International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	TRUST AND RELIABILITY OF AI SYSTEMS IN EARLY CANCER DIAGNOSTICS: A NARRATIVE REVIEW
DOI	https://doi.org/10.31435/ijitss.2(50).2026.5906
RECEIVED	13 March 2026
ACCEPTED	12 June 2026
PUBLISHED	15 June 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

TRUST AND RELIABILITY OF AI SYSTEMS IN EARLY CANCER DIAGNOSTICS: A NARRATIVE REVIEW

Emil Pałyga *(Corresponding Author, Email: emilpalyga212@gmail.com)*
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0000-6614-964X

Mateusz Kwiatkowski
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0008-1099-1676

Zofia Leżańska
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0006-6808-5006

Katarzyna Marcinkowska
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0005-2930-1805

Aleksandra Cieślak
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0006-4901-4341

Sara Demkow
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0007-7192-7435

Karolina Siemińska
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0009-7712-4259

Joanna Sowińska
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0007-9507-6639

Natalia Paluszkiewicz
Medical University of Warsaw, Warsaw, Poland
ORCID ID: 0009-0001-6367-1018

Sandra Bryg
Medical University of Silesia in Katowice, Katowice, Poland
ORCID ID: 0009-0003-6539-6595

ABSTRACT

Artificial intelligence (AI) tools are now embedded in many early cancer diagnostic workflows, with selected systems matching or surpassing human readers in mammography, low-dose computed tomography, colonoscopy, dermoscopy, and digital pathology. Whether such tools can be safely scaled depends on more than headline accuracy. Clinicians, patients, regulators, and health systems must be able to trust them and rely on their behaviour across diverse populations and over time. This narrative review draws together evidence from PubMed and PubMed Central published between 2018 and 2026, with priority given to randomised trials, large prospective studies, systematic reviews, and authoritative regulatory analyses. We examine six interrelated dimensions of trust and reliability: diagnostic performance against clinical reference standards; external validity and generalisability across populations, scanners, and time; algorithmic bias and health equity; transparency through explainable AI methods; human–AI interaction, including automation bias; and the evolving regulatory landscape under the United States Food and Drug Administration and the European Union AI Act. Recent evidence shows that well-validated AI can raise cancer detection rates and reduce reader workload in real-world screening. Reliability, however, is constrained by limited external validation, demographic under-representation, opaque decision-making, and uneven post-market surveillance. Trust in this domain is not an intrinsic property of any single model. It emerges from rigorous validation, transparent reporting, equitable performance, regulatory oversight, and meaningful human oversight maintained throughout the lifecycle of every deployed tool.

KEYWORDS

Artificial Intelligence, Early Cancer Detection, Algorithmic Bias, Explainable AI, Clinical Validation, Regulatory Oversight

CITATION

Emil Pałyga, Mateusz Kwiatkowski, Zofia Leżańska, Katarzyna Marcinkowska, Aleksandra Cieślak, Sara Demkow, Karolina Siemińska, Joanna Sowińska, Natalia Paluszkiewicz, Sandra Bryg. (2026) Trust and Reliability of AI Systems in Early Cancer Diagnostics: A Narrative Review. *International Journal of Innovative Technologies in Social Science*. 2(50). doi: 10.31435/ijitss.2(50).2026.5906

1. Introduction

Cancer is among the leading causes of morbidity and mortality worldwide. The World Health Organization estimated approximately 20 million new cases and nearly 10 million deaths in 2022 (Bray et al., 2024). Outcomes are strongly stage-dependent: five-year survival for localised breast, colorectal, and lung cancer substantially exceeds that observed when disease is detected at advanced stages. That gradient has driven decades of investment in population-based screening and early diagnostic pathways. Yet the workforce supporting these pathways—radiologists, pathologists, endoscopists, dermatologists—faces a widening gap between rising demand and limited capacity, with downstream consequences for interpretive variability, missed cancers, and clinician burnout.

Within the last decade, AI—and in particular deep learning models trained on large medical imaging or histopathology datasets—has emerged as a candidate technology for closing parts of that gap. AI-based tools have reached performance approaching or exceeding board-certified specialists in several modalities. Mammography is the most mature example (Lång et al., 2023; Hernström et al., 2025). Comparable evidence now exists for low-dose computed tomography (CT) of the lung (Geppert et al., 2024), real-time computer-aided detection during colonoscopy (Soleymanjahi et al., 2024), dermoscopic image classification of pigmented and non-pigmented skin lesions (Esteva et al., 2017; Tschandl et al., 2019), and Gleason grading of prostate biopsies (Bulten et al., 2022). By October 2023,the United States Food and Drug Administration (FDA) had authorised 691 AI- and machine-learning-enabled medical devices, the vast majority in radiology (Joshi et al., 2024).

Enthusiasm regarding diagnostic performance has been tempered by substantive concerns. Two related but distinct concepts dominate the debate. Reliability is a technical property: it captures whether an AI system delivers consistent, accurate, and safe outputs across new patient populations, different scanners, evolving

clinical practice, and the passage of time. Trust is relational and sociotechnical. It captures the extent to which clinicians, patients, regulators, and health systems are willing to act on an AI system's outputs given what they know about its development, validation, governance, and oversight (Hassan et al., 2024; van Kolfschooten & van Oirschot, 2024). The two are not equivalent. A model can be reliable yet poorly trusted, or trusted on the basis of promotional claims without robust evidence of reliability.

Several recurring challenges complicate this relationship in early cancer diagnostics. High-performing models are often validated on retrospective, single-centre datasets and may degrade when transposed to other populations, vendors, or screening intervals (Anderson et al., 2022; Schurz et al., 2025). Training data frequently under-represent groups defined by skin tone, age, sex, or socioeconomic status, raising the prospect of systematic bias and disparate harm (Daneshjou et al., 2022; Drukker et al., 2023). The deep learning architectures behind most state-of-the-art systems function as opaque black boxes, which limits clinicians' ability to scrutinise individual predictions and has fuelled work on explainable AI (XAI) methods such as Shapley Additive exPlanations (SHAP) and class activation maps (Ghasemi et al., 2024). Decision support tools can also induce automation bias, leading readers to follow incorrect AI suggestions (Dratsch et al., 2023; Kromrey et al., 2024). On top of all this, the regulatory architecture for AI-enabled medical devices is still maturing. The FDA's 2024 final guidance on predetermined change control plans and the European Union's AI Act of 2024 both aim to introduce lifecycle oversight for systems whose behaviour can evolve over time (United States Food and Drug Administration, 2024; Vardas et al., 2025).

This narrative review synthesises evidence from roughly 2018 to 2026 to map how trust and reliability are currently established—or undermined—for AI systems used in early cancer diagnostics. We pursue three objectives. The first is to summarise diagnostic performance across the main early-detection modalities. The second is to identify the principal technical, clinical, ethical, and regulatory determinants of reliability and trust. The third is to draw implications for clinicians, developers, and policymakers seeking to embed AI safely in cancer screening and early-diagnosis pathways. The review is written for a multidisciplinary readership across the social and health sciences, in line with the scope of the International Journal of Innovative Technologies in Social Science.

2. Methodology

The article was prepared as a narrative literature review, with reporting informed by the principles of the Scale for the Assessment of Narrative Review Articles (SANRA) (Baethge et al., 2019). A narrative design suits material that spans multiple imaging modalities, regulatory jurisdictions, and disciplinary traditions—radiology, oncology, pathology, computer science, ethics, and health policy. A single-modality systematic review would have been ill suited to that breadth.

Searches were run between January and May 2026 in PubMed and PubMed Central as primary sources, with Google Scholar and forward citation tracking used to identify additional papers. Search terms combined the concepts "artificial intelligence," "machine learning," "deep learning," "convolutional neural network," or "explainable AI" with "cancer," "oncology," "screening," "early detection," "mammography," "low-dose CT," "lung cancer screening," "colonoscopy," "computer-aided detection," "dermoscopy," "melanoma," "digital pathology," or "Gleason grading," and with "trust," "reliability," "bias," "fairness," "generalizability," "external validation," "automation bias," "regulation," "FDA," or "EU AI Act." Priority was given to full-text articles available through PubMed Central or other open-access channels.

Sources were selected with preference for, in roughly descending order, randomised clinical trials and large prospective cohort studies of AI in cancer screening; recent systematic reviews and meta-analyses; authoritative regulatory guidance and consensus statements from professional societies; and qualitative studies of clinician and patient attitudes towards AI in oncology. Earlier landmark studies were kept where they constitute foundational evidence for a given claim. Studies focused solely on AI for cancer treatment planning, drug discovery, or genomic profiling without diagnostic application were excluded, as were preprints not yet published in peer-reviewed venues. Citation details were checked against PubMed records, and reference formatting follows the seventh edition of the Publication Manual of the American Psychological Association (American Psychological Association, 2020), consistent with the journal's requirements.

3. Results

3.1. Diagnostic performance of AI systems in early cancer detection

3.1.1. Breast cancer screening with mammography

Breast cancer screening has produced the most extensive prospective evidence base for AI in early cancer diagnostics. The Mammography Screening with Artificial Intelligence (MASAI) trial, a randomised, controlled, single-blinded, non-inferiority study run within Sweden's population-based screening programme, randomised more than 80,000 women aged 40–80 years to AI-supported screening using the Transpara system or to standard double reading. The prespecified clinical safety analysis, published in The Lancet Oncology, found that AI-supported screening yielded a cancer detection rate of 6.1 per 1,000 screened participants, compared with 5.1 per 1,000 with double reading, with comparable recall rates and a 44.3% reduction in screen-reading workload (Lång et al., 2023). The screening performance analysis in The Lancet Digital Health confirmed a 29% relative increase in screen-detected cancers without a meaningful rise in false positives, and reported an increased detection of invasive cancers (270 vs 217, proportion ratio 1.24), mainly small T1 and lymph-node-negative tumours (Hernström et al., 2025).

Other prospective and retrospective programmes have reinforced the picture. Lauritzen and colleagues described a Danish AI-based mammography screening protocol that maintained accuracy while reducing the reader workload required to maintain double reading (Lauritzen et al., 2022). Ng and colleagues reported that prospective deployment of an AI tool across a Hungarian screening service improved early detection compared with the pre-implementation period (Ng et al., 2023). An independent external validation review concluded that contemporary commercial systems showed stand-alone AUC improvements of 0.02–0.13 over radiologists in some studies, with several matching or exceeding the average performance of experienced radiologists (Anderson et al., 2022). The benefits are largely realised when AI is used as a supplemental or triage reader, rather than as an autonomous decision-maker, and remain dependent on local calibration.

3.1.2. Lung cancer screening with low-dose computed tomography

Lung cancer screening is the second-best-evidenced modality. A systematic review of test-accuracy studies of CE-marked AI software for nodule and cancer detection on low-dose CT, conducted by Geppert and colleagues and published in Thorax, found that AI assistance was generally faster than unaided reading and increased sensitivity for actionable and malignant nodules by approximately 5–20% and 3–15% respectively, at the cost of a 3–8% reduction in specificity (Geppert et al., 2024). Modelling at an assumed 0.5% cancer prevalence suggested that AI-assisted reading could yield 150–750 additional cancers detected per million screened individuals, against an additional 59,700–79,600 individuals undergoing unnecessary CT surveillance. The authors emphasised that most evidence is retrospective and at high or unclear risk of bias, and called for prospective in-practice evaluations.

3.1.3. Computer-aided detection in colonoscopy

Real-time computer-aided detection (CADe) during colonoscopy is one of the few oncology applications evaluated through multiple randomised controlled trials. A systematic review and meta-analysis published in the Annals of Internal Medicine pooled data from 44 randomised trials and reported that CADe significantly increased the adenoma detection rate (ADR), with a relative gain of approximately 21% (rate ratio 1.21, 95% CI 1.15–1.28) over standard colonoscopy and consistent benefit across the different neural network architectures evaluated (Soleymanjahi et al., 2024). The same review found improvements in polyp detection rate and in the detection of sessile serrated lesions, although the long-term effect on colorectal cancer incidence and mortality remains to be confirmed through interval-cancer outcome studies.

3.1.4. Dermoscopic and clinical image analysis

Deep learning models for dermoscopic image classification have repeatedly approached dermatologist-level accuracy in melanoma recognition. The foundational study by Esteva and colleagues used a convolutional neural network trained on nearly 130,000 clinical images to classify skin lesions at expert level (Esteva et al., 2017). In the international, web-based diagnostic study by Tschandl and colleagues, top machine-learning algorithms outperformed expert dermatologists across a 7-class pigmented-lesion task using dermoscopic images, though the authors stressed that the data were curated and that real-world dermatology workflows differ substantially from such test conditions (Tschandl et al., 2019). Hybrid AI–dermatologist workflows have since been shown to outperform either component alone for skin cancer triage, particularly for less experienced clinicians.

3.1.5. AI in digital pathology

Digital pathology has produced one of the most ambitious external evaluations of AI in oncology. The Prostate cANcer graDe Assessment (PANDA) challenge was the largest histopathology competition to date, with 1,290 developers from 1,010 teams analysing more than 10,600 prostate biopsies. A diverse set of submitted algorithms reached pathologist-level performance on independent United States and European validation cohorts, with quadratically weighted Cohen's κ values of 0.862 (95% confidence interval 0.840–0.884) and 0.868 (0.835–0.900) against expert uropathologist consensus (Bulten et al., 2022). An independent external validation of a commercial prostate cancer detection algorithm reported strong diagnostic accuracy at an independent institution, with sensitivity of 97.7% and specificity of 99.3% on biopsies processed outside the algorithm's original training site (Perincheri et al., 2021). For breast, lung, and colorectal histopathology, AI-based tools have shown high concordance with expert pathologists, although prospective trials and outcome data remain comparatively scarce (van der Laak et al., 2021).

3.2. Reliability: external validation, generalisability, and dataset shift

Performance in development cohorts is not equivalent to reliability under real-world clinical deployment. Independent external validation studies of mammography AI algorithms have repeatedly noted that the supporting cohorts are mostly homogeneous in race, ethnicity, vendor, and screening setting, which limits the inferences that can be drawn about behaviour in more diverse populations (Anderson et al., 2022).

Experimental evidence of dataset shift comes from the retrospective VAIB study, which simulated mismatches between AI calibration and the target population by manipulating temporal, demographic, breast-density, and vendor characteristics of a Swedish mammography cohort. The reported distortions in cancer detection rate were substantial and varied by mismatch type: acquisition-year mismatch shifted detection by −3% to +19%, age mismatch by −0.2% to +27%, breast-density mismatch by +1% to +16%, and vendor mismatch by −32% to +33%, with parallel swings in false positive rate (Schurz et al., 2025). These findings carry a substantial clinical implication: a single algorithm can behave markedly differently when transferred between centres or screening eras, even within one country, indicating that local recalibration is a precondition for safe use rather than an optional refinement.

Generalisability has also been tested in lower- and middle-income settings, where data availability and patient demographics differ markedly from high-income screening programmes. Mammography-derived AI risk models trained on European screening cohorts have been validated in United States cohorts of White and Black women with age-adjusted AUCs around 0.67–0.70, offering some reassurance about cross-population transportability while also showing the modest performance ceiling of current breast cancer risk-prediction tools (Gastounioti et al., 2022). Beyond risk prediction, broader reviews of AI in oncology imaging have repeatedly recommended that any deployment outside the original training distribution include a documented recalibration step (Hosny et al., 2018; van der Laak et al., 2021).

3.3. Algorithmic bias, fairness, and health equity

Algorithmic bias in cancer diagnostics tends to follow the contours of the training data. The clearest empirical illustration comes from dermatology. In a curated, pathologically confirmed clinical image set spanning the full Fitzpatrick skin-type spectrum, AI algorithms developed on standard public dermoscopy datasets dropped substantially in performance on darker skin tones and on uncommon diagnoses. Modern robust training methods alone could not close the gap without diverse training data (Daneshjou et al., 2022).

The clinical stakes are high. Black patients with melanoma already face worse survival than White patients despite lower incidence, with five-year survival of roughly 70% versus 94% (Norori et al., 2021). An AI tool that systematically underdetects melanoma in darker skin would not be a neutral failure; it would compound an existing inequity.

Dermatology is not unique. Demographic under-representation has been documented in breast cancer risk prediction, where many models were trained mostly on women of European ancestry (Gastounioti et al., 2022). A wider review of bias in AI for medical imaging argued that bias enters at every stage of the pipeline—data collection, annotation, model selection, training, evaluation, and deployment—and that bias auditing and inclusive dataset construction must be integral, not optional, components of clinical AI development (Drukker et al., 2023).

3.4. The black-box problem and explainable AI

The opacity of deep learning models is repeatedly cited as an impediment to clinician trust. Explainable AI (XAI) methods aim to mitigate this limitation by providing human-interpretable rationales for individual predictions or for overall model behaviour. A systematic scoping review by Ghasemi and colleagues examined 30 studies applying XAI to breast cancer detection and risk prediction, and identified SHAP as the dominant model-agnostic technique. SHAP was most often paired with tree-based ensemble learners to explain biomarker importance, diagnostic classifications, and survival predictions (Ghasemi et al., 2024). The authors concluded that XAI can meaningfully improve transparency, interpretability, fairness, and trustworthiness, but only when implemented as part of an iterative collaboration between clinicians and AI engineers rather than as a superficial transparency feature.

Two caveats temper the enthusiasm. Post-hoc explanation methods can themselves be unstable or unfaithful to the underlying model logic, and may give clinicians a false sense of understanding. Beyond technique, the broader landscape of FDA-cleared AI imaging devices documented by Joshi and colleagues is dominated by radiology and 510(k) substantial-equivalence pathways, with explainability and subgroup behaviour only partially addressed in many public approval summaries (Joshi et al., 2024). Useful clinical transparency therefore needs more than XAI methods. It also needs model cards, audit trails, and standardised product labels that communicate intended population, performance limits, subgroup performance, and uncertainty.

3.5. Human–AI interaction and automation bias

Even an accurate AI tool can degrade clinical performance when used inappropriately by clinicians. Automation bias—the tendency to over-rely on automated systems and to discount one's own judgment—has now been documented across several cancer-related AI applications. In a multi-reader study of breast cancer screening by Dratsch and colleagues, radiologists at three experience levels read mammograms with purported AI Breast Imaging Reporting and Data System (BI-RADS) suggestions, some of which were deliberately incorrect. The proportion of correctly rated mammograms fell sharply for every group when AI suggestions were wrong. Inexperienced radiologists were significantly more likely to follow incorrectly higher BI-RADS suggestions than very experienced readers, who were affected to a lesser degree, although no experience level was wholly immune to the effect (Dratsch et al., 2023).

Outside breast imaging, similar dynamics have been described in chest radiography. Kromrey and colleagues compared 1,506 consecutive chest X-rays in a maximum-care hospital, re-evaluating them with a commercial AI software. The system detected thoracic pathologies more often than the radiologists augmented by AI (18.5% vs 11.1% positive readings), but also generated many AI-only findings that radiologists did not confirm, illustrating both the diagnostic added value of AI as an adjunct reader and the risk of over-detection when AI output is incorporated without robust workflow integration (Kromrey et al., 2024).

The implications for early cancer detection are direct. Tools that generate risk scores, heatmaps, or confidence indicators need careful attention to how those outputs are surfaced to clinicians. Recurrent recommendations in this literature include mandatory training in AI tool handling during residency, embedding AI in a single clinical platform to minimise ergonomic friction, dynamic confidence information that supports trust calibration, and periodic audits of human–AI agreement and disagreement at the level of the individual reader (Dratsch et al., 2023; Kromrey et al., 2024).

3.6. Regulatory frameworks: FDA and the EU AI Act

Few areas of medical regulation have moved as quickly as AI-enabled medical devices since 2021. In the United States, the FDA had authorised 691 AI/ML-enabled medical devices by October 2023, with annual approvals rising sharply since 2018 (Joshi et al., 2024); cumulative authorisations have continued to climb since, passing the one-thousand mark by late 2024 according to the FDA's publicly maintained list of authorised AI/ML-enabled medical devices (United States Food and Drug Administration, 2025). The majority are radiology-focused and approved through the 510(k) pathway, although a growing minority of breakthrough devices proceed through De Novo or Premarket Approval routes. Cancer-relevant applications—particularly mammography, low-dose CT, and dermatology—account for a substantial share of recent approvals.

On December 4, 2024, the FDA finalised its guidance on Predetermined Change Control Plans (PCCPs) for AI-enabled device software functions. The mechanism allows manufacturers to pre-specify and validate future software updates within authorised boundaries, without resubmitting marketing applications for each

change (United States Food and Drug Administration, 2024). The aim is to bring lifecycle thinking into a regulatory architecture historically built around static devices.

The European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689), which entered into force on August 1, 2024, complements the existing Medical Device Regulation by introducing horizontal, risk-based obligations for AI systems used as, or as safety components of, regulated medical devices, including diagnostic imaging software for cancer (van Kolfschooten & van Oirschot, 2024). High-risk medical AI systems must satisfy obligations on data governance, technical documentation, transparency, human oversight, accuracy, robustness, and cybersecurity, alongside MDR conformity assessment. Implementation across member states has so far been uneven, and the dual conformity-assessment burden has raised concerns about disproportionate costs for smaller developers and academic spin-offs (Vardas et al., 2025).

Both frameworks now anticipate the lifecycle nature of AI medical devices. One-time approval is no longer treated as sufficient for systems whose behaviour can shift with data and clinical context.

3.7. Synthesis of key AI applications in early cancer diagnostics

Table 1 brings together the principal AI applications discussed above, with representative evidence, reported performance benefits, and the main trust- and reliability-related concerns highlighted in the literature. Table 2 then summarises the two principal regulatory frameworks currently governing AI cancer diagnostic tools.

Table 1. Representative AI applications in early cancer diagnostics, with reported performance and principal reliability and trust concerns.

Modality / cancer type	Representative evidence	Reported performance benefit	Main reliability / trust concerns
Mammography / breast cancer	MASAI RCT (Lång et al., 2023; Hernström et al., 2025); prospective deployment studies (Lauritzen et al., 2022; Ng et al., 2023); external validation review (Anderson et al., 2022)	Up to ~29% increase in screen-detected cancers; ~44% reduction in screen-reading workload; non-inferior recall rate	Calibration mismatch across vendors and populations; under-representation of non-European cohorts; automation bias
Low-dose CT / lung cancer	Geppert et al., 2024 (Thorax systematic review)	+5–20% sensitivity for actionable nodules; faster reading; potential 150–750 additional cancers per million screened	3–8% specificity loss; large additional workload of CT surveillance; mostly retrospective evidence
Colonoscopy / colorectal cancer	Soleymanjahi et al., 2024 meta-analysis of 44 RCTs	~21% relative increase in ADR (RR 1.21); improved PDR and detection of sessile serrated lesions	Uncertain long-term impact on cancer incidence and mortality; high CADe false-positive rates; operator variability
Dermoscopic and clinical image analysis / melanoma and non-melanoma skin cancer	Esteva et al., 2017; Tschandl et al., 2019; Daneshjou et al., 2022 (fairness)	Dermatologist-level AUROC on curated datasets; improved triage for non-specialists	Severe performance drop on Fitzpatrick V–VI; sparse prospective real-world data; risk of widening inequities
Digital pathology / prostate and other solid tumours	PANDA challenge (Bulten et al., 2022); independent external validation (Perincheri et al., 2021)	Pathologist-level $\kappa \approx 0.86$ across cross-continental cohorts; high sensitivity for cancer detection	Limited prospective outcome data; scanner and stain variability; need for inherently interpretable models
Breast cancer risk prediction (mammography-derived)	Gastounioti et al., 2022	AUC ≈ 0.67–0.70 in external validation; outperforms classical clinical risk models in some settings	Modest absolute discrimination; limited ethnic diversity; unclear integration into screening intervals

ADR = adenoma detection rate; AUC / AUROC = area under the receiver operating characteristic curve; CADe = computer-aided detection; CT = computed tomography; PDR = polyp detection rate; RCT = randomised controlled trial.

Table 2. Principal regulatory frameworks for AI-enabled medical devices in early cancer diagnostics: comparison of the United States Food and Drug Administration (FDA) and the European Union AI Act / Medical Device Regulation.

Dimension	United States — FDA framework	European Union — AI Act + MDR
Core legal instrument	Food, Drug, and Cosmetic Act; FDA guidance documents including the 2024 final guidance on Predetermined Change Control Plans (PCCPs)	Regulation (EU) 2024/1689 (AI Act, in force from 1 August 2024) combined with Regulation (EU) 2017/745 (MDR)
Risk classification	Class I–III medical device classification; majority of cancer-relevant AI tools cleared via the 510(k) pathway, with De Novo and PMA routes for novel or higher-risk devices	Horizontal risk tiers (unacceptable, high, limited, minimal); diagnostic AI for cancer falls under high-risk, requiring additional obligations beyond the MDR conformity route
Adaptive / learning AI	PCCP allows manufacturers to pre-specify and validate future software changes within authorised boundaries without new marketing submissions (FDA, 2024)	AI Act explicitly addresses systems that learn after deployment; substantial modifications generally require new conformity assessment under MDR and AI Act provisions
Transparency obligations	Labelling and Summary of Safety and Effectiveness; reporting of intended use, performance, and population; remaining inconsistencies in subgroup data disclosure (Joshi et al., 2024)	Detailed technical documentation, instructions for use, transparency to deployers; specific obligations on data quality, traceability, and explanation of decisions affecting individuals
Human oversight	Embedded in device labelling and intended-use indications; clinician retains responsibility for final clinical decision	Explicit AI Act requirement of effective human oversight measures, with named responsible persons within deployer organisations
Post-market surveillance	MedWatch and Manufacturer and User Facility Device Experience (MAUDE); evolving expectations on real-world performance monitoring for AI devices	Continuous post-market surveillance and post-market clinical follow-up under the MDR; AI Act adds serious incident reporting and conformity re-assessment after substantial change
Fairness and bias	Good Machine Learning Practice principles and FDA guidance encourage subgroup analyses; not always reflected in public approval summaries	AI Act mandates data governance to detect and mitigate bias; bias monitoring is a core obligation for high-risk systems
Geographic scope and enforcement	Federal enforcement via FDA; state-level licensure of healthcare delivery; class actions and civil litigation as additional accountability layers	EU-wide application with national competent authorities; significant administrative fines for non-compliance; market surveillance by national bodies

FDA = United States Food and Drug Administration; MDR = Medical Device Regulation; PCCP = Predetermined Change Control Plan; PMA = Premarket Approval. Synthesised from Joshi et al. (2024), United States Food and Drug Administration (2024), van Kolfschooten and van Oirschot (2024), and Vardas et al. (2025).

4. Discussion

4.1. Trust as a sociotechnical, not purely technical, property

The evidence reviewed here supports a now-familiar conclusion: the deployability of AI in early cancer diagnostics depends less on incremental gains in headline accuracy than on the broader sociotechnical conditions of trust and reliability. Prospective trials such as MASAI demonstrate that, in mature mammography programmes with well-trained radiologists, double reading combined with carefully integrated AI can match or exceed standard practice while easing workload (Lång et al., 2023; Hernström et al., 2025). The same algorithms, transposed to populations or scanners outside their calibration distribution, may behave very differently (Schurz et al., 2025). Trustworthiness cannot be summarised by a single AUC.

From a social-science perspective, trust in AI for early cancer diagnostics is best framed as a relational property of a system that includes the model, its developers, the institutions deploying it, the regulators overseeing it, and the clinicians and patients interacting with it. Patients typically anchor their trust in the physician and the wider care pathway, with AI judged in relation to that anchor (Hassan et al., 2024). The implication for developers is that performance evidence, clinician presence, disclosure of training data composition, and visible governance arrangements are stronger determinants of acceptance than the mere deployment of AI itself.

4.2. Reliability as a lifecycle property

Reliability is also a lifecycle, rather than a release-time, property. The VAIB simulation by Schurz and colleagues makes the point concretely: shifts in calibration population can move cancer detection by tens of percentage points without any change to the underlying algorithm (Schurz et al., 2025). The external validation literature for mammography AI reinforces this picture, showing that most published evaluations are conducted in narrow cohorts that do not reflect the heterogeneity of real-world screening (Anderson et al., 2022). Mandatory prospective external validation, sex- and age-stratified performance reporting, and real-world monitoring of AI cancer diagnostic systems should therefore become standard expectations rather than aspirational extras.

A lifecycle view also requires the systematic adoption of contemporary reporting standards. The TRIPOD+AI statement, published in the BMJ in April 2024, supersedes earlier prediction-model reporting guidance with a unified 27-item checklist that explicitly covers fairness, reproducibility, and patient and public involvement (Collins et al., 2024). The CONSORT-AI extension for AI clinical trials and its companion SPIRIT-AI protocol guidance describe how trials of AI interventions should be designed and reported (Liu et al., 2020; Cruz Rivera et al., 2020). The updated CLAIM 2024 checklist provides a focused tool for medical imaging studies (Park & Suh, 2024). Embedding these standards into peer review and regulatory submission would substantially improve comparability across AI cancer diagnostic studies and reduce the over-representation of proof-of-concept studies that currently characterises the literature.

4.3. Bias mitigation and equitable deployment

Bias mitigation deserves special attention in early cancer diagnostics because the populations most likely to benefit from improved screening efficiency are often those least represented in training data. Three categories of intervention emerge from the literature. Dataset construction should be inclusive by design, with deliberate over-sampling of under-represented groups and explicit documentation of population composition, as exemplified by the Diverse Dermatology Images dataset (Daneshjou et al., 2022). Fairness should be evaluated using stratified performance metrics across protected attributes, ideally pre-specified rather than tested post hoc, and reported transparently in regulatory and peer-reviewed documents (Drukker et al., 2023). Post-market surveillance should include continuous monitoring of subgroup performance, in line with the EU AI Act's data governance obligations and the FDA's evolving expectations on real-world performance (van Kolfschooten & van Oirschot, 2024; Vardas et al., 2025).

4.4. From explainability to clinically meaningful transparency

The XAI literature reviewed here suggests that explainability is necessary but insufficient for trust. Tools such as SHAP, LIME, and saliency-based methods provide useful local rationales, especially for tree-based ensembles and certain convolutional architectures (Ghasemi et al., 2024). They do not by themselves guarantee that a clinician can judge whether a tool is appropriate for the patient in front of them. Closing that gap calls for more than technical advances. Model cards, audit trails, and standardised product labels that communicate intended population, performance limits, subgroup performance, and uncertainty are equally important pieces of the transparency puzzle.

4.5. Human–AI integration and clinician training

Evidence on automation bias has direct implications for early cancer diagnostics, where false positives generate downstream harms (biopsies, anxiety, costs) and false negatives can be lethal. In Dratsch and colleagues' multi-reader study, even very experienced radiologists were significantly affected by incorrect AI BI-RADS suggestions, although less so than their less experienced colleagues (Dratsch et al., 2023). The chest X-ray data of Kromrey and colleagues add a complementary lesson: high AI sensitivity in a real-world clinical environment can come with large numbers of unconfirmed findings that demand careful triage (Kromrey et al., 2024). Healthcare organisations adopting AI should therefore plan for structured training in critical appraisal of AI outputs, explicit protocols for handling AI–reader discordance, and routine performance audits of human–AI agreement at the level of the individual reader. Such audits already have precedent in mammography and colonoscopy quality-assurance programmes; extending them to AI use is a natural next step.

4.6. Limitations of the present review

Several limitations should be acknowledged. The search strategy, although structured, was not systematic in the PRISMA sense, and the selection of evidence reflects the authors' disciplinary perspectives. The review focused on the most evidence-rich oncology applications—breast, lung, colorectal, skin, and prostate cancers—and devoted less attention to emerging areas such as pancreatic, ovarian, and hematologic malignancies, where AI evidence is growing fast but is currently sparser. Regulatory frameworks evolve quickly, and specific guidance documents may have been further revised by the time of publication. Most evidence on patient and clinician trust derives from high-income settings, which constrains inferences for low- and middle-income countries. Finally, the article is not a quantitative synthesis; pooled effect estimates were avoided in favour of a narrative integration across heterogeneous study designs.

5. Conclusions

AI has moved from laboratory benchmarks into prospective trials and clinical practice in early cancer diagnostics. Mammography, low-dose CT, colonoscopy, dermoscopy, and digital pathology now provide the strongest evidence base. Where AI tools are well-validated and carefully integrated, they can raise cancer detection, ease reader workload, and support less experienced clinicians. Headline performance metrics alone, however, cannot establish trustworthiness or reliability. Both depend on the diversity and quality of training data, the rigour and breadth of external validation, the transparency of model behaviour and labelling, the design of human–AI interaction, and the maturity of regulatory and post-market surveillance mechanisms.

Three priorities emerge for the next phase of development. Robust prospective and external validation in diverse populations, reported according to contemporary standards such as TRIPOD+AI and CONSORT-AI, should become the norm. Fairness audits, subgroup performance reporting, and lifecycle surveillance should be embedded in regulatory pathways under the FDA's PCCP framework and the EU AI Act, with explicit attention to populations most vulnerable to underdiagnosis. Clinicians, patients, and policymakers should be engaged early and continuously in the design and oversight of AI tools, recognising that trust is co-produced rather than unilaterally claimed.

Future research should prioritise prospective outcome studies linking AI-supported early diagnostics to cancer incidence, stage distribution, and mortality; comparative effectiveness studies of competing AI systems in identical clinical contexts; structured evaluations of human–AI interaction, including training and trust calibration; and equity-focused evaluations across low- and middle-income settings. Realising the potential of AI in early cancer diagnostics will require treating trust and reliability not as promotional claims but as engineering, regulatory, and ethical achievements maintained throughout the lifetime of each deployed system.

ACKNOWLEDGMENTS

No financial support was received for this work. The authors thank colleagues for helpful discussions on AI in clinical oncology and on the evolving regulatory environment in the European Union.

Conflict of interest: The authors declare no conflicts of interest.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions:
Conceptualization: Emil Pałyga.
Methodology: Emil Pałyga, Mateusz Kwiatkowski.
Validation: Sara Demkow, Aleksandra Cieślak, Katarzyna Marcinkowska.
Formal analysis: Sandra Bryg, Joanna Sowińska.
Investigation: Natalia Paluszkiewicz, Karolina Siemińska.
Resources: Emil Pałyga, Zofia Leżańska.
Data curation: Mateusz Kwiatkowski, Sandra Bryg.
Writing — original draft preparation: Emil Pałyga, Karolina Siemińska, Aleksandra Cieślak.
Writing — review and editing: Emil Pałyga, Natalia Paluszkiewicz.
Visualization: Mateusz Kwiatkowski, Zofia Leżańska.
Supervision: Sara Demkow, Katarzyna Marcinkowska.
Project administration: Emil Pałyga.

Ethics approval: Not applicable. This article is a narrative review of previously published, publicly available studies and does not involve new research with human or animal subjects.

REFERENCES

1. American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000
2. Anderson, A. W., Marinovich, M. L., Houssami, N., Lowry, K. P., Elmore, J. G., Buist, D. S. M., Hofvind, S., & Lee, C. I. (2022). Independent external validation of artificial intelligence algorithms for automated interpretation of screening mammography: A systematic review. *Journal of the American College of Radiology, 19*(2), 259–273. https://doi.org/10.1016/j.jacr.2021.11.008
3. Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA—a scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review, 4*, Article 5. https://doi.org/10.1186/s41073-019-0064-8
4. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 74*(3), 229–263. https://doi.org/10.3322/caac.21834
5. Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M. B., Evans, A. J., van der Kwast, T., Allan, R., Humphrey, P. A., Grönberg, H., Samaratunga, H., ... Litjens, G. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine, 28*(1), 154–163. https://doi.org/10.1038/s41591-021-01620-2
6. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, Article e078378. https://doi.org/10.1136/bmj-2023-078378
7. Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine, 26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7
8. Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. A. C., Okata-Karigane, U., Zou, J., & Chiou, A. S. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances, 8*(32), Article eabq6147. https://doi.org/10.1126/sciadv.abq6147
9. Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto Dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology, 307*(4), Article e222176. https://doi.org/10.1148/radiol.222176
10. Drukker, K., Chen, W., Gichoya, J. W., Gruszauskas, N. P., Kalpathy-Cramer, J., Koyejo, S., Myers, K. J., Sá, R. C., Sahiner, B., Whitney, H., Zhang, Z., & Giger, M. L. (2023). Toward fairness in artificial intelligence for medical image analysis: Identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging, 10*(6), Article 061104. https://doi.org/10.1117/1.JMI.10.6.061104
11. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

12. Gastounioti, A., Eriksson, M., Cohen, E. A., Mankowski, W., Pantalone, L., Ehsan, S., McCarthy, A. M., Kontos, D., Hall, P., & Conant, E. F. (2022). External validation of a mammography-derived AI-based risk model in a U.S. breast cancer screening cohort of White and Black women. *Cancers, 14*(19), Article 4803. https://doi.org/10.3390/cancers14194803
13. Geppert, J., Asgharzadeh, A., Brown, A., Stinton, C., Helm, E. J., Jayakody, S., Todkill, D., Gallacher, D., Ghiasvand, H., Patel, M., Auguste, P., Tsertsvadze, A., Chen, Y.-F., Grove, A., Shinkins, B., Clarke, A., & Taylor-Phillips, S. (2024). Software using artificial intelligence for nodule and cancer detection in CT lung cancer screening: Systematic review of test accuracy studies. *Thorax, 79*(11), 1040–1049. https://doi.org/10.1136/thorax-2024-221662
14. Ghasemi, A., Hashtarkhani, S., Schwartz, D. L., & Shaban-Nejad, A. (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. *Cancer Innovation, 3*(5), Article e136. https://doi.org/10.1002/cai2.136
15. Hassan, M., Kushniruk, A., & Borycki, E. (2024). Barriers to and facilitators of artificial intelligence adoption in health care: Scoping review. *JMIR Human Factors, 11*, Article e48633. https://doi.org/10.2196/48633
16. Hernström, V., Josefsson, V., Sartor, H., Schmidt, D., Larsson, A.-M., Hofvind, S., Andersson, I., Rosso, A., Hagberg, O., & Lång, K. (2025). Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): A randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. *The Lancet Digital Health, 7*(3), e175–e183. https://doi.org/10.1016/S2589-7500(24)00267-X
17. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer, 18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5
18. Joshi, G., Jain, A., Araveeti, S. R., Adhikari, S., Garg, H., & Bhandari, M. (2024). FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: An updated landscape. *Electronics, 13*(3), Article 498. https://doi.org/10.3390/electronics13030498
19. Kromrey, M.-L., Steiner, L., Schön, F., Gamain, J., Roller, C., & Malsch, C. (2024). Navigating the spectrum: Assessing the concordance of ML-based AI findings with radiology in chest X-rays in clinical settings. *Healthcare, 12*(22), Article 2225. https://doi.org/10.3390/healthcare12222225
20. Lång, K., Josefsson, V., Larsson, A.-M., Larsson, S., Högberg, C., Sartor, H., Hofvind, S., Andersson, I., & Rosso, A. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): A clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology, 24*(8), 936–944. https://doi.org/10.1016/S1470-2045(23)00298-X
21. Lauritzen, A. D., Rodríguez-Ruiz, A., von Euler-Chelpin, M. C., Lynge, E., Vejborg, I., Nielsen, M., Karssemeijer, N., & Lillholm, M. (2022). An artificial intelligence-based mammography screening protocol for breast cancer: Outcome and radiologist workload. *Radiology, 304*(1), 41–49. https://doi.org/10.1148/radiol.210948
22. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x
23. Ng, A. Y., Oberije, C. J. G., Ambrózay, É., Szabó, E., Serfőző, O., Karpati, E., Fox, G., Glocker, B., Morris, E. A., Forrai, G., & Kecskemethy, P. D. (2023). Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nature Medicine, 29*(12), 3044–3049. https://doi.org/10.1038/s41591-023-02625-9
24. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns, 2*(10), Article 100347. https://doi.org/10.1016/j.patter.2021.100347
25. Park, S. H., & Suh, C. H. (2024). Reporting guidelines for artificial intelligence studies in healthcare (for both conventional and large language models): What's new in 2024. *Korean Journal of Radiology, 25*(8), 687–690. https://doi.org/10.3348/kjr.2024.0598
26. Perincheri, S., Levi, A. W., Celli, R., Gershkovich, P., Rimm, D., Morrow, J. S., Rothrock, B., Raciti, P., Klimstra, D., & Sinard, J. (2021). An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Modern Pathology, 34*(8), 1588–1595. https://doi.org/10.1038/s41379-021-00794-x
27. Schurz, H., Solander, K., Åström, D., Cossío, F., Choi, T., Dustler, M., Lundström, C., Gustafsson, H., Zackrisson, S., & Strand, F. (2025). Simulating mismatch between calibration and target population in AI for mammography: The retrospective VAIB study. *npj Digital Medicine, 8*, Article 259. https://doi.org/10.1038/s41746-025-01623-0
28. Soleymanjahi, S., Huebner, J., Elmansy, L., Rajashekar, N., Lüdtke, N., Paracha, R., Thompson, R., Grimshaw, A. A., Foroutan, F., Sultan, S., & Shung, D. L. (2024). Artificial intelligence-assisted colonoscopy for polyp detection: A systematic review and meta-analysis. *Annals of Internal Medicine, 177*(12), 1652–1663. https://doi.org/10.7326/ANNALS-24-00981
29. Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M. A., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., ... Kittler, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology, 20*(7), 938–947. https://doi.org/10.1016/S1470-2045(19)30333-X

30. United States Food and Drug Administration. (2024). *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and Food and Drug Administration staff*. U.S. Department of Health and Human Services. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence
31. United States Food and Drug Administration. (2025). *Artificial intelligence and machine learning (AI/ML)-enabled medical devices*. U.S. Department of Health and Human Services. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
32. van der Laak, J., Litjens, G., & Ciompi, F. (2021). Deep learning in histopathology: The path to the clinic. *Nature Medicine, 27*(5), 775–784. https://doi.org/10.1038/s41591-021-01343-4
33. van Kolfschooten, H., & van Oirschot, J. (2024). The EU Artificial Intelligence Act (2024): Implications for healthcare. *Health Policy, 149*, Article 105152. https://doi.org/10.1016/j.healthpol.2024.105152
34. Vardas, E. P., Marketou, M., & Vardas, P. E. (2025). Medicine, healthcare and the AI Act: Gaps, challenges and future implications. *European Heart Journal—Digital Health, 6*(4), 833–839. https://doi.org/10.1093/ehjdh/ztaf041