CHEARS Chronical
CHEARS

Volume 13 · Issue 4 · Oct–Dec 2026

TECHNOLOGY

Artificial Intelligence in Healthcare: Transforming Disease Diagnosis, Clinical Decision-Making, and Patient Outcomes

Haruka Takahashi*, Ateendra Jha

* *Correspondence: Haruka Takahashi*

Received: 30 October 2026 **Published:** 31 October 2026 **Volume 13, Issue 4** (Oct–Dec 2026)
Article no.: e225 **Open access** · free to read

ABSTRACT

Artificial intelligence (AI), particularly deep learning, has moved from research into regulated clinical use. By December 2025, the US Food and Drug Administration (FDA) had authorized 1,430 AI/ML-enabled medical devices, and radiology accounted for 77.5% of authorizations in 2023–2025.[1] Objective: To synthesize evidence on AI in disease diagnosis, clinical decision-making and patient outcomes, and to appraise the quality of that evidence. Findings: A meta-analysis of imaging studies found pooled sensitivity of 87.0% for deep learning versus 86.4% for health-care professionals, although few studies used external validation.[2] In the MASAI randomized screening trial (about 106,000 women), AI-supported mammography reading was associated with 12% fewer interval cancers.[3] In contrast, an independent validation of a widely deployed proprietary sepsis model reported an area under the curve (AUC) of 0.63 and missed 67% of sepsis cases.? A commercial risk algorithm under-identified Black patients for extra care (17.7% of those flagged, versus 46.5% if corrected),? and an observational study reported a 6.0-percentage-point fall in unassisted adenoma detection after AI exposure.? Conclusions: Evidence supports AI as an assistive tool for well-defined tasks. Evidence of improved patient outcomes remains limited, uneven and often observational. Prospective validation, bias auditing, post-deployment monitoring and clinician-skill safeguards are prerequisites for responsible scale-up.

Keywords: *artificial intelligence, deep learning, clinical decision support, medical imaging, diagnostic accuracy, algorithmic bias, sepsis prediction, large language models*

1. Introduction

Artificial intelligence describes computational systems that perform tasks ordinarily requiring human judgement. Most clinically deployed systems use machine learning (ML), and particularly deep learning (DL), in which multilayer neural networks learn patterns directly from images, waveforms or electronic health record (EHR) data. Large language models (LLMs) have more recently extended AI to free-text tasks. Regulatory activity illustrates the pace of change: the FDA authorized a mean of 2.0 AI/ML-enabled devices per year between 1995 and 2015, compared with a mean of 264 per year between 2023 and 2025, including 331 in 2025 alone.[1]

The World Health Organization (WHO) recognizes that AI can improve the speed and accuracy of diagnosis and screening, while cautioning against overestimating its benefits.? That caution reflects a persistent gap between regulatory authorization and clinical evidence. In an analysis of 130 FDA-approved AI devices (2015–2020), 126 were evaluated only retrospectively, and none of the 54 high-risk devices had undergone prospective evaluation.? Whether AI improves outcomes that matter to patients, rather than benchmark accuracy, is therefore the central question.

Objective and scope. This narrative review examines (i) the regulatory landscape, (ii) diagnostic performance, (iii) clinical decision support, and (iv) evidence on patient outcomes, followed by a critical analysis of validation, bias and safety. Sources were identified through targeted searches of peer-reviewed literature and official sources in October 2026, prioritizing systematic reviews, randomized trials, prospective multicentre studies and

independent external validations. This is not a formal systematic review, so selection bias is possible. Evidence is labelled as established, emerging or contested throughout.

2. The Regulatory Landscape for Clinical AI

2.1 Growth and specialty concentration

Between September 1995 and December 2025, the FDA authorized 1,430 AI/ML-enabled devices from 576 manufacturers. Of these, 96.2% (1,376/1,430) were cleared via the 510(k) pathway, 2.6% via De Novo and 1.2% via premarket approval.[1] Radiology has led in every era, but its share has begun to fall (Table 1).

Table 1. FDA-authorized AI/ML-enabled devices by era, 1995–2025 (N = 1,430)

Era (years)	Devices, n	Share of all devices (%)	Radiology share of era (%)	Specialty concentration (HHI*)
1 (1995–2015)	42	2.9	35.7	0.181
2 (2016–2019)	190	13.3	62.1	0.427
3 (2020–2022)	406	28.4	85.5	0.738
4 (2023–2025)	792	55.4	77.5	0.612

*Herfindahl-Hirschman Index (range 0–1; higher = more concentrated). Source: Lee and Youn, 2026.[1]

Suggested Figure 1: A stacked-bar chart of specialty share by era (radiology, cardiovascular, neurology, other) using the Table 1 data, captioned “Diversification of FDA-authorized clinical AI beyond radiology, 1995–2025.”

Among the 629 radiology devices whose imaging modality could be identified (of 1,094 radiology devices), computed tomography (n = 200), ultrasound (n = 124) and magnetic resonance imaging (n = 109) were most common.[1] Outside radiology, neurology authorizations rose from 6 (2020–2022) to 45 (2023–2025), and cardiovascular devices formed the largest non-radiology category (136 devices).[1]

2.2 Interpreting the numbers

The FDA registry is not exhaustive, counts authorizations rather than unique products, and does not capture clinical adoption or real-world impact; its authors stress that authorization is not adoption.[1] Other trackers report slightly different totals because of update timing and counting rules, so the figures above are a snapshot through December 2025.

3. AI in Disease Diagnosis

3.1 Medical imaging: pooled evidence

Liu and colleagues searched studies published between January 2012 and June 2019 comparing deep learning with health-care professionals in imaging-based diagnosis. Of 82 studies in the review, 25 performed out-of-sample external validation and 14 compared algorithms and clinicians on the same sample.[2] In those 14 studies, pooled sensitivity was 87.0% (95% CI 83.0–90.2) for deep learning and 86.4% (79.9–91.0) for clinicians; pooled specificity was 92.5% (85.1–96.4) and 90.5% (80.6–95.7), respectively.[2] The overlapping confidence intervals support a conclusion of broad equivalence, not superiority. The authors also noted poor reporting, and only four studies gave clinicians the additional clinical information they would normally use.[2] *Established:* DL can match clinicians on selected imaging tasks under experimental conditions. *Not established:* that this holds across populations and in routine workflow.

3.2 Autonomous screening for diabetic retinopathy

In a pivotal trial, Abràmoff and colleagues enrolled 900 adults with diabetes and no prior history of diabetic retinopathy at 10 US primary care sites, comparing an autonomous AI system with Wisconsin Fundus Photograph Reading Center grading.? Sensitivity was 87.2% (95% CI 81.8–91.2%), specificity 90.7% (88.3–92.7%) and imageability 96.1% (94.6–97.3%), exceeding prespecified endpoints (>85% and >82.5%).?

The FDA subsequently authorized it as the first autonomous AI diagnostic system in any field of medicine.? The significance is access: specialty-level screening can be delivered in primary care. The trial, however, measured diagnostic accuracy, not reduction in vision loss.

3.3 Population breast-cancer screening

The Mammography Screening with Artificial Intelligence (MASAI) trial, conducted in Sweden, randomized screening examinations 1:1 to AI-supported reading or standard double reading, with AI used to triage examinations to single or double reading.[1]? Final results, published in *The Lancet* in 2026, showed that AI-supported screening was non-inferior for the primary endpoint, interval cancer rate, with consistently more favourable outcomes.[3] According to the investigators, there were 12% fewer interval cancers (82 versus 93), and sensitivity increased without more false positives.[1]? Earlier trial reports had shown 29% more cancers detected and a 44% reduction in screen-reading workload.[1]? The authors state that the results do not support replacing radiologists, since at least one human reader remained part of the workflow.[3] Limitations include a single country, one AI system and a surrogate endpoint; breast-cancer mortality was not the trial's primary outcome.

4. AI in Clinical Decision-Making

4.1 Predictive early-warning systems: the sepsis example

Decision-support models embedded in EHRs show both the promise and the hazards of deployment. Wong and colleagues externally validated the Epic Sepsis Model, a proprietary tool implemented at hundreds of US hospitals, in 38,455 hospitalizations (27,697 patients) at one academic centre between December 2018 and October 2019.? Sepsis occurred in 2,552 patients (7%). The model's AUC was 0.63, below the 0.76–0.83 reported by its developer; it generated alerts for 18% of all hospitalizations yet failed to identify 67% of patients with sepsis, and it flagged only 183 patients (7% of those with sepsis) who had not received timely antibiotics.? The study demonstrated alert fatigue and poor sensitivity, but it was retrospective and single-centre, so generalizability is uncertain.

A contrasting finding comes from the Targeted Real-time Early Warning System (TREWS). In a prospective multisite cohort across five hospitals, 590,736 patients were monitored, and 6,877 had sepsis identified by the alert before antibiotics.[11] When a clinician confirmed the alert within 3 hours, adjusted in-hospital mortality was lower, with an absolute reduction of 3.3% (95% CI 1.7–5.1) and a relative reduction of 18.7% (95% CI 9.4–27.0).[11] Because this was an observational comparison of timely versus untimely clinician response, it shows association, not causation: clinicians who respond quickly may differ systematically from those who do not.

4.2 Large language models

LLMs are rapidly entering clinical documentation and question-answering, but evaluation lags behind. A systematic review of 519 studies published between January 2022 and February 2024 found that only 5% used real patient care data.[12] Medical-knowledge assessments such as licensing-examination questions made up 44.5% of tasks and diagnosis 19.5%; accuracy was the primary metric in 95.4% of studies, while fairness, bias and toxicity (15.8%) and deployment considerations (4.6%) were rarely measured.[12] Performance on examinations therefore cannot be assumed to predict safety or benefit in care. This evidence is *emerging*, and the field has advanced since the review's search date.

5. Effects on Patient Outcomes

Evidence of benefit can be ordered by how close it lies to what patients experience: diagnostic accuracy, then process measures (workload, time to treatment), then surrogate clinical endpoints (interval cancers), and finally patient outcomes (mortality, vision loss, quality of life). Most AI evidence sits in the first two tiers.

- **Accuracy:** pooled imaging studies[2] and the diabetic retinopathy pivotal trial.?
- **Process:** a 44% reduction in screen-reading workload in MASAI;[1]? earlier antibiotic action after TREWS alerts.[11]

- **Surrogate endpoints:** 12% fewer interval cancers in a randomized design.[3,1]?
- **Patient outcomes:** a mortality association in an observational sepsis cohort,[11] but no randomized evidence of mortality benefit in the sources reviewed here.

Suggested Figure 2: An "evidence ladder" diagram with the four tiers above and the studies discussed placed at their highest demonstrated tier.

6. Challenges, Risks and Limitations

6.1 The validation gap

Beyond the retrospective-only pattern among approved devices,? Wu and colleagues found that 93 of 130 devices had no publicly reported multi-site assessment, and among the 71 devices reporting sample size, the median evaluation sample was 300 patients.? Similarly, the Liu meta-analysis found few externally validated studies.[2] Single-site validation limits conclusions about bias and robustness across equipment and populations.

6.2 Algorithmic bias

Obermeyer and colleagues audited a commercial algorithm that, by industry estimates, is used to assess roughly 200 million people annually in the United States.? Because it predicted health-care costs rather than illness, and less money is spent on Black patients with equal need, Black patients at a given risk score were considerably sicker than White patients. Correcting the disparity would raise the share of Black patients receiving additional help from 17.7% to 46.5%.? The algorithm did not use race as an input, so bias arose from the choice of outcome label, not from an explicit variable.

6.3 Automation and deskilling

In a multicentre observational study of four endoscopy centres in Poland, the adenoma detection rate for colonoscopies performed without AI fell from 28.4% (226/795) before AI exposure to 22.4% (145/648) afterwards, an absolute difference of ?6.0 percentage points (95% CI ?10.5 to ?1.6).? Detection with AI assistance was 25.3%.? The design was observational, so the findings are hypothesis-generating; they nevertheless raise the possibility that continuous AI exposure may erode unassisted skills and may also affect the comparator arms of earlier trials.? This is an *emerging, contested* signal requiring replication.

6.4 Governance and ethics

The WHO guidance identifies six principles: protecting human autonomy; promoting well-being, safety and the public interest; ensuring transparency, explainability and intelligibility; fostering responsibility and accountability; ensuring inclusiveness and equity; and promoting AI that is responsive and sustainable.? These principles map directly onto the failures described above: opaque proprietary performance claims,? inequitable outcomes? and weak evaluation standards.?,[12]

7. Comparative Summary of Key Evidence

Table 2. Selected studies on AI performance, outcomes and limitations

Study (ref.)	Design and setting	Key quantitative result	Principal limitation
Liu et al., 2019[2]	Systematic review and meta-analysis; 82 studies (25 externally validated; 14 same-sample comparisons)	Pooled sensitivity 87.0% vs 86.4%; specificity 92.5% vs 90.5% (DL vs clinicians)	Few external validations; poor reporting
Abràmoff et al., 2018?	Prospective pivotal trial; 900 adults; 10 US primary care sites	Sensitivity 87.2%; specificity 90.7%; imageability 96.1%	Accuracy only; no outcome data
MASAI, 2026[3,1]?	Randomized trial; about 106,000 women; Sweden	12% fewer interval cancers (82 vs 93); earlier reports: 29% more cancers detected, 44% lower workload	One country and AI system; mortality not primary endpoint

Study (ref.)	Design and setting	Key quantitative result	Principal limitation
Wong et al., 2021?	Retrospective external validation; 38,455 hospitalizations; one US centre	AUC 0.63; 67% of sepsis cases not identified; alerts for 18% of hospitalizations	Single centre; retrospective
Adams et al., 2022[11]	Prospective multisite cohort; 590,736 monitored; 6,877 analysed	Adjusted absolute mortality reduction 3.3% (relative 18.7%) with timely alert confirmation	Observational; association only
Budzy? et al., 2025?	Retrospective observational; 4 centres, Poland	Unassisted detection 28.4% to 22.4% (?6.0 points)	Observational; short windows
Obermeyer et al., 2019?	Algorithm audit; large US academic medical centre	Share of Black patients receiving extra care: 17.7% vs 46.5% if corrected	One algorithm; cost-proxy specific
Bedi et al., 2025[12]	Systematic review; 519 LLM studies	5% used real patient data; 95.4% used accuracy as primary metric	Search ended February 2024
Wu et al., 2021?	Analysis of 130 FDA-approved devices (2015–2020)	126 retrospective only; 0 of 54 high-risk devices evaluated prospectively	Covers pre-2021 approvals

Methodological note: the sensitivity and specificity values in rows 1–2 derive from different reference standards, populations and tasks and should not be compared directly with one another.

8. Discussion and Critical Analysis

Convergent findings. Across imaging tasks, AI performs at approximately the level of clinicians under controlled conditions.[2],? The more consequential benefits in the best-designed evidence concern workflow and access: reduced reading workload in screening[1]? and specialty-level screening delivered in primary care.?

Apparent contradictions. A positive randomized trial in mammography[3] and a poorly performing proprietary sepsis model? are not contradictory. In our interpretation, they illustrate that performance is task-, site- and implementation-specific. MASAI evaluated a defined workflow prospectively within a screening programme, whereas the sepsis model was deployed widely before independent validation. Even within sepsis, TREWS and the Epic model produced very different results in different studies, which differ in design, site and endpoint and so cannot be compared directly.?,[11]

Causation and bias in the evidence. The strongest mortality signal (TREWS) and the most concerning safety signal (deskilling) both come from observational data.?,[11] Several studies were authored or co-authored by individuals with industry ties, as the authors of the Liu meta-analysis and the TREWS study disclosed,[2],[11] which does not invalidate the findings but strengthens the case for independent replication.

Equity. The Obermeyer finding shows that high predictive accuracy for the wrong target can embed inequity.? Standard accuracy metrics did not reveal it; testing whether equal scores correspond to equal need did.

Limitations of this review. The review is narrative, uses a limited set of sources, and relies on published reports; some quantitative details for the MASAI trial were taken from the investigators' institutional summary[1]? alongside the primary publication[3] and should be confirmed against the full text.

9. Future Perspectives and Research Directions

- **Prospective, outcome-focused trials.** Randomized or pragmatic designs measuring mortality, morbidity and patient-reported outcomes, not only accuracy, are the most pressing need.?,[11]
- **Post-market surveillance.** Monitoring for performance drift and unintended bias is needed beyond premarket studies;? the 2024 Predetermined Change Control Plan framework for adaptive algorithms has not yet been empirically evaluated.[1]

- **Broader specialties.** Authorizations are diversifying into neurology, gastroenterology-urology, anesthesiology and pathology,[1] so validation infrastructure must follow.
- **LLM evaluation using real clinical data,** with bias, calibration and deployment measures, as called for in the systematic review.[12]
- **Skill-preservation research.** Replication of deskilling findings and trials of training and monitoring strategies.?
- **Equity-by-design.** Auditing of outcome labels and subgroup performance before deployment.?,?

10. Conclusion

AI is now an established part of regulated medical technology, with strongest evidence in imaging-based diagnosis and screening, where it matches clinicians in controlled studies and, in one large randomized trial, reduced workload while improving a surrogate cancer-screening endpoint. In clinical decision support, results are heterogeneous: some tools have performed poorly under independent validation, while others show promising but observational associations with better outcomes. Evidence of benefit on patient outcomes remains limited, and risks such as algorithmic bias, weak validation standards and potential clinician deskilling are documented. A balanced view is that AI should currently be regarded as a powerful assistive technology whose value depends on rigorous external validation, equitable design, transparent reporting and ongoing monitoring within human-supervised workflows.

Competing interests

The authors declare that they have no competing interests.

References

1. Lee YS, Youn BY. Clinical specialty expansion of AI-enabled and machine learning–enabled medical devices authorized by the US Food and Drug Administration from 1995 to 2025: longitudinal content analysis. J Med Internet Res. 2026;28:e103040. doi: 10.2196/103040.
2. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1:e271–97. Available from: https://doaj.org/article/c73287f0c1f24d9abbb2cb81ce6add3e
3. Lång K, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet. 2026;407:505–14. doi: 10.1016/S0140-6736(25)02464-X.
4. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065–70. doi: 10.1001/jamainternmed.2021.2626.
5. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. doi: 10.1126/science.aax2342.
6. Budzy? K, Roma?czyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. Lancet Gastroenterol Hepatol. 2025;10(10):896–903. doi: 10.1016/S2468-1253(25)00133-5.
7. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. Available from: https://who.int/publications/i/item/9789240029200
8. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27(4):582–4. doi: 10.1038/s41591-021-01312-x.
9. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39. doi: 10.1038/s41746-018-0040-6.
10. Lund University. AI support in breast cancer screening: fewer missed cancer cases [Internet]. Lund: Lund University; 2026 [cited 2026 Oct 30]. Available from: https://www.lunduniversity.lu.se/article/ai-support-breast-cancer-screening-fewer-missed-cancer-cases
11. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med.

2022;28(7):1455–60. doi: 10.1038/s41591-022-01894-0.

12. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025;333(4):319–28. doi: 10.1001/jama.2024.21700.

How to cite this article (Vancouver)
Takahashi H, Jha A. Artificial Intelligence in Healthcare: Transforming Disease Diagnosis, Clinical Decision-Making, and Patient Outcomes. CHEARS Chronical. 2026;13(4):e225. Available from: https://chears.in/journal/volume-13/issue-4/artificial-intelligence-in-healthcare-transforming-disease-diagnosis-clinical-de-225
Published in CHEARS Chronical.
https://chears.in/journal/volume-13/issue-4/artificial-intelligence-in-healthcare-transforming-disease-diagnosis-clinical-de-225