart disease, pulmonary hypertension, cardiomegaly), and the evaluative dimension (external validation, generalisability, algorithmic bias, reporting standards, prospective evaluation). The full query set is reproduced in Table 2. Reference lists of included studies and of relevant reviews were screened for additional records.

2.4 Study selection and data extraction

Records were de-duplicated on digital object identifier and title. Titles and abstracts were screened against the eligibility criteria, and records surviving that stage were assessed in full text. A standardised extraction form captured: bibliographic identifiers; country and number of contributing sites; study design and data provenance; sample size at the image and patient level; index modality and projection; model architecture; reference standard and its ascertainment; the target condition and any thresholds applied; internal and external discrimination with confidence intervals; calibration; sensitivity and specificity at the reported operating point; whether performance was disaggregated by demographic subgroup; the explainability method employed, if any; and whether any comparison against human readers or any prospective evaluation was undertaken.

2.5 Risk of bias and quality appraisal

Risk of bias was appraised using the four signalling domains shared by PROBAST+AI and QUADAS-2 — participants and data sources, predictors or index test, outcome or reference standard, and analysis — with each domain rated as low, high or unclear risk [50,52]. PROBAST+AI was preferred because it distinguishes explicitly between model development and model evaluation, a distinction that is central to this literature [50]. Adherence to reporting standards was appraised against TRIPOD+AI [49] and, for studies claiming early clinical evaluation, DECIDE-AI [51]. Because the majority of included studies are prediction-model development studies rather than classical diagnostic-accuracy studies, QUADAS-2 was applied as a supplementary rather than a primary instrument.

2.6 Data synthesis

Clinical heterogeneity across target conditions, reference standards, thresholds and case-mix was substantial, and the reference standards themselves are not interchangeable — an Agatston score derived from gated CT, an echocardiographic ejection fraction and an angiographic stenosis threshold measure different constructs. Formal meta-analytic pooling of discrimination across these targets would therefore have produced a summary statistic without a coherent clinical referent, and was not attempted. Synthesis proceeded instead by structured narrative grouping into four clinical domains, with performance tabulated within each.

One pre-specified quantitative analysis was undertaken. For every study reporting discrimination in both an internal validation set and at least one geographically or temporally independent external set for the same task, the paired estimates were extracted and the difference (ΔAUC = external − internal) computed. Where a study reported several thresholds for a single construct, the arithmetic mean across thresholds was used. Results are summarised as median and range and are reported as a descriptive, exploratory analysis rather than a meta-analysis; no pooling, weighting or inferential testing was performed, and the small number of contributing studies precludes any assessment of small-study effects.

2.7 The Translational Readiness Matrix

To characterise evidence maturity, each included study was positioned on two orthogonal axes. The efficacy axis adapts the established hierarchical model of imaging efficacy, which has previously been applied to commercial radiological AI [63], collapsing technical and diagnostic-accuracy efficacy into a single level (E1–E2) and retaining diagnostic-thinking efficacy (E3), therapeutic efficacy (E4), patient-outcome efficacy (E5) and societal or economic efficacy (E6). The generalisability axis was developed for this review and grades the evidentiary strength of the validation design: G0, internal split or cross-validation only; G1, temporal validation or a single external site within the same health system; G2, multi-site external validation; G3, external validation in a different country or health system; G4, prospective or interventional evaluation in the intended workflow. Each study was assigned the highest level attained on each axis, and the resulting joint distribution is presented as an evidence

map. The matrix is descriptive rather than a quality score: a study may be methodologically excellent and still sit at E1–E2/G0 by design.

Table 1. Eligibility criteria specified using a PICOTS framework.

Element	Included	Excluded
Population	Adults or children undergoing chest radiography or chest CT for any clinical or screening indication	Phantom-only or purely synthetic-image studies
Index test	Any AI or machine-learning model taking a thoracic radiological image as primary input, alone or fused with routine clinical covariates	Echocardiography, ECG-only, digital auscultation, tabular-only models, non-thoracic imaging
Comparator	Reference standard, clinical risk score, or human reader; comparator not mandatory	—
Outcome	Discrimination (AUC), sensitivity, specificity, agreement (ICC, κ), or event-based hazard ratios	Studies without any extractable quantitative performance estimate
Target condition	Coronary artery calcium; obstructive CAD; LV size, mass and systolic function; valvular disease; pulmonary arterial or venous hypertension; congenital heart disease; cardiovascular events or mortality	Purely pulmonary, oncological, infectious or musculoskeletal targets
Timing / design	Retrospective or prospective primary studies; development, validation or clinical-evaluation designs; 2019–2026	Narrative reviews, editorials, correspondence, protocols without results
Setting	Any (primary care, outpatient, inpatient, emergency, screening cohort)	—

Table 2. Search strategy: platform, query and yield. Searches executed 4 September 2026.

#	Platform	Query	Filter	Records
Q1	Consensus	deep learning chest radiograph cardiovascular disease risk prediction	≥2018	20
Q2	Elicit	artificial intelligence chest X-ray detection of cardiomegaly, valvular heart disease, left ventricular systolic dysfunction diagnostic accuracy	≥2019	10
Q3	scite	"external validation" AND "deep learning" AND ("chest radiograph" OR "chest X-ray") AND generalizability	≥2019	10
Q4	Consensus	artificial intelligence chest radiograph detection valvular heart disease pulmonary hypertension left ventricular dysfunction	≥2020	20
Q5	Consensus	algorithmic bias underdiagnosis chest X-ray artificial intelligence health equity underserved populations	≥2020	20
Q6	Elicit	reporting guidelines and risk of bias assessment for medical imaging AI: CLAIM, TRIPOD+AI, PROBAST-AI, QUADAS-2, DECIDE-AI	≥2018	10
Q7	Consensus	opportunistic screening coronary artery calcium routine chest CT deep learning clinical implementation cost-effectiveness	≥2020	20
Q8	Consensus	prospective clinical trial randomized artificial intelligence radiology workflow implementation regulatory approval imaging devices	≥2021	20
		Total retrieved		130

3. RESULTS

3.1 Study selection

The eight queries returned 130 records. Fourteen duplicates were removed, leaving 116 unique records for title and abstract screening. Forty-four were excluded at that stage, and 72 records were assessed in full text. A further 28 were excluded: 13 were narrative reviews, editorials or correspondence; six addressed a non-cardiovascular target; seven used an input that was not a thoracic radiological image; and two reported no extractable quantitative performance estimate. Forty-four studies met all criteria and were included (Figure 1). Twenty-eight used chest radiography as the index modality and 16 used chest CT.

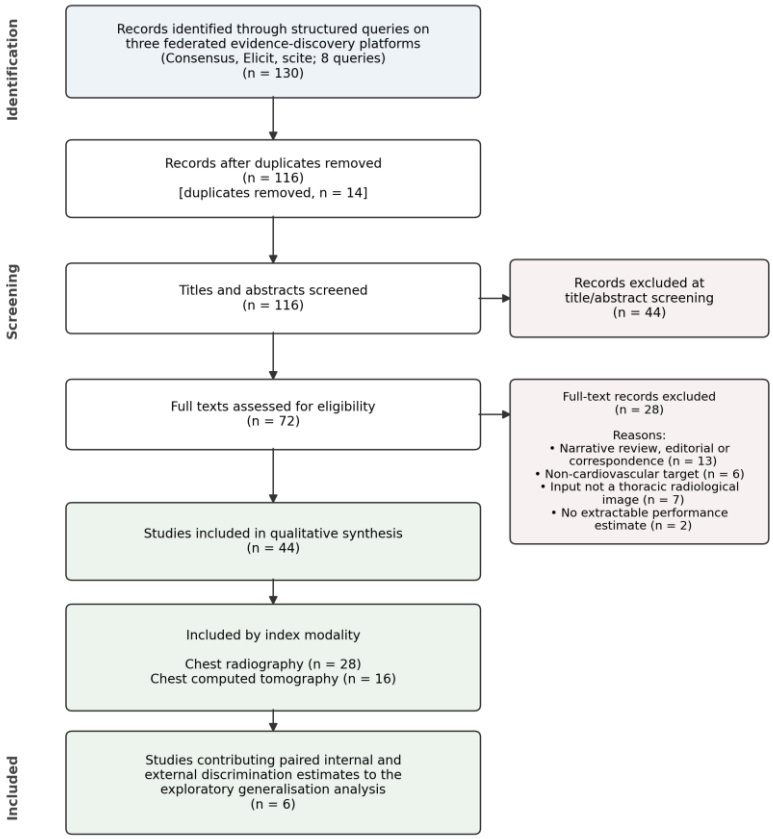

Figure 1. PRISMA 2020 flow diagram of study identification, screening and inclusion.

3.2 Characteristics of included studies

Included studies were published between 2020 and 2026, with two-thirds appearing from 2023 onwards. Data provenance was overwhelmingly retrospective: 42 of 44 studies used retrospectively assembled institutional or trial cohorts, most commonly the National Lung Screening Trial, the Prostate, Lung, Colorectal and Ovarian screening trial, MIMIC-CXR, CheXpert, or single-centre picture-archiving extracts. Sample sizes spanned four orders of magnitude, from 55 paired examinations in a prospective calcium-scoring substudy [31] to 116,035 radiographs used for model development [11] and 795,055 radiographs in a multi-country classification study [71]. Convolutional architectures predominated — DenseNet, ResNet and Inception variants — with U-Net-derived segmentation used where an anatomical measurement rather than a label was the objective [21,23,25], and transformer or vision-language architectures appearing only in the most recent work [19,70,72]. Table 3 summarises the characteristics of representative included studies grouped by clinical domain.

Table 3. Characteristics and principal performance findings of representative included studies, grouped by clinical domain. AUC, area under the receiver operating characteristic curve; CAC, coronary artery calcium; CACS, coronary artery calcium score; CXR, chest radiograph; DL, deep learning; ICC, intraclass correlation coefficient; LV, left ventricle; LVEF, left ventricular ejection fraction; MACE, major adverse cardiovascular events; NLST, National Lung Screening Trial; PLCO, Prostate, Lung, Colorectal and Ovarian screening trial; VSD, ventricular septal defect.

Study (year)	Modality	n	Target	Reference standard	Principal finding
Kamel (2021) [4]	CXR	1,689	CAC > 0; CAC ≥ 100	Gated cardiac CT Agatston	AUC 0.73 frontal, 0.70 lateral for CAC > 0;

Study (year)	Modality	n	Target	Reference standard	Principal finding
					0.74 for CAC ≥ 100; predicted CAC independent of traditional risk factors
Weiss (2024) [5]	CXR	11,001 external	10-year MACE	Incident MACE (EHR follow-up)	Adjusted HR 1.73 (1.47–2.03) in patients with incalculable ASCVD score; incremental to the ASCVD score (HR 1.88)
Hong (2025) [6]	CXR	10,230	CACS 0 / 100 / 400 thresholds	CACS within 6 months	AUC 0.74–0.79 image-only; 0.77–0.82 with age, sex, BMI; external 0.78–0.81
Gallone (2025) [7]	CXR	550	CAC > 0	Chest CT within 3 months	AUC 0.90 internal, 0.77 temporally independent external; sensitivity > 92%; ASCVD events 13.5% vs 3.4% by AI-CAC status
Jeong (2025) [8]	CXR + demographics	2,121 + 2 external	CAC 0 vs > 400	Gated CT Agatston	AUC 0.86 internal; 0.77 and 0.82 in two external cohorts with differing ethnic composition
D'Ancona (2022) [9]	CXR	7,728	Stenosis ≥ 70% (≥ 50% left main)	Invasive coronary angiography	Sensitivity 0.90 at specificity 0.31; AUC 0.73, rising to 0.77 with angina status
Raghu (2021) [11]	CXR	46,381 test	All-cause and CV mortality	Observed mortality (PLCO, NLST)	CXR-Age HR 2.45 per 5 years for CV mortality vs 1.82 for chronological age
Ieki (2022) [12]	CXR	>100,000	Radiograph-estimated age	Chronological age; outcomes	Elevated X-ray age associated with heart-failure readmission and death
Ueda (2023) [15]	CXR	22,551	LVEF, 8 valvular targets	Paired echocardiography	External AUC 0.92 for LVEF < 40%; 0.83–0.92 across valvular targets; 4-site design with independent external institution
Bhave (2024) [16]	CXR	71,589 + 8,003 external	Severe LV hypertrophy; dilated LV	Echocardiographic labels	AUC 0.79–0.80; outperformed all 15 board-certified radiologists on the composite label
Hsiang (2022) [17]	CXR	Multi-centre	LV systolic dysfunction	Echocardiography	AI-enabled CXR identified reduced systolic function with clinically usable discrimination

Study (year)	Modality	n	Target	Reference standard	Principal finding
Sogancioglu (2020) [21]	CXR	65,000 + 367 test	Cardiomegaly	Two independent radiologists	Segmentation AUC 0.977 vs classification 0.941; segmentation matched an independent expert reader using 100× fewer annotations
Lee JG (2024) [18]	CXR	96,129 normal + 44,567 diseased	Cardiovascular border z-scores	Disease labels; 5-year outcomes	RA + LV combination AUC 0.82 vs cardiothoracic ratio 0.79 for valve disease; z ≥ 2 associated with adjusted HR 3.73
White (2025) [19]	CXR	1,682	Pulmonary venous hypertension stage	Doppler echocardiographic LVDD grade	Transformer-based ranking tracked LVDD grade more strongly than expert consensus staging ($p < 0.001$)
Liu PY (2024) [20]	CXR + ECG	Multi-centre	Systolic PAP > 50 mmHg	Transthoracic echocardiography	AUC 0.85 CXR alone, 0.86 combined internal, 0.87 external; NPV 98%
Han (2023) [26]	CXR	3,255	Congenital heart disease; PAH-CHD	Clinical/surgical diagnosis	AUC 0.948 for CHD, 0.778 for PAH-CHD; AI assistance improved five radiologists' AUC (Δ +0.096 and +0.066)
Zeleznik (2021) [29]	Chest CT	20,084	CAC; cardiovascular events	Manual Agatston; adjudicated events	Automated score predicted events independent of risk factors (adjusted HR up to 4.3) across four cohorts
Chao (2021) [30]	Low-dose CT	30,286 train / 2,085 test	CVD and CVD mortality	Registry outcomes; gated CT markers	AUC 0.871 for CVD, 0.768 for CVD mortality
Eng (2021) [31]	Chest CT	341 train / 303 external	CAC ≥ 1; CAC ≥ 100	Paired gated CT	Sensitivity 71–94% and PPV 88–100% for CAC ≥ 100 across four geographically distinct health systems
van Velzen (2020) [32]	Chest CT ×6 protocols	7,240	CAC and thoracic aortic calcification	Manual Agatston / volume	ICC 0.79–0.97 at baseline, rising to 0.85–0.99 with protocol-specific training; κ 0.90 for risk category
Hagopian (2025) [35]	Non-gated CT	446 train; 795 paired; 8,052 screening	CAC 0 / ≥ 100 / > 400	Paired gated CT within 1 year	Accuracy 89.4% and 87.3%; 10-year mortality 25.4% vs 60.2% for CAC 0 vs > 400; developed across 98 medical centres
Peng (2023) [36]	Non-gated CT	5,678	All-cause death; death/MI/stroke	Adjudicated outcomes	DL-CAC ≥ 100 associated with

Study (year)	Modality	n	Target	Reference standard	Principal finding
					adjusted HR 1.51 for death; only 26% of this group were receiving statins
Sandhu (2022) [37]	Non-gated CT	173 randomised	Statin initiation at 6 months	Prescription record	51.2% vs 6.9% statin initiation after clinician and patient notification ($p < 0.001$); randomised quality-improvement design
Raikhelkar (2026) [44]	Non-gated CT	34,058	LVEF < 50%	Paired echocardiography reports	AUROC 0.786 hold-out, 0.762 external; exceeded expert radiologists in accuracy and reading time

3.3 Domain A: subclinical atherosclerosis and cardiovascular risk from chest radiography

Ten studies addressed the inference of atherosclerotic burden or event risk from a plain radiograph. Performance was consistent and consistently moderate. Discrimination for the presence of any coronary calcium clustered between 0.73 and 0.90, with the higher figures obtained in internal validation of smaller, tightly curated cohorts [7] and the lower figures in larger or externally validated series [4,8]. Discrimination improved reliably at higher calcium thresholds — one large study reported AUC rising from 0.74 for zero versus non-zero to 0.79 for the 400 Agatston threshold, and to 0.82 when age, sex and body-mass index were fused with the image [6] — which is consistent with the physical intuition that a heavier calcium burden produces a more radiographically apparent cardiomediastinal signature.

The more clinically interesting finding is that these models carry prognostic information that is not reducible to the calcium score they were trained to approximate. A deep-learning risk estimate derived from a single radiograph predicted ten-year major adverse cardiovascular events with an adjusted hazard ratio of 1.73 among 8,869 outpatients whose conventional risk score could not be computed because the necessary inputs were missing, and remained predictive over and above the conventional score in patients for whom it could be computed [5]. Radiograph-derived biological age behaved similarly: a five-year increment in estimated age carried a greater hazard for cardiovascular mortality than a five-year increment in chronological age (2.45 versus 1.82) across two large screening trials [11], and the same construct predicted heart-failure readmission and death in hospitalised cohorts [12,13] and improved established intensive-care mortality scores when substituted for chronological age [14]. Radiograph-derived and polygenic risk estimates were independently and additively associated with incident coronary disease [45], and radiograph risk scores correlated with directly measured calcium, systolic pressure, ankle–brachial index and angiographic stenosis in two independent external cohorts [46].

One study departed instructively from the deep-learning paradigm by extracting 32 interpretable geometric measurements — cardiomediastinal dimensions, aortic calcification and tortuosity — from segmented radiographs. A model built from these measurements plus age and sex performed comparably to a deep-learning risk model over a 19-year median follow-up, and identified two previously unreported prognostic features: mediastinal width at valve level and maximal lateral displacement of the descending aorta [47]. This suggests that a substantial portion of the radiographic cardiovascular signal is geometrically explicit rather than irreducibly latent, which has direct consequences for the interpretability problem discussed below.

3.4 Domain B: chamber morphology, systolic function and valvular disease

Eight studies addressed cardiac structure and function. The most methodologically complete was a four-institution study of 22,551 radiographs paired with contemporaneous echocardiograms, externally tested at a fifth site, which reported AUCs of 0.92 for an ejection fraction below 40%, 0.89 for moderate-to-severe mitral regurgitation, 0.92 for tricuspid regurgitation, 0.83 for aortic stenosis and 0.85 for elevated tricuspid regurgitant velocity [15]. Specificity at the reported operating points was generally 70–86% and sensitivity 73–91%, a profile appropriate to a triage instrument that defers a definitive answer to echocardiography rather than to a confirmatory test. Commentators have noted that the practical value of such a model depends critically on the pre-test prevalence of the population in which it is deployed, and that a screening application would require careful attention to positive predictive value in low-prevalence primary-care settings [68].

A second large study developed a model for severe left ventricular hypertrophy and dilated left ventricle from 71,589 radiographs with echocardiographic labels, achieving AUCs of 0.79–0.80 with comparable performance on 8,003 externally sourced radiographs. Critically, it also compared the model against 15 board-certified radiologists reading the same images, and the model exceeded every individual reader on the composite label,

achieving 71% sensitivity against the consensus reader vote's 66% at matched specificity [16]. This is one of the few results in the corpus that establishes not merely that the model works but that it recovers information human readers do not reliably extract.

Cardiomegaly, the most-studied single target, illustrates a methodological point rather than a clinical one. Reported accuracies ranged from 84% to 95% [22,24,25], but the comparison that matters was made in a study that pitted anatomical segmentation against image-level classification on the same data: segmentation reached an AUC of 0.977 against 0.941 for classification, matched an independent expert reader, and required roughly one hundred times fewer annotated radiographs while producing an inspectable output [21]. A subsequent study measuring cardiothoracic ratio and transverse cardiac diameter from U-Net segmentations against cardiac magnetic resonance found accuracies of only 62–74% depending on the magnetic-resonance definition of cardiomegaly adopted [23] — a reminder that apparent performance in this task is substantially a function of which reference standard is chosen. A related approach abandoned the cardiothoracic ratio altogether, deriving age- and sex-specific z-scores for eight cardiovascular border segments from 96,129 normal radiographs; the combination of right atrial and left ventricular borders outperformed the cardiothoracic ratio for detecting valvular disease, and border z-scores differentiated mitral from aortic stenosis in a pattern consistent with their distinct pathophysiology [18].

3.5 Domain C: pulmonary vascular and congenital disease

Six studies addressed pulmonary hypertension or congenital disease. A transformer-based model assigning pulmonary vascular patterns tracked echocardiographic diastolic-dysfunction grade more strongly than expert cardiothoracic radiologist consensus, in a design that explicitly quantified intra- and inter-reader variability first and thereby established the human baseline the model was to be judged against [19]. Fusing radiograph and electrocardiogram improved detection of systolic pulmonary artery pressure above 50 mmHg from an AUC of 0.85 for the radiograph alone to 0.86 internally and 0.87 externally, with a negative predictive value of 98% in both cohorts and hazard ratios above four for subsequent left ventricular dysfunction and cardiovascular death [20]. In paediatric populations, models detected congenital heart disease with an AUC of 0.948 but associated pulmonary arterial hypertension with only 0.778; the reader study embedded in that work is notable because AI assistance improved the mean AUC of five radiologists by 0.096 for congenital disease and 0.066 for pulmonary hypertension, both statistically significant [26]. Comparable results were obtained for post-operative pulmonary hypertension in ventricular septal defect [27,28].

3.6 Domain D: automated calcium quantification and risk stratification on chest CT

Sixteen studies addressed chest CT, and this is where the technical case is strongest. Automated Agatston-equivalent quantification on non-gated CT agreed closely with manual gated scoring across every validation reported: intraclass correlations of 0.90–0.99 [32,34,38], correlation coefficients above 0.92 against gated reference standards [33], and risk-category agreement (Cohen's κ) between 0.65 and 0.95 [32,34,38,39]. Robustness across acquisition protocols was directly examined in a study spanning six examination types including PET attenuation-correction and radiotherapy planning CT, where baseline intraclass correlations of 0.79–0.97 rose to 0.85–0.99 when protocol-specific images were added to training [32]. The largest deployment-oriented evaluation developed a model across 98 medical centres of a national health system and reported accuracies of 89.4% and 87.3% for the zero and 100-Agatston thresholds against paired gated CT, with ten-year mortality of 25.4% versus 60.2% between the zero and above-400 groups; four cardiologists reviewing a random sample of high-scoring low-dose screening examinations judged that 527 of 531 patients would benefit from lipid-lowering therapy [35].

Prognostic validity was established independently of agreement with the gated standard. Automated calcium scoring predicted events across asymptomatic and symptomatic cohorts with adjusted hazard ratios up to 4.3 [29]; incidental calcium of 100 or more on routine ungated CT was associated with an adjusted hazard ratio of 1.51 for all-cause death and 1.57 for the composite of death, myocardial infarction and stroke, in a cohort where the mean ten-year risk was 24% but only 26% were receiving statins [36]. A deep-learning model trained directly on low-dose lung-cancer screening CT achieved an AUC of 0.871 for cardiovascular disease and 0.768 for cardiovascular mortality [30], and a model applied to segmented heart and aortic volumes predicted 12-year cardiovascular mortality beyond risk factors, calcium score and radiologist findings [43]. Extension to abnormal ejection fraction on static non-gated CT reached an AUROC of 0.786 internally and 0.762 externally, exceeding expert radiologists in both accuracy and reading time [44].

The single most consequential study in the entire corpus is also the only one to reach therapeutic efficacy. In a randomised quality-improvement project, patients with algorithm-detected and radiologist-confirmed incidental calcium on a prior non-gated CT were randomised to clinician and patient notification or usual care; statin prescription at six months was 51.2% in the notification arm against 6.9% under usual care, with coronary testing also increasing (15.1% versus 2.3%) [37]. This demonstrates that the detection–action gap is closable, and that closing it — not further improving the detector — is where the marginal return currently lies.

3.7 Exploratory analysis of generalisation

Six studies reported paired internal and external discrimination for the same task and contributed to the pre-specified exploratory analysis (Figure 2). The median change on external validation was −0.045 AUC points (range −0.130 to +0.033; interquartile range −0.077 to +0.001). Four of six studies lost discrimination and two gained it. The largest loss (−0.130) occurred in the study with the smallest development cohort, where an internal

AUC of 0.90 for any coronary calcium fell to 0.77 in a temporally independent set from the same institution — a decrement obtained without any change of site, scanner or population [7]. The largest gain (+0.033) occurred in the study with the largest development cohort and the most systematic cross-validation [6]. This inverse relationship between development sample size and generalisation penalty, though observed in only six studies and therefore no more than hypothesis-generating, is consistent with the roughly six-percentage-point average external decrement reported across 535 radiological AI studies [60].

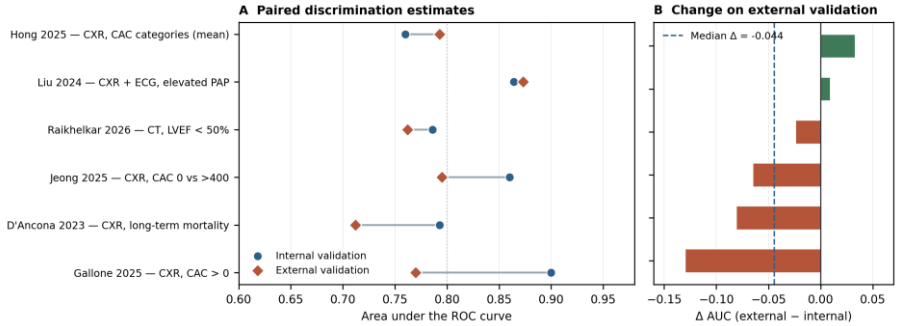

Figure 2. Paired internal and external discrimination in the six included studies reporting both (A), and the corresponding change in area under the ROC curve (B). Negative values indicate loss of discrimination outside the development environment. This is a descriptive summary, not a meta-analysis; estimates are unweighted and no inferential testing was performed.

Two further observations qualify this figure. First, ΔAUC understates the practical problem, because a model may retain rank-ordering while losing calibration entirely — and calibration was reported in only four of 44 included studies, a reporting gap that TRIPOD+AI explicitly targets [49]. Second, external validation on a curated research dataset is not equivalent to external validation in the wild. A multi-architecture study of chest radiograph models found that internal robustness rankings reversed completely on external data: the domain-specific vision-language model with the smallest internal sensitivity gap between anteroposterior and posteroanterior projections had the largest gap externally, rising from 14.3% to 48.9%, while view type explained 61–100% of performance variance across models — far more than age (0–38%) or sex (< 4%) [72]. Technical acquisition parameters, in other words, may dominate the demographic variables that fairness auditing conventionally targets.

3.8 Risk of bias and reporting quality

Risk of bias was high or unclear in at least one domain for 39 of 44 studies (Table 4). The analysis domain was the most frequently compromised: 26 studies were rated high risk, most often because performance was reported only on an internally split sample, because calibration was not assessed, or because the effective number of outcome events was small relative to model complexity. The participants and data sources domain was rated high or unclear in 21 studies, chiefly through retrospective convenience sampling from a single institution's picture-archiving system with incompletely described inclusion. The reference-standard domain fared better, since most studies anchored on an objectively measured comparator — gated CT Agatston score, echocardiographic ejection fraction, or angiographic stenosis — although the interval between index test and reference standard was frequently permissive, extending to six months [6] or a full year [16,35], during which the underlying condition may genuinely have changed.

Reporting quality against TRIPOD+AI was uneven. Sample-size justification was absent in 38 studies; calibration was reported in four; a data-availability statement permitting independent replication accompanied nine; and the flow of participants was fully traceable in 19. Explainability was addressed in 24 studies, almost exclusively through saliency-based post hoc methods — gradient-weighted class activation mapping predominantly, with Shapley values used in a minority — a distribution that mirrors the wider cardiovascular imaging literature, in which such methods are evaluated qualitatively rather than quantitatively and no standardised assessment exists [66].

Demographic disaggregation was the most conspicuous omission. Only 11 of 44 studies (25%) reported any performance stratified by sex, age or ethnicity, and only three examined intersectional subgroups. This matters because the fairness properties of chest radiograph models are now well characterised and unfavourable: classifiers trained on the major public datasets selectively underdiagnose under-served populations, with the largest true-positive-rate gaps at demographic intersections such as Black female patients [54,56], vision-language foundation models reproduce the pattern relative to board-certified radiologists [57], and models can infer self-reported race from images alone [58]. The interpretation of these disparities is contested — the use of test data drawn from the same biased distribution as the training data complicates causal attribution [55] — and part of the effect appears to be mediated by acquisition and processing parameters rather than by patient characteristics per se, since demographics-independent calibration reduces the measured bias [58]. But the practical implication is not contested: a model whose subgroup performance has never been measured cannot be assumed to be equitable, and 33 of the 44 studies in this review make no such measurement.

Table 4. Risk of bias across the four PROBAST+AI signalling domains and selected reporting items (n = 44 studies).

Domain / item	Low	High	Unclear	Most frequent reason for downgrading
Participants and data sources	23	9	12	Retrospective single-site convenience sampling; incompletely specified inclusion
Predictors / index test	31	5	8	Pre-processing and image-selection steps not fully specified
Outcome / reference standard	29	6	9	Permissive interval (up to 12 months) between index test and reference standard
Analysis	12	26	6	Internal split only; calibration not assessed; low events-per-parameter
Overall (worst domain)	5	27	12	—
Reported calibration	4	—	—	Discrimination reported in isolation
Sample-size justification	6	—	—	Convenience sample size
Subgroup performance reported	11	—	—	No demographic disaggregation
Explainability method applied	24	—	—	Saliency-based, qualitatively evaluated
Data or code available	9	—	—	No availability statement

3.9 Translational readiness

Positioning every included study on the two-axis matrix produces a distribution that is highly asymmetric (Figure 3, Table 5). Thirty-seven of 44 studies (84%) reached only technical or diagnostic-accuracy efficacy. Six studies reached diagnostic-thinking efficacy, principally through reader studies demonstrating that AI assistance changed radiologist performance [16,26] or that the model exceeded expert readers [44]. One study reached therapeutic efficacy by demonstrating a change in prescribing [37]. No included study demonstrated patient-outcome efficacy — that is, that an AI-informed pathway reduces cardiovascular events — and none reported a formal economic evaluation, although several asserted cost-effectiveness as a prospective claim [6,38,39].

On the generalisability axis, 17 studies remained at internal validation only, 13 achieved temporal or single-site external validation, eight achieved multi-site external validation, four validated across national boundaries [8,15,34,35], and two involved prospective or interventional data collection [31,37]. The upper-right quadrant of the matrix — the region occupied by evidence sufficient to support clinical commissioning — contains a single study.

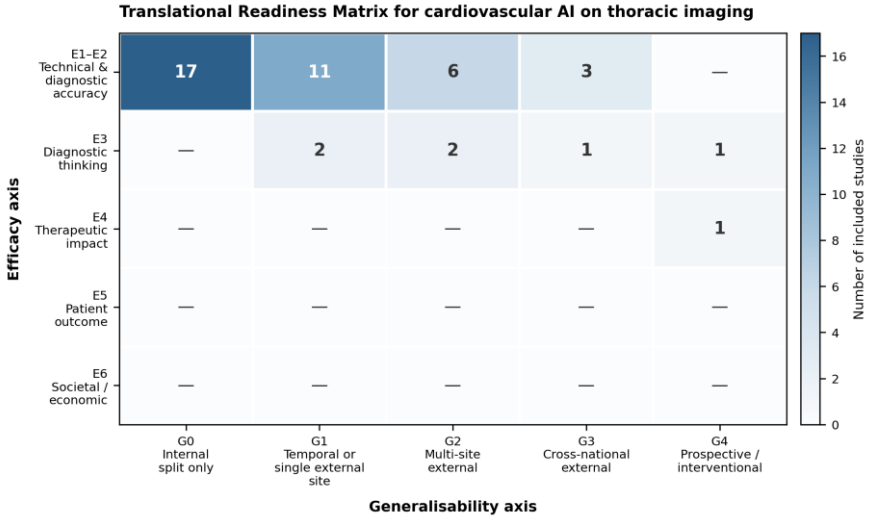

Figure 3. Translational Readiness Matrix. Each included study is positioned at the highest level attained on the efficacy axis (adapted imaging-efficacy hierarchy) and the generalisability axis (validation-design ladder). Cell values are study counts; dashes denote empty cells. The concentration in the upper-left quadrant, and the emptiness of the lower-right, characterise the field's principal evidence gap.

Table 5. Definitions of the two axes of the Translational Readiness Matrix, with the number of included studies attaining each level and a representative example.

Level	Label	Definition	n	Example
E1–E2	Technical and diagnostic accuracy	Model output agrees with a reference standard; discrimination, agreement or calibration reported	37	[32]
E3	Diagnostic thinking	Model output demonstrably changes the diagnostic assessment of a clinician, or exceeds reader performance in a formal comparison	6	[16,26]
E4	Therapeutic impact	Model output demonstrably changes management (testing, prescribing, referral)	1	[37]
E5	Patient outcome	AI-informed pathway shown to change morbidity, mortality or quality of life	0	—
E6	Societal and economic	Formal cost-effectiveness or health-system impact evaluation	0	—
G0	Internal only	Random or cross-validated split of a single dataset	17	[22,24]
G1	Temporal or single external site	Temporally separated cohort, or one external site within the same system	13	[7,44]
G2	Multi-site external	Two or more independent external sites	8	[29,31]
G3	Cross-national external	External validation in a different country or health system	4	[8,15,34,35]
G4	Prospective or interventional	Data collected prospectively in the intended workflow, or randomised	2	[31,37]

4. DISCUSSION

4.1 Principal findings

This review of 44 studies supports three conclusions. First, artificial intelligence applied to thoracic radiological images can recover clinically meaningful cardiovascular information that routine interpretation discards, and the strength of that recovery varies systematically and predictably with the physical directness of the signal. Coronary calcium is directly visualised on CT, and automated quantification accordingly reaches agreement with the gated reference standard that is close to the ceiling of the reference standard's own reproducibility. Coronary calcium is not directly visualised on a radiograph, and radiograph-based inference accordingly plateaus around an AUC of 0.75–0.80 across independent groups, datasets and architectures. This is not a failure of engineering; it is an information-theoretic boundary, and the consistency with which independent teams converge on it is itself evidence that the boundary is real.

Second, the most valuable radiographic models are not the ones that best approximate an existing test. Radiograph-derived risk and biological-age estimates outperform chronological age and add information beyond conventional risk equations [5,11], and they do so most usefully in exactly the population where conventional equations fail — patients whose risk score cannot be computed because the inputs are missing [5]. Framed as a surrogate calcium score, a radiograph model is a poor substitute for CT. Framed as a zero-marginal-cost risk signal available for millions of patients who will never receive a calcium score, it is something CT cannot be.

Third, and decisively, the binding constraint on clinical adoption is no longer discriminative accuracy. Eighty-four percent of the corpus stops at diagnostic accuracy; a single study demonstrates that anything downstream changes. The bottleneck is the evidence architecture, and the corrective is not a better model.

4.2 Three distinct value propositions, requiring three distinct evidence bases

The word "optimization" conceals at least three different claims, which the literature routinely conflates and which require different evidence to substantiate.

- Substitution — the model replaces a test that would otherwise be performed. This requires non-inferiority against the displaced test in the population where substitution would occur, and is the strongest claim in the corpus only for automated calcium scoring on gated CT, where agreement with manual scoring approaches the reproducibility limit of manual scoring itself [31,32].
- Triage — the model reorders who receives a downstream test. This requires a defensible operating point rather than a good AUC, and specifically a negative predictive value adequate to the prevalence of the deployment population. The valvular and pulmonary-hypertension models are triage instruments, and the two studies that

report negative predictive values above 98% [20] make the strongest triage case in this review. Discrimination alone cannot support this claim, because AUC is prevalence-invariant and the clinical question is not.

- Opportunistic detection — the model surfaces information in an image acquired for an unrelated reason, in a patient who was never going to be tested. This is the proposition with the largest population reach and the weakest evidentiary tradition, because its benefit is realised only if detection changes management. The randomised notification study [37] is the only study in this review that tests that link, and its effect size — a sevenfold increase in statin initiation — indicates both that the link is real and that it is not automatic: the algorithm had already detected calcium in the usual-care arm, and 93% of those patients still received no statin.

Conflating these three propositions produces the characteristic pathology of the field, in which a model validated as a substitute is proposed as a screening tool, or an AUC obtained in a high-prevalence angiographic referral cohort is offered as evidence for primary-care use. Systematic reviews of commercial radiological AI find precisely this pattern: the proportion of evidence addressing clinical decision-making, patient outcome or economic impact has remained flat at about a quarter even as product numbers have grown, while vendor independence and prospective design have declined [63].

4.3 The generalisation penalty and its sources

The median external decrement of 0.045 AUC points observed here is modest in isolation but concerning in context, for three reasons. It was measured in the subset of studies that chose to report external validation at all — a group that is by construction more methodologically committed than the 17 studies that reported none, so the true field-wide decrement is almost certainly larger. It reflects discrimination only, and calibration typically degrades faster than discrimination under distribution shift, yet was reported in four studies. And it was measured against curated external datasets rather than deployment conditions, where the dominant sources of variance turn out to be technical rather than clinical: projection type alone explained 61–100% of performance variance in a multi-architecture comparison, an order of magnitude more than patient age or sex, and internal robustness rankings reversed entirely on external data [72].

Three implications follow. Foundation-model pretraining does not confer robustness by itself; domain-specific pretraining conferred the best internal robustness and the worst external robustness in the same comparison [72]. Acquisition-parameter auditing — projection, scanner, dose, reconstruction, post-processing — belongs alongside demographic auditing in any regulatory submission, and is currently absent from both. And protocol-specific fine-tuning is an effective and cheap mitigation: adding a few hundred protocol-matched images raised intraclass correlations from 0.79–0.97 to 0.85–0.99 in a six-protocol CT study [32], and a recalibration procedure recovered 53–85% of the performance lost when digital radiograph models were applied to smartphone photographs of radiographs [3] — a deployment scenario of direct relevance to settings where images are not natively digital.

4.4 Equity is a design specification, not a post hoc audit

The strongest ethical argument for opportunistic cardiovascular screening is that it reaches people whom conventional pathways miss — patients with no lipid panel, no continuity of care, and no realistic prospect of a dedicated calcium score. That argument fails if the model performs worst in exactly those groups. The evidence that chest radiograph classifiers underdiagnose under-served populations is now substantial and reproduced across datasets, architectures and paradigms [54,56,57], and models encode protected characteristics whether or not they are supplied as inputs [56,58]. The mechanism is contested and the measurement is genuinely difficult — testing on data drawn from the same biased distribution as training confounds attribution [55], and correcting shortcuts to achieve local fairness within a training distribution does not produce fairness in new deployment settings, where models encoding demographic attributes least tend to generalise best [57].

None of that ambiguity licenses omission, and 75% of the studies in this review omit. The minimum defensible standard is not a fairness guarantee but a fairness measurement: performance disaggregated by sex, age band and, where recorded, ethnicity and insurance status, reported for every validation cohort, with intersectional strata where sample size permits, alongside the acquisition-parameter distribution across those strata. This is a reporting requirement that costs nothing beyond discipline, and it is already codified in TRIPOD+AI [49] and in the fairness recommendations developed specifically for radiology [59,67].

4.5 From accuracy to action

The distribution in Figure 3 is not peculiar to this domain. Across all of clinical medicine, a scoping review identified only 86 randomised controlled trials of AI in clinical practice, of which 81% reported positive primary endpoints but the great majority were single-centre, with sparse demographic reporting and inconsistent measurement of operational efficiency [61]. Meta-analysis of 48 real-world imaging implementations found that although two-thirds of studies reported reduced task time, three separate pooled analyses showed no significant effect after AI implementation, and only three studies attempted a workload calculation [65]. A prospective pre–post study of a computer-aided detection system found that implementation lengthened reading time for high-suspicion cases (15.7 to 23.1 minutes, $p = 0.02$) and left self-reported workload and stress unchanged [74].

This should temper expectations without deflating them. The randomised notification result [37] is instructive precisely because the intervention was not the algorithm: the algorithm had already found the calcium. What changed outcomes was a notification pathway with a radiologist confirmation step, a patient-specific image, and an explicit guideline recommendation delivered to a named clinician. The generalisable lesson is that the unit of evaluation should be the pathway, not the model — which is what DECIDE-AI was designed to formalise for

early-stage clinical evaluation [51], and what the four-phase safety–efficacy–effectiveness–monitoring framework proposed for health-system deployment operationalises [73].

4.6 Recommendations

On the basis of this synthesis we propose six requirements for studies in this domain.

- Report the Translational Readiness position explicitly. Stating a study's efficacy level and generalisability level in the abstract would prevent the routine over-claiming in which an E1/G0 result is discussed as though it were an E4/G3 result, and would give commissioners a single comparable coordinate.
- Report calibration alongside discrimination, with a calibration plot and slope, in every validation cohort. Four of 44 studies did so. A miscalibrated model with excellent discrimination is unusable for the risk-threshold decisions that cardiovascular prevention guidelines actually specify.
- Disaggregate performance by sex, age and, where available, ethnicity in every validation cohort, and report the acquisition-parameter distribution across those strata. Report intersectional strata where power permits, and state explicitly where it does not.
- Specify and justify the operating point for the intended deployment prevalence, and report positive and negative predictive values at that prevalence rather than only prevalence-invariant discrimination.
- Evaluate the pathway, not the model. Where an opportunistic-detection claim is made, the study should specify the notification mechanism, the human confirmation step, and the intended clinical action, and should measure whether that action occurs.
- Prioritise external validation over architectural novelty. The marginal scientific return on a new architecture evaluated on a single internal split is now close to zero in this domain; the marginal return on independent, cross-national, acquisition-diverse validation of an existing open-source model is high, and several such models are publicly available [8,16,35].

5. Limitations

Several limitations qualify these findings. The search was executed through three federated evidence-discovery platforms rather than through native database interfaces with controlled vocabulary; although these platforms jointly index MEDLINE, Scopus, Semantic Scholar and arXiv, semantic retrieval does not guarantee the exhaustiveness of a MeSH-anchored Boolean strategy, and a confirmatory search in MEDLINE and Embase with controlled vocabulary is recommended before this synthesis is treated as complete. The review was not prospectively registered, which weakens the guarantee against post hoc modification of the protocol. Screening, extraction and risk-of-bias assessment were not performed in duplicate by independent reviewers, so no inter-rater agreement statistic can be reported; this is a meaningful departure from best practice and readers should weight the risk-of-bias ratings accordingly.

English-language restriction may have excluded relevant work, particularly from East Asian centres that are prominent in this field. Five conference abstracts were retained on methodological grounds but report incomplete methods and have not undergone full peer review; excluding them would not alter any principal conclusion. The exploratory generalisation analysis rests on six studies, is unweighted, and cannot support inference; it is presented as an observation, not an estimate. Finally, the Translational Readiness Matrix is a novel instrument that has not been validated, and the assignment of studies to levels involved judgement, particularly in distinguishing E3 from E1–E2 where reader comparisons were informal. It should be read as a structuring device for discussion rather than as a measurement.

6. CONCLUSIONS

Thoracic radiological images acquired for unrelated indications contain a cardiovascular signal that artificial intelligence can extract at scale and at essentially zero marginal cost. For coronary artery calcium on non-gated chest CT, the technical question is settled: automated quantification agrees with the gated reference standard within its own reproducibility, predicts events independently of established risk factors, and has been demonstrated across a national health system. For chest radiography, performance is more modest and appears to be bounded by the physics of the modality rather than by the sophistication of the model, but the resulting risk signal carries prognostic information beyond chronological age and beyond conventional risk equations — and it is available precisely for the patients whom conventional equations cannot score.

What the field has not established is that any of this makes patients better off. Eighty-four percent of the studies synthesised here stop at diagnostic accuracy; three-quarters never measure whether performance holds across demographic subgroups; and one study of 44 demonstrates a change in clinical management. That single study is also the most encouraging, because it shows the detection-to-action gap is closable by unglamorous means — a notification, a confirmation step, a named clinician. The next decade of work in this area should be judged less by the discrimination it reports than by the distance it travels up and to the right of the matrix presented here.

REFERENCES

1. Farina J, Chao CJ, Ghurbhurrun A, et al. Artificial intelligence-based prediction of cardiovascular diseases from chest radiography. J Imaging. 2023;9(11):236. doi:10.3390/jimaging9110236
2. Foraker R, Blaha MJ, Nasir K, et al. Opportunistic detection of coronary artery calcium on non-cardiac chest computed tomography: an emerging tool for cardiovascular disease prevention. Circulation. 2025. doi:10.1161/CIR.0000000000001382
3. Kuo PC, Tsai CC, López DM, et al. Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph. NPJ Digit Med. 2021;4(1):25. doi:10.1038/s41746-021-00393-9
4. Kamel PI, Yi PH, Sair HI, Lin CT. Prediction of coronary artery calcium and cardiovascular risk on chest radiographs using deep learning. Radiol Cardiothorac Imaging. 2021;3(3):e200486. doi:10.1148/ryct.2021200486
5. Weiss J, Raghu VK, Paruchuri K, et al. Deep learning to estimate cardiovascular risk from chest radiographs: a risk prediction study. Ann Intern Med. 2024;177(4):409-417. doi:10.7326/M23-1898
6. Hong Y, Kim J, Chang HJ, et al. Predicting categories of coronary artery calcium scores from chest X-ray images using deep learning. J Cardiovasc Comput Tomogr. 2025. doi:10.1016/j.jcct.2025.03.010
7. Gallone G, Iodice F, Presta A, et al. Detection of subclinical atherosclerosis by image-based deep learning on chest X-ray. Eur Heart J Digit Health. 2025;6(4):567-576. doi:10.1093/ehjdh/ztaf033
8. Jeong J, Patel B, Banerjee I, et al. Artificial intelligence chest X-ray opportunistic screening model for coronary artery calcium deposition: a multi-objective model with multimodal data fusion. Mayo Clin Proc Digit Health. 2025;3:100300. doi:10.1016/j.mcpdig.2025.100300
9. D'Ancona G, Massussi M, Savardi M, et al. Deep learning to detect significant coronary artery disease from plain chest radiographs AI4CAD. Int J Cardiol. 2022;370:435-441. doi:10.1016/j.ijcard.2022.10.154
10. D'Ancona G, Savardi M, Massussi M, et al. Deep learning to predict long-term mortality from plain chest radiographs in patients referred for suspected angina. Eur Heart J. 2023;44(Suppl 2):ehad655.1464. doi:10.1093/eurheartj/ehad655.1464
11. Raghu VK, Weiss J, Hoffmann U, Aerts HJWL, Lu MT. Deep learning to estimate biological age from chest radiographs. JACC Cardiovasc Imaging. 2021;14(11):2226-2236. doi:10.1016/j.jcmg.2021.01.008
12. Ieki H, Ito K, Saji M, et al. Deep learning-based age estimation from chest X-rays indicates cardiovascular prognosis. Commun Med. 2022;2:159. doi:10.1038/s43856-022-00220-6
13. Liao HC, Lin C, Fang WH, et al. The deep learning algorithm estimates chest radiograph-based sex and age as independent risk factors for future cardiovascular outcomes. Digit Health. 2023;9:20552076231191055. doi:10.1177/20552076231191055
14. Deguchi R, Mitsuyama Y, Walston SL, et al. External validation study of a chest radiograph-derived aging biomarker for ICU mortality prediction. Preprint. 2025. doi:10.21203/rs.3.rs-6708615/v1
15. Ueda D, Matsumoto T, Ehara S, et al. Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study. Lancet Digit Health. 2023;5(8):e525-e533. doi:10.1016/S2589-7500(23)00107-3
16. Bhave S, Rodriguez V, Poterucha T, et al. Deep learning to detect left ventricular structural abnormalities in chest X-rays. Eur Heart J. 2024;45(22):2002-2012. doi:10.1093/eurheartj/ehad782
17. Hsiang CW, Lin C, Liu WC, et al. Detection of left ventricular systolic dysfunction using an artificial intelligence-enabled chest X-ray. Can J Cardiol. 2022;38(6):763-773. doi:10.1016/j.cjca.2021.12.019
18. Lee JG, Kim H, Kang JW, et al. Automated, standardized, quantitative analysis of cardiovascular borders on chest X-rays using deep learning for assessing cardiovascular disease. medRxiv. 2024. doi:10.1101/2024.07.17.24310314
19. White RD, Erdal BS, Demirer M, et al. Artificial intelligence improves detection and classification of pulmonary venous hypertension related to left ventricular diastolic dysfunction by chest radiography. Sci Rep. 2025;15:22026. doi:10.1038/s41598-025-22026-x
20. Liu PY, Lin C, Lin CS, et al. A deep-learning-enabled electrocardiogram and chest X-ray for detecting pulmonary arterial hypertension. J Imaging Inform Med. 2024. doi:10.1007/s10278-024-01225-4
21. Sogancioglu E, Murphy K, Çallı E, Scholten ET, Schalekamp S, van Ginneken B. Cardiomegaly detection on chest radiographs: segmentation versus classification. IEEE Access. 2020;8:94631-94642. doi:10.1109/ACCESS.2020.2995567
22. Bougias H, Georgiadou E, Malamateniou C, Stogiannos N. Identifying cardiomegaly in chest X-rays: a cross-sectional study of evaluation and comparison between different transfer learning methods. Acta Radiol. 2021;62(12):1601-1609. doi:10.1177/0284185120973630
23. Matusik P, Tabor Z, Kucybała I, Jarczewski JD, Popiela TJ. New approaches to AI methods for screening cardiomegaly on chest radiographs. Appl Sci. 2024;14(24):11605. doi:10.3390/app142411605
24. Lee KH, Choi JW, Park C, Han DH, Kang M. A development and validation of an AI model for cardiomegaly detection in chest X-rays. Appl Sci. 2024;14(17):7465. doi:10.3390/app14177465
25. Bouslama A, Laaziz Y, Tali A. Diagnosis and precise localization of cardiomegaly disease using U-NET. Inform Med Unlocked. 2020;19:100306. doi:10.1016/j.imu.2020.100306
26. Han P, Liu X, Zhang Y, et al. Artificial intelligence-assisted diagnosis of congenital heart disease and associated pulmonary arterial hypertension from chest radiographs: a multi-reader multi-case study. Eur J Radiol. 2023;169:111277. doi:10.1016/j.ejrad.2023.111277

27. Li Z, Luo G, Ji Z, Wang S, Pan W. Explanatory deep learning to predict elevated pulmonary artery pressure in children with ventricular septal defects using standard chest x-rays: a novel approach. Front Cardiovasc Med. 2024;11:1330685. doi:10.3389/fcvm.2024.1330685
28. Luo G, Li Z, Wang S, et al. A multimodal multipath AI system for assessing PAH after VSD correction on echocardiography and chest radiography images. Sci Rep. 2025;15:19584. doi:10.1038/s41598-025-19584-5
29. Zeleznik R, Foldyna B, Eslami P, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nat Commun. 2021;12:715. doi:10.1038/s41467-021-20966-2
30. Chao H, Shan H, Homayounieh F, et al. Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography. Nat Commun. 2021;12:2963. doi:10.1038/s41467-021-23235-4
31. Eng D, Chute C, Khandwala N, et al. Automated coronary calcium scoring using deep learning with multicenter external validation. NPJ Digit Med. 2021;4:88. doi:10.1038/s41746-021-00460-1
32. van Velzen SGM, Lessmann N, Velthuis BK, et al. Deep learning for automatic calcium scoring in CT: validation using multiple cardiac CT and chest CT protocols. Radiology. 2020;295(1):66-79. doi:10.1148/radiol.2020191621
33. van Assen M, Martin SS, Varga-Szemes A, et al. Automatic coronary calcium scoring in chest CT using a deep neural network in direct comparison with non-contrast cardiac CT: a validation study. Eur J Radiol. 2021;134:109428. doi:10.1016/j.ejrad.2020.109428
34. Xu C, Guo H, Xu M, et al. Automatic coronary artery calcium scoring on routine chest computed tomography: comparison of a deep learning algorithm and a dedicated calcium scoring CT. Quant Imaging Med Surg. 2022;12(4):2684-2695. doi:10.21037/qims-21-1017
35. Hagopian R, Strait T, Kim E, et al. AI opportunistic coronary calcium screening at Veterans Affairs hospitals. NEJM AI. 2025. doi:10.1056/AIoa2400937
36. Peng AW, Dudum R, Jain SS, et al. Association of coronary artery calcium detected by routine ungated CT imaging with cardiovascular outcomes. J Am Coll Cardiol. 2023;82(12):1192-1202. doi:10.1016/j.jacc.2023.06.040
37. Sandhu AT, Rodriguez F, Ngo S, et al. Incidental coronary artery calcium: opportunistic screening of prior non-gated chest CTs to improve statin rates (NOTIFY-1 project). Circulation. 2023;147(9):703-714. doi:10.1161/CIRCULATIONAHA.122.062746
38. Mineo E, Nomura CH, Assunção AN, et al. Automated coronary artery calcium scoring using deep learning: validation across diverse chest CT protocols. Acad Radiol. 2025. doi:10.1016/j.acra.2025.09.009
39. Rathore S, Gautam S, Agarwal V, et al. Fully automated coronary artery calcium score and risk categorization from chest CT using deep learning and multiorgan segmentation: a validation study from the National Lung Screening Trial. Int J Cardiol Heart Vasc. 2025;56:101593. doi:10.1016/j.ijcha.2024.101593
40. Kim NY, Hong Y, Suh YJ, et al. Opportunistic assessment of coronary artery calcium volume and density from non-electrocardiogram-gated chest CT using artificial intelligence: prognostic implications in a screening cohort. Korean J Radiol. 2026. doi:10.3348/kjr.2025.1614
41. Sartoretti E, Sartoretti T, Gutzwiller A, et al. Opportunistic deep learning powered calcium scoring in oncologic patients with very high coronary artery calcium undergoing 18F-FDG PET/CT. Sci Rep. 2022;12:15927. doi:10.1038/s41598-022-20005-0
42. Anifowose JO, Ezhil V, Nair A, et al. Automated opportunistic cardiovascular risk assessment in non-small cell lung cancer patients on routine chest CT using an optimised nnU-net framework. BMC Med Imaging. 2026. doi:10.1186/s12880-026-02252-z
43. Raghu VK, Weiss J, Aerts HJWL, Lu MT. Opportunistic assessment of cardiovascular risk using deep learning of the heart and aorta on non-contrast chest computed tomography. Circulation. 2024;150(Suppl 1):A4141180. doi:10.1161/circ.150.suppl_1.4141180
44. Raikhelkar J, Fiman A, Kagen A, et al. An artificial intelligence model to detect abnormal ejection fraction from non-contrast chest computed tomography: the CT–LVEF study. Eur Heart J Digit Health. 2026. doi:10.1093/ehjdh/ztag088
45. Paruchuri K, Weiss J, Raghu VK, et al. Polygenic risk score and deep learning using chest radiographs: complementary prediction of incident coronary artery disease. Circulation. 2021;144(Suppl 1):A12397. doi:10.1161/circ.144.suppl_1.12397
46. Chandra J, Raghu VK, Weiss J, et al. Opportunistic screening for cardiovascular risk using chest X-rays and deep learning: associations with coronary artery disease in the Project Baseline Health Study and Mass General Brigham Biobank. Circulation. 2024;150(Suppl 1):A4138647. doi:10.1161/circ.150.suppl_1.4138647
47. Du R, Lin M, Yeo SY, et al. Machine-extractable markers in chest radiograph to predict cardiovascular risk in screening population. Circulation. 2024;150(Suppl 1):A4121454. doi:10.1161/circ.150.suppl_1.4121454
48. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71
49. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378

50. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505
51. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28:924-933. doi:10.1038/s41591-022-01772-9
52. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the quality assessment of diagnostic accuracy studies using AI (QUADAS-AI): protocol for a qualitative study. JMIR Res Protoc. 2024;13:e58202. doi:10.2196/58202
53. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26:1320-1324. doi:10.1038/s41591-020-1041-y
54. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27:2176-2182. doi:10.1038/s41591-021-01595-0
55. Bernhardt M, Jones C, Glocker B. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. Nat Med. 2022;28:1157-1158. doi:10.1038/s41591-022-01846-8
56. Glocker B, Jones C, Bernhardt M, Winzeck S. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. EBioMedicine. 2023;89:104467. doi:10.1016/j.ebiom.2023.104467
57. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. Nat Med. 2024;30:2838-2848. doi:10.1038/s41591-024-03113-4
58. Lotter W. Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. Nat Commun. 2024;15:7620. doi:10.1038/s41467-024-52003-3
59. Banerjee I, Bhattacharjee K, Burns JL, et al. "Shortcuts" causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. J Am Coll Radiol. 2023;20(9):842-851. doi:10.1016/j.jacr.2023.06.025
60. Kelly BS, Judge C, Bollard SM, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol. 2022;32:7998-8007. doi:10.1007/s00330-022-08784-6
61. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health. 2024;6(5):e367-e373. doi:10.1016/S2589-7500(24)00047-5
62. Windecker D, Baj G, Shiri I, et al. Generalizability of FDA-approved AI-enabled medical devices for clinical use. JAMA Netw Open. 2025;8(4):e258052. doi:10.1001/jamanetworkopen.2025.8052
63. Antonissen N, Tulk AS, Kwee TC, et al. Artificial intelligence in radiology: 173 commercially available products and their scientific evidence. Eur Radiol. 2025. doi:10.1007/s00330-025-11830-8
64. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health. 2021;3(3):e195-e203. doi:10.1016/S2589-7500(20)30292-2
65. Wenderott K, Krups J, Zaruchas F, Weigl M. Effects of artificial intelligence implementation on efficiency in medical imaging: a systematic literature review and meta-analysis. NPJ Digit Med. 2024;7:265. doi:10.1038/s41746-024-01248-9
66. Haupt M, Wehrse E, Kim SH, et al. Explainable artificial intelligence in radiological cardiovascular imaging: a systematic review. Diagnostics. 2025;15(11):1399. doi:10.3390/diagnostics15111399
67. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. 2024;42:3-15. doi:10.1007/s11604-023-01474-3
68. Nayebirad S, Javadi M, Tavolinejad H. Challenges of using artificial intelligence to detect valvular heart disease from chest radiography. Lancet Digit Health. 2024;6(2):e88. doi:10.1016/S2589-7500(23)00223-6
69. Lawrence R, Fulop NJ, Ramsay AIG, et al. Artificial intelligence for diagnostics in radiology practice: a rapid systematic scoping review. EClinicalMedicine. 2025;85:103228. doi:10.1016/j.eclinm.2025.103228
70. Hsieh CF, Peng HH, Tsai YH, et al. An adaptive deep learning framework for multi-label chest X-ray diagnosis using a hybrid CNN–transformer architecture and class-wise ensemble fusion. Diagnostics. 2026;16(8):1227. doi:10.3390/diagnostics16081227
71. Liu Z, Xu J, Yin C, et al. Development and external validation of an artificial intelligence-based method for scalable chest radiograph diagnosis: a multi-country cross-sectional study. Research. 2024;7:0426. doi:10.34133/research.0426
72. Farquhar H. Foundation model robustness to technical acquisition parameters in chest X-ray AI: a multi-architecture comparative study with external validation. medRxiv. 2026. doi:10.64898/2026.01.25.26344809
73. You JG, Hernandez-Boussard T, Pfeffer MA, Landman A, Mishuris RG. Clinical trials informed framework for real world clinical implementation and deployment of artificial intelligence applications. NPJ Digit Med. 2025;8:147. doi:10.1038/s41746-025-01506-4
74. Wenderott K, Krups J, Luetkens JA, Weigl M. Prospective effects of an artificial intelligence-based computer-aided detection system for prostate imaging on routine workflow and radiologists' outcomes. Eur J Radiol. 2023;170:111252. doi:10.1016/j.ejrad.2023.111252