

State of clinical AI in 2026

John Emmett Worth,¹ Anastasia Perez ,² David JH Wu,³ Peter Brodeur,⁴ Emily Tat,⁵ Arjun Manrai,⁶ David Wu,⁷ Chase Walton,^{8,9} Liam G McCoy,^{4,10,11} Priyank Jain,¹² Rebecca Handler,² Jason Hom,¹³ Laura Zwaan,¹⁴ Vishnu Ravi,² Brian Han,² Kevin Schulman,² Kathleen Lacar,² Kameron Black,² Adrian D Haimovich,¹⁵ Adam Rodman,⁴ Ethan Goh ,^{16,17} Jonathan H Chen^{16,17,18}

To cite: Worth JE, Perez A, Wu DJH, *et al.* State of clinical AI in 2026. *BMJ Digit Health* 2026;**2**:e000113. doi:10.1136/bmjdh-2026-000113

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjdh-2026-000113>).

JEW, AP, DJW and PB contributed equally.

AR, EG and JHC are joint senior authors.

Received 9 May 2026
Accepted 31 August 2026



© Author(s) (or their employer(s)) 2026. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

Correspondence to
Dr Ethan Goh;
ethangoh@stanford.edu

ABSTRACT

Clinical artificial intelligence (AI) has advanced rapidly, with frontier large language models now matching or exceeding physician performance on simulated diagnostic reasoning and clinical decision-support tasks. Yet adoption has outpaced the evidence base: fewer than 5% of cleared U.S. Food and Drug Administration (FDA) AI/machine learning (ML) devices have undergone peer-reviewed evaluation and prospective randomised trials remain scarce. This narrative review synthesises evidence from ARISE's State of Clinical AI Report 2026 across six domains: model performance and human benchmarking, evaluation methodology and benchmark validity, foundational methods including multimodal and multiagent systems, clinical workflow integration, patient-facing applications and applied domain-specific AI. We conducted a targeted literature search of PubMed, medRxiv and arXiv supplemented by expert nomination, including empirical studies of AI in clinical contexts. Across domains, frontier models demonstrate strong clinical reasoning capabilities but exhibit persistent failures in uncertainty calibration, metacognition and robustness to distributional shift. Standard benchmarks have saturated, prompting development of multidimensional evaluation frameworks that assess safety, real-world workflows and agentic capabilities. New trends in foundational methods include converting medical data into tokens and developing multiagent, multimodal systems. In clinical workflows, early prospective trials in clinical decision support and diagnostic imaging show promising results, though human-AI collaboration remains suboptimal and risks of automation bias and clinician deskilling are emerging. In tandem, patient-facing AI is progressing with more personalised health assistance; with this comes the risk of patient overtrust, raising the bar for guardrails. The most credible translational progress occurs on narrowly defined tasks with clear endpoints, while broader clinical autonomy awaits advances in evaluation and human-AI interaction.

INTRODUCTION

Clinical artificial intelligence (AI) has reached an inflection point. Over the past 2 years, AI systems have expanded across clinical reasoning, administrative workflows and patient-facing interaction.¹ Frontier models now achieve superhuman performance on

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence (AI) systems have demonstrated strong performance on medical licensing examinations and controlled clinical reasoning tasks, leading to rapid adoption across healthcare. However, prior reviews have typically focused on narrow subtopics, such as individual specialties or specific model architectures, without synthesising evidence across model capabilities, evaluation methodologies, human-AI interaction and patient safety factors.

WHAT THIS STUDY ADDS

⇒ This review synthesises evidence across six domains: model performance, benchmarks and evaluation, foundational methods, AI in clinical workflows, patient-facing AI and applied AI. Key themes and pitfalls are explored for each domain. Across domains, a consistent theme emerges: model capabilities have progressed faster than the research and implementation systems needed for safe deployment. For example, benchmarks are saturating without capturing real-world complexity, current human-AI collaboration underperforms model-only baselines and real-world prospective trials, while promising, remain limited in scope and generalisability.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ More research is needed in real-world clinical settings. These findings support directing research investment towards prospective evaluations embedded in authentic workflows, with greater emphasis on safety, equity, metacognitive calibration and human-AI collaboration alongside accuracy. They also support governance structures that require demographic reporting, bias assessment and ongoing post-deployment monitoring of performance and clinical outcomes for all clinical AI tools.

controlled reasoning tasks, yet exhibit persistent failures in uncertainty calibration (does expressed confidence match actual accuracy), metacognition (eg, 'recognizing the limits of one's own knowledge') and robustness under distributional shift (performance degradation when patient populations, care

settings or data formats differ from those seen during training).² The gap between benchmark performance and real-world reliability remains one of the field's defining tensions.

The volume of literature has grown accordingly, with systematic reviews spanning nearly all medical specialties and the rate of large language model (LLM) studies published increasing rapidly since ChatGPT's release.^{3–7} This expansion in volume, however, has not translated into depth of evidence. Adoption of AI-enabled medical devices and consumer health is outpacing the generation of prospective, real-world evidence validating their safety and effectiveness in clinical workflows.^{8–12} Regulatory oversight has similarly struggled to keep pace, with more than 1200 FDA-cleared AI/ML devices on the market and fewer than 5% having undergone rigorous peer-reviewed evaluation.¹³

This review is a narrative rendition of the AI Research and Science Evaluation (ARISE) network's State of Clinical AI Report 2026 to provide clinicians, researchers and policymakers with a critical appraisal across six domains: (1) model performance, (2) benchmarks and evaluation, (3) foundational methods, (4) AI in clinical workflows, (5) patient-facing AI and (6) applied AI (figure 1).

METHODOLOGY

The authors conducted a targeted literature search from October 2025 to January 2026 in PubMed, medRxiv, and arXiv, supplemented by expert nominations. The primary emphasis was on general-purpose and generative AI systems, including LLMs, multimodal foundation

models and agentic AI systems. Search terms included “large language models in medicine,” “AI in medicine,” “diagnostic reasoning,” “management reasoning,” “diagnostic error,” “clinical benchmarks” and “patient-facing AI.” Eligible studies required high-quality empirical evaluation of AI in clinical contexts with quantitative or qualitative outcomes (eg, diagnostic accuracy, bias, calibration, workflow and user performance). Purely technical papers without clinician-facing or patient-facing evaluation, editorials and non-clinical AI studies (eg, drug discovery, biotech) were excluded.

In addition, sources and recommendations from outside the scope of the targeted literature review were solicited from clinicians and AI researchers at academic institutions via email and social media. Using a Delphi process among the authors, these studies were reviewed and included if they were felt to be of high clinical impact. As such, inclusion decisions were resolved through discussion among author-reviewers to reach consensus on final study selection. This established a final cohort of 82 studies: 40 (49%) identified through database searches, 31 (38%) nominated by the author team on the basis of expert knowledge and 11 (13%) submitted by external clinicians and AI researchers.

After the final study list was generated, the authors conducted an additional Delphi process to establish core themes. This consisted of a structured consensus process across three meetings to identify, refine and organise the core themes emerging from the included literature. During these meetings, studies were reviewed collectively, candidate themes were proposed and iteratively revised

State of Clinical AI 2026 — Landscape Map

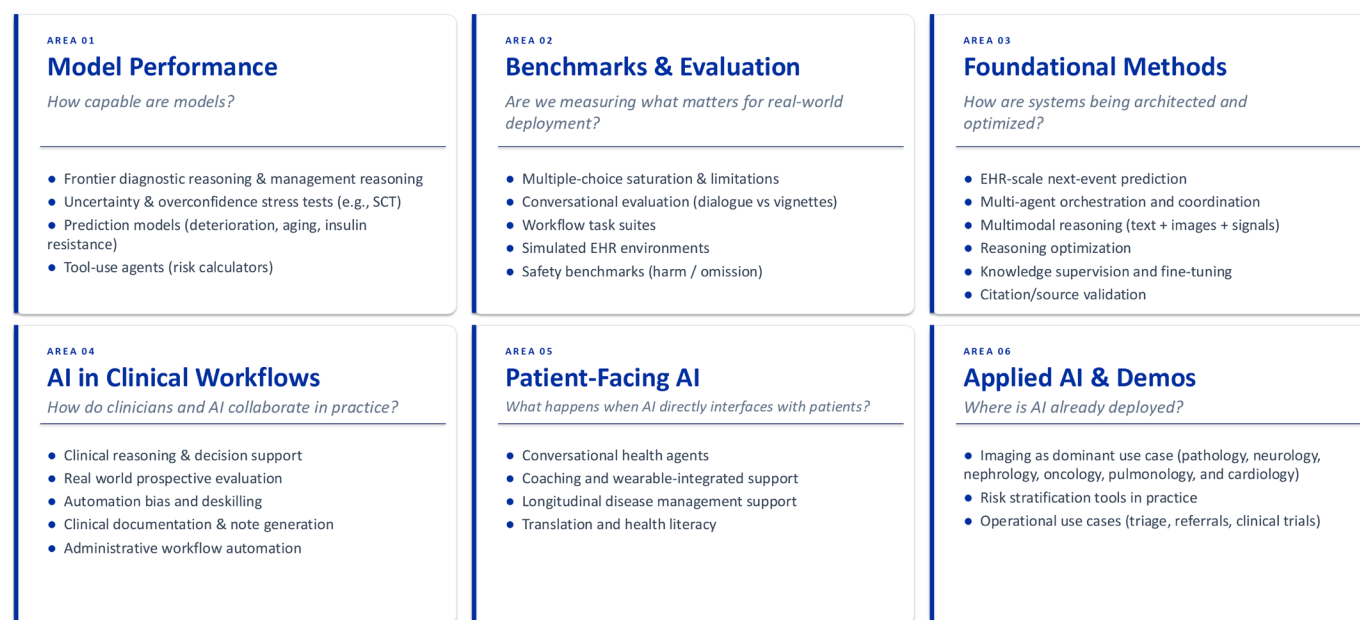


Figure 1 Summary of the included literature by area. Areas include model performance, benchmarks and evaluation, foundational methods, AI in clinical workflows, patient-facing AI and applied AI and demos. AI, artificial intelligence; EHR, electronic health record; SCT, script concordance testing.

and disagreements were resolved through group discussion until consensus was reached on the final thematic framework. Our six domains were derived through the Delphi process described above and reflect where the 2025–2026 empirical evidence was most concentrated; they are not intended as an exhaustive map of clinical AI. Several important areas, notably equity and algorithmic bias, regulation and governance, data privacy and security and implementation science, warrant dedicated review and fall outside the scope of this synthesis (online supplemental figure S1).

RESULTS: THE SIX DOMAINS OF CLINICAL AI

Model performance

In 2025, frontier LLMs have shown capabilities in controlled clinical reasoning tasks, including diagnostic reasoning and clinical decision support, that frequently match or exceed human physician performance. Following these general-purpose models, specialised prediction models can now provide clinicians with insights previously unavailable through traditional statistical methods. These studies typically evaluate the performance of models in different scenarios as a ‘one-time measure’, which is distinct from benchmarks, which are deliberately constructed to be reused for evaluation of emerging models.

Diagnostic reasoning and management

Across studies, newer models have shown marked improvements in diagnostic reasoning, reasoning presentation and disease management capabilities compared with earlier AI systems and physicians. For instance, one evaluation of frontier models on historically difficult *New England Journal of Medicine* (NEJM) clinicopathological conference (CPC) cases found OpenAI’s o3 model achieved 60% top-1 and 84% top-10 diagnostic accuracy and selected the correct next test 98% of the time.¹⁴ In another study using CPCs, OpenAI’s o1 model achieved 78% diagnostic accuracy and selected the next-best test 87% of the time.² Compared with human performance, a study using 302 CPCs reported 24% top-1 and 45% top-10 accuracy among physicians with unrestricted electronic search.¹⁴ Newer models now extend beyond diagnosis to support disease management by integrating reasoning and retrieval capabilities. For example, Google’s Articulate Medical Intelligence Explorer (AMIE) system was found to be non-inferior to 20 primary care physicians in management reasoning (treatment planning, longitudinal care), producing more precise plans and outperforming physicians on medication-reasoning questions.¹⁵

Scalable prediction and longitudinal clinical modelling

Clinical AI is increasingly enabling scalable prediction and longitudinal health monitoring, including assessments of risk (disease progression, clinical deterioration, adverse event, etc) and population health. Agentic systems, or AI models that autonomously plan and execute multistep tasks, have demonstrated clinical utility

in controlled settings. AgentMD, for example, enabled models to autonomously select from a library of over 2000 clinical calculators for a given patient, significantly outperforming general-purpose models in individual risk prediction accuracy on the 350-question RiskQA benchmark (87.7% vs 40.9%).¹⁶ Furthermore, the combination of deep learning models with continuous wearable data has been shown to augment clinical insight. In some studies/instances, models detected surgical patient deterioration up to 17 hours in advance and proactively identified metabolic precursors of insulin resistance.^{17 18} One study applying LLM-based frameworks to population-scale health data found that the predicted ‘biological age’ outperformed established ageing biomarkers like telomere length or epigenetic clocks.¹⁹

Persistent failure modes

Despite strong benchmark performance, a recurring limitation is the fragility of model performance in unfamiliar scenarios. When multiple-choice tests replaced the correct answers with ‘None of the other answers’, a study found a drop in performance, ranging from 9% to 38%, across all models.²⁰ This indicates that near-perfect accuracy on standard benchmarks may reflect statistical pattern-matching or shortcut behaviour, where models learn to associate structural or contextual cues of multiple-choice exam formatting with likely correct answers, rather than robust clinical logic. Moreover, in tests designed to reflect the dynamism of clinician reasoning when new information is introduced, models demonstrated systematic overconfidence and difficulty updating judgments under uncertainty.²¹ Even highly optimised systems show significant room for improvement in high-stakes tasks like medication reasoning, where accuracy varies between open-book and closed-book environments.¹⁵ Some note frontier models as displaying a ‘jagged’ intelligence: superhuman on some tasks yet failing at others humans find trivial, and prone to collapse when confronted with excessive complexity or out-of-distribution data. Overall, for clinicians, the practical implication is that trust should be calibrated to the specific task at hand rather than to a general impression of competence, since capability can vary sharply between superficially similar problems.

Benchmarks and evaluation

In contrast to model performance alone, benchmarks are standardised instruments deliberately constructed for repeated use, enabling consistent comparison of capability as new models emerge. Near-perfect scores on medical licensing examinations do not capture the ‘messiness’ of real-world scenarios; therefore, clinical AI evaluation has begun to shift toward more ecologically valid assessments of real-world capability (figure 2, table 1).

The saturation of knowledge-centric benchmarking

In response to benchmark saturation and the need for more trustworthy models, clinical AI evaluation has

transitioned from multiple-choice benchmarking towards more naturalistic assessments. The current landscape overemphasises knowledge-based performance while neglecting realistic deployment scenarios. A systematic review (n=519 studies from 2022 to 2024) found that LLM evaluations overwhelmingly focused on medical knowledge tasks, with 95% of studies relying on accuracy as the primary evaluation metric.²² The most common tasks in these studies were answering medical licensing examination questions (45%) and making clinical diagnoses (19%), with only 5% of studies incorporating real patient data.²² The saturation of knowledge-centric benchmarking and lack of task diversity have opened the door for multidimensional evaluation frameworks that seek to represent a holistic view of clinical practice. As benchmarks improve, we find that determining the 'right' answer in medicine is itself difficult: this can often require integrating biology, ethics and the values and priorities of individual patients; an option that appears 'wrong' on paper may be optimal for a particular patient given unrecorded context; and 'correct' answers are not stable as evidence evolves. Benchmarks that assume a single determinate answer therefore capture only part of clinical quality.

Broadening the horizon (multidimensional frameworks and models of clinical realities)

Benchmarks introduced in 2025 expanded the complexity of evaluation practices to mirror the complexity of actual clinical workflows. A key example of this is Medical Holistic Evaluation of Large Language Models (MedHELM), which developed five physician-validated taxonomies for frontier models across 121 distinct clinical tasks.²³ This broader evaluation approach builds on a shift towards more realistic, physician-verified assessments using synthetically generated multiturn health conversations, as evidenced in OpenAI's HealthBench and Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD) studies.^{24 25} Together, they found steady progress from GPT-3.5 (16%) to GPT-4 (60%) on complex health reasoning and excellent performance in documentation and patient communication. By

approximating realistic clinical scenarios, these multi-dimensional frameworks have advanced the evaluation paradigm beyond knowledge testing. Because these frameworks are harder and more realistic than licensing-exam benchmarks, they avoid saturation and create a meaningful spread between models of differing capabilities.

Gaps in safety, performance and the metacognitive frontier

Despite the increased sophistication of these assessments, the transition to dynamic utility has exposed critical shortcomings. MedHELM's findings highlight that the tasks most desired by practitioners to reduce burnout, such as administration and workflow support, remain the models' weakest areas.²³ Similarly, MedAgentBench used a virtual electronic health record (EHR) environment to test agentic capabilities across 300 patient-specific tasks, revealing that frontier models perform relatively poorly on multiturn, clinical tasks compared with their medical reasoning performance.²⁶ Findings suggest that Fast Healthcare Interoperability Resources Application Programming Interface (API) navigation, interpretation of users' prompt intent and long-term planning remain immature for live clinical deployment. Another recurring limitation is the disconnect between benchmark accuracy and clinical safety. The Numerous Options Harm Assessment for Risk in Medicine (NOHARM) framework revealed that in approximately 22% of primary care consultation cases, frontier models produced responses that could lead to clinical harm.²⁷ Specifically, models exhibited 'sins of omission', frequently failing to suggest critical diagnostic tests or treatments that a human physician would deem essential for safety.²⁷ This limitation is further highlighted by the CRAFT-MD benchmark, where models, which appeared expert-level on structured summaries, frequently missed key history details, and diagnostic accuracy dropped substantially when cases were converted to dialogue (eg, GPT-4 from 0.82 to 0.63 and to 0.49 without multiple-choice cues).²⁵ Finally, the consensus in 2025 reemphasises that models lack the ability to self-assess or recognise when they are wrong. The MetaMedQA benchmark revealed that LLMs

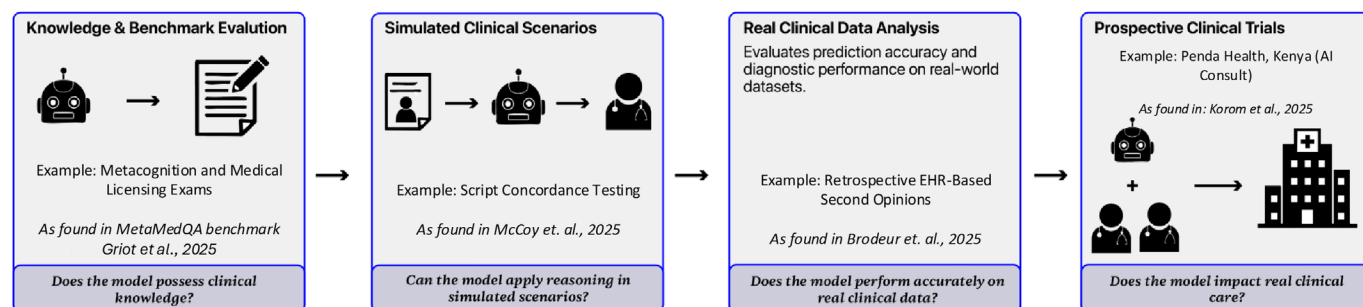


Figure 2 Clinical AI evaluation methodology spans across four levels with increasing amounts of real-world validity: benchmark testing of clinical knowledge, reasoning assessment through simulated vignettes, retrospective validation on real patient datasets and prospective clinical trials in live care settings. Each stage increases the amount of methodological rigour and therefore provides more information about safety and proximity to real clinical deployment. AI, artificial intelligence; EHR, electronic health record.

Table 1 Summary table of selected benchmarks from 2025

Study	Benchmark name	Country of origin	Purpose	Methods
Aali <i>et al</i> ⁷³	MedVAL-Bench	USA	Evaluate the factual correctness of medical text generation	A dataset of 840 physician-annotated outputs across six medical tasks collected from public medical datasets, with outputs annotated by physicians.
Arora <i>et al</i> ²⁴	HealthBench	USA	Evaluate LLM performance across realistic patient-doctor conversation tasks	5000 multiturn conversations between a model and an individual user or healthcare professional using conversation-specific rubrics created by physicians.
Bedi <i>et al</i> ²³	MedHELM	USA	Evaluate LLM performance for real clinical tasks across clinical decision support, clinical note generation, patient communication, medical research and administration	Developed a benchmark suite that includes 19 existing benchmarks, five reformulated benchmarks based on existing datasets and 13 new benchmarks. Used accuracy-based metrics and an 'LLM-jury' for scores.
Dai <i>et al</i> ⁷⁴	Clinical Large-Scale Integrative Multimodal Benchmark	USA	Evaluate models across multimodal clinical data in realistic patient scenarios	Comprised of 4.51 million patient samples across 2D imaging, 3D video, time series, graphs and multimodal data, and models testing against a ground truth comparison of expert-validated answers.
Diego <i>et al</i> ⁷⁵	MedPI	USA	Evaluate LLMs in simulated multiturn, patient-doctor conversations in primary care	Simulated data from 366 distinct 'AI patients' and 7097 patient-doctor conversations were scored against 105 dimensions and judged by LLM judges.
McCoy <i>et al</i> ²¹	Script concordance testing (SCT)	USA	Evaluate models' ability to produce consistent reasoning and answers, especially when new information is introduced	750 SCT questions across multiple specialties, testing the change in diagnosis or management depending on added data. Scored against expert-panel responses.
Mehandru <i>et al</i> ⁷⁶	ER-Reason	USA	Evaluate LLM-based clinical reasoning and decision-making in the emergency room (ER)	Data from 3984 patients, including clinical notes, progress notes, history and physical exams, consults, echo reports, imaging notes and ER provider documentation, with LLM results evaluated against clinician answers.
Nori <i>et al</i> ³¹	SDBench	USA	Evaluate iterative clinical diagnosis to benchmark diagnostic agents (human or AI) decision-making across multiple steps	Converted 304 NEJM-CPC cases into stepwise diagnostic encounters, with performance assessed by diagnostic accuracy and total cost
Wang <i>et al</i> ⁵	Clinical Safety-Effectiveness Dual-Track Benchmark	China	Evaluate LLMs against 30 metrics with a focus on safety and effectiveness	Synthesised 2069 clinical scenario questions across 26 specialties and scored responses on safety and effectiveness dimensions, combining into a total score.
Wu <i>et al</i> ²⁷	NOHARM	USA	Evaluate the frequency and severity of harm from AI-generated medical recommendations	Physician-validated benchmark of 100 real primary-care to specialist consultation cases, covering 10 specialties. Given the patient case description and physician question, the LLM must list a complete management plan in free text.
Zuo <i>et al</i> ⁷⁷	MedXpertQA	China	Evaluate expert-level medical knowledge and advanced reasoning	Included 4460 questions across 17 specialties. Evaluated LLMs in two subsets: text and multimodal.

This is a non-exhaustive list of benchmarks selected by expert review.

AI, artificial intelligence; 2D, two-dimensional; 3D, three-dimensional; LLM, large language model ; MedHELM, Medical Holistic Evaluation of Large Language Models; NEJM-CPC, New England Journal of Medicine clinicopathological conference; NOHARM, Numerous Options Harm Assessment for Risk in Medicine.

consistently fail to recognise their own knowledge limitations, providing highly confident answers even when the correct option is absent from a question.²⁸ This failure in recall suggests that models cannot yet be trusted to admit uncertainty, which is a prerequisite for safe clinical translation. Collectively, these findings demonstrate that better evaluation, specifically frameworks that prioritise safety, longitudinal workflows and metacognition, is the essential prerequisite for trustworthy clinical AI.

Automated evaluations, an aside

A substantial share of the benchmark and evaluation studies reviewed here rely on automated 'LLM-as-judge' scoring, including multiagent designs (eg, HealthBench, CRAFT-MD) in which language models simulate patient, clinician and adjudicator. This contrasts with scripted, ground-truth-anchored schemes such as *NOHARM* which uses correction algorithms where models are constrained in their adjudication allowing repeatable external verification. The reliability and potential bias of LLM judges is itself a growing area of interest and concern; for now, we suggest that results which depend on automated evaluation should be interpreted with corresponding caution.

Foundational methods

Advancements in foundational methods such as LLM/AI system architecture and setup suggest that more structured statistical grounding can expand the utility of clinical AI.

Prediction: the era of medical event foundation models

Foundation models such as Delphi-2M and Epic's Cosmos Medical Event Transformer (CoMET) can convert a patient's entire medical history into structured, sequentially readable units (a process called tokenisation), enabling these systems to predict future clinical events at the population scale. These systems are trained on hundreds of thousands to 100 million patients and have been shown to forecast future diagnoses and future clinical events across thousands of conditions.^{29 30} Delphi-2M, for instance, achieves an average area under the curve (AUC) of 0.76 in predicting the timing and occurrence of over 1000 diseases, with the unique capacity to simulate realistic lifetime health trajectories up to 20 years into the future. AUC alone is not enough to base real-world decision-making, which requires external validation and prospective studies. The model's developers note these limitations, cautioning against assuming the patterns reflect causation and stressing that any individual's prediction remains heavily influenced by chance. This caution extends to other models in the space. CoMET uses 115 billion clinical events to autoregressively predict subsequent diagnoses and readmissions, matching or exceeding the performance of task-specific models without traditional fine-tuning. By capturing the natural history of disease through these tokenised timelines, these models provide a pioneering level of support for anticipatory medicine, surfacing critical windows for

preventative intervention. However, as the clinical scope of these models expands, the field is beginning to move beyond single-model architectures towards the complex orchestration of specialised AI agents.

Multiagent systems: orchestration and interaction challenges

Emerging multiagent clinical AI systems suggest that the way models are organised and interact may matter as much as the capabilities of the individual models themselves. Microsoft's MAI Diagnostic Orchestrator demonstrates this principle by simulating a panel of five physicians that sequentially request diagnostic information until reaching high confidence, achieving about 80% accuracy on NEJM case presentations while reducing diagnostic costs by up to 70% compared with off-the-shelf o3.³¹ Critically, this performance arises not from a single superior model but from orchestration: agents collaboratively select the most informative next diagnostic test rather than exhaustively exploring all options. Multiagent conversation frameworks reinforce this finding, showing that diagnostic accuracy improves when rare-disease cases are evaluated through structured disagreement among several specialised agents and a supervising agent (78% vs 58% for standalone models).³² However, this approach has demonstrated shortcomings. Testing 26 multiagent configurations on 2400 real emergency department cases revealed an 'optimisation paradox', where the multiagent model that was assembled from the highest performing individual agents ('best-of-breed' system) actually underperformed. It hit 86% component-level accuracy but only 68% diagnostic accuracy, well below a less-optimised system's 77%.³³ This gap in performance has been attributed to breakdowns in the flow of information between agents, demonstrating that optimising individual agents does not guarantee end-to-end system performance.

Multimodal systems: statistical and clinical grounding

Multimodal clinical AI systems process multiple types of data, such as clinical notes, lab results and medical imaging, simultaneously. This system of integrating diverse medical data through a single model has been shown to improve diagnosis, prognosis and treatment selection in some studies. Systems like AMIE and Multimodal transformer with Unified maSKed modeling demonstrate the superiority of this approach, leveraging multimodal reasoning to outperform primary care physicians in diagnostic accuracy and surpass established molecular biomarkers in precision oncology.^{15 34} Furthermore, real-time clinical copilots such as EyeFM have validated these advancements in randomised controlled trials, showing dramatic improvements in referral accuracy and a significant increase in patient compliance with follow-up care.³⁵ Yet, the impressive capabilities of these models do not protect them from fundamental errors. Notably, one study found that multimodal medical AI models can produce highly confident and detailed explanations even when their diagnoses are incorrect, at times describing visual features that are not actually present.³⁶

This 'illusion of readiness' captures how easily shortcut behaviours can be mistaken for genuine clinical and visual grounding. Overall, the literature suggests that high benchmark performance does not translate into clinical readiness, even with maturing coordination of modalities.

Reasoning models: optimising clinical logic and trust

While frontier models possess significant mathematical and computational proficiency, clinical reasoning is a distinct cognitive domain that requires independent optimisation, considering gaps and promise for improvement. Primarily, the literature suggests that current biomedical benchmarks often overestimate true reasoning ability: one study found only 33% of questions require multi-step reasoning, while models lose 40–60% accuracy when forced to revise incorrect hypotheses.³⁷ To address important limitations in optimised clinical reasoning, Med-process reward modeling (PRM) introduced a medical domain-specific process reward model. It scored each reasoning step for factual accuracy, logical coherence and clinical relevance rather than evaluating only final answers. Using a retrieval-augmented generation-as-a-judge labelling pipeline where medical evidence supports the scoring of each reasoning chain link, Med-PRM achieves 72–73% accuracy across seven benchmarks, substantially outperforming open-source models with the strongest gains on reasoning-heavy tasks.³⁸

AI in clinical workflows

While real-world AI deployments have improved diagnostic accuracy and reduced administrative burden, the question of how to optimise human-AI collaboration, without eroding clinical skills, remains unresolved.

Emerging evidence from real-world clinical deployment

Recent studies suggest rapid improvement in the clinical reasoning capabilities of AI systems, although real-world deployment remains constrained by workflow complexity. AI has substantially improved since 2022, in many cases outperforming clinicians on diagnostic and management tasks in controlled settings. However, in some of these studies, clinicians were evaluated without access to clinical decision support tools they would use in practice, which could overstate AI's relative advantage.^{39 40} Clinicians were typically strongest at adapting to inconsistent information and physical exam findings, highlighting the challenges that AI systems face in real-world clinical environments.³⁹ In 2025, a Penda Health and OpenAI large-scale quality improvement study with LLM clinical decision-support conducted across 15 clinic sites in Kenya (39 849 visits) reported 16% and 13% relative reductions in physician-rated diagnostic and treatment errors, though patient-reported outcomes did not differ significantly.⁴¹ Further studies are needed to assess clinical outcomes and generalisability, given the prompt-heavy and context-heavy design.

In diagnostic interpretation, models can serve as a helpful copilot for clinicians. In a prospective, real-world implementation study in Germany across 12 sites, radiologists who opted to use AI support achieved a higher breast cancer detection rate (6.7 vs 5.7 per 1000; 17.6% relative increase, CI 5.7% to 30.8%) without a significant increase in recall rate.⁴² In a primary care spirometry interpretation study, clinicians using AI performed better at identifying the top diagnosis (mean difference, 9.0; CI 4.5 to 13.3), establishing a diagnosis of chronic obstructive pulmonary disease (mean difference, 15.9; CI 9.0 to 22.7) or including the diagnosis in the differential diagnosis (mean difference, 7.3; CI 3.0 to 11.7), as determined by expert pulmonologists.⁴³

Administrative and workflow automation

While much research focuses on diagnostic reasoning and clinical decision support, the most mature real-world deployments of AI in healthcare today remain administrative. Several studies showed that ambient scribes reduce administrative burden, with physicians reporting reduced cognitive workload, improved patient attention and less time spent on documentation.⁴⁴ However, measured time savings have been modest, with some studies reporting around 20 s per note.^{45 46} Accuracy and the rates of clinically meaningful errors also remain a concern, though current evidence is largely limited to retrospective or benchmark study designs.^{18 47 48} Areas such as clinical inbox management and discharge summaries show similar trends of growing adoption but still require further testing to understand their impact across different clinical contexts.

Human-AI collaboration remains underdeveloped

Even with advances in real-world deployment and diagnostic interpretation, evidence suggests that a limiting factor may lie in human-AI interaction, rather than model capability. In a meta-analysis of 52 studies, human-AI teaming demonstrated higher diagnostic reliability than clinicians alone. However, full complementarity, or when the combined performance of clinicians and models working together is greater than each in isolation, was rarely achieved.⁴⁹ A randomised controlled trial of 70 physicians showed that a custom collaborative GPT-4 system designed to create a joint synthesis of human and AI reasoning found improvements in diagnostic accuracy when AI served as a second opinion (82% vs 75%; mean difference 6.8 percentage points, CI 4.0 to 9.6) or first opinion (85% vs 75%; mean difference 9.9 percentage points, CI 4.7 to 15.0), with no significant difference from AI-alone performance (87%; $p=0.20$).⁵⁰ Similar results were observed in another randomized controlled trial (RCT) study, which tested Open AI GPT-4's ability to aid physicians with management decisions.⁴⁰ Despite this, human-AI teaming did not outperform the model alone.

Risks of automation bias and clinician deskilling

Alongside the opportunities from human-AI collaboration, the risks of routine AI exposure to clinical expertise remain largely unexamined. In 2025, research showed that AI literacy courses can improve physician-AI teamwork.⁵¹ However, those benefits may be complicated by risks of automation bias. The randomised trial, conducted from June to August 2025, reported that physicians exposed to errors scored 14 percentage points (95% CI -18.9 to -9.1) lower in diagnostic accuracy than their peers in the control group.⁵¹ These findings, together with a retrospective observational study reporting declines in adenoma detection rates among clinicians following AI-assisted colonoscopies,⁵² suggest that skill redistribution under sustained AI use follows a pattern consistent with prior waves of technology adoption, rather than a novel risk. The discussion now evolves to identifying which critical skills are worth deliberately preserving, and on creating safety structures to compensate for those that will diminish. Both over-reliance on automation and clinician deskilling warrant further research to understand their causes and identify strategies that preserve clinical expertise alongside AI adoption.

Patient facing AI

Patient-facing AI shows early promise for improving access and health behaviours, but patient preference is an unreliable proxy for safety, and evaluation frameworks have yet to catch up.

Patient-facing conversational AI

In a randomised, double-blind study of 159 simulated patient cases, Google's AMIE outperformed 20 physicians on nearly all measures of diagnostic accuracy, communication quality and empathy.¹⁵ Similarly, specialised systems continue to demonstrate performance gains over general models. Google's Personal Health LLM, a Gemini-based system fine-tuned for personalised sleep and fitness recommendations, performed at or above the level of human experts and meaningfully outperformed base models, underscoring the continued value of specialisation even as general model capabilities improve.⁵³

Beyond diagnostic performance, randomised and controlled evaluations tested whether patient-facing AI influences behavioural and clinical outcomes. In separate studies, participants preferred AI-generated responses over human-generated content in medical advice and health coaching contexts, largely because the responses were easier to understand, used more positive and encouraging language and provided clearer, more actionable guidance.^{54 55} One RCT involving 368 adults found that an AI diabetes prevention programme was non-inferior to human coaching in terms of weight, hemoglobin A1c (HbA1c) and physical activity outcomes.⁵⁶ Additional applications include patient-facing support for understanding clinical notes and medical language translation, with early studies demonstrating improved understanding.^{57 58} These findings provide early evidence

that patient-facing AI can influence health behaviours and clinical outcomes, though interpretation is limited by study design and potential human bias toward more empathetic-perceived responses.

Limits of patient-mediated safety validation

Importantly, patients may not reliably distinguish AI-generated medical advice from that authored by clinicians and frequently rate AI responses, even low-accuracy outputs, as equally or more trustworthy and complete. Beyond rating AI responses as comparable to physicians, participants reported being just as likely to follow potentially harmful low-accuracy recommendations as on high-accuracy AI or physician responses, suggesting that the burden of medical validation could increasingly shift onto patients as more technology becomes patient-facing.⁵⁵

Applied AI

Research on model capability is extensive, with studies across multiple medical specialties routinely showing incremental, task-specific improvements. Randomised prospective trials have already commenced and should be the next wave of evidence in late 2026 and 2027 (table 2).

Specialty-tuned models for high-signal clinical tasks

By 2025, the most clinically translatable AI advances came from specialty-tuned models solving specific problems with clear endpoints. Imaging remained the dominant use case, with high-impact studies across pathology, neurology, nephrology, oncology, pulmonology and cardiology. Several studies collectively demonstrated that AI achieved specialist-level accuracy in coeliac disease diagnosis, carotid ultrasound, structural heart disease detection and breast cancer malignancy detection.^{59–62}

The rise of domain-specific foundational models

The more consequential development may be the shift from single-task models toward domain-specific foundation models. LCTfound, a lung CT vision foundation model, achieved an area under the receiver operating characteristic curve (AUROC) of 0.95 for rare diseases, performed state-of-the-art non-small cell lung cancer (NSCLC) prognosis stratification and received top Dice scores for mediastinal tumour segmentation.⁶³ EchoPrime, trained on 12.1 million echocardiogram videos and cardiologist reports, matched or exceeded task-specific models across five health systems (mean AUC 0.85–0.92).⁶⁴ Woollie, an oncology-specific LLM trained on Memorial Sloan Kettering data, generalised credibly to an external UCSF cohort (AUROC 0.97 internal; 0.88 on the UCSF cohort).⁶⁵ Merlin, a three-dimensional vision-language model (VLM) for abdominal CT, similarly demonstrated strong generalisation across 752 individual tasks and over 44 000 external scans, outperforming two-dimensional VLMs, CT foundational models and off-the-shelf radiology models.⁶⁶ The common theme is that a single well-constructed foundation model can reduce the need to build new architectures and labelling pipelines

Table 2 Summary table of selected randomised controlled trials from 2025

Study	Title	Country of origin	AI system used	Purpose
Unlu <i>et al</i> ⁷⁰	Manual vs AI-Assisted Prescreening for Trial Eligibility Using Large Language Models—A Randomised Clinical Trial	USA	Large Language Model (LLM) (GPT-4) processing unstructured patient data.	To compare the efficiency and accuracy of AI-assisted versus manual clinician prescreening for heart failure trial eligibility.
Wang <i>et al</i> ⁷⁸	An adaptive AI-based virtual reality sports system for adolescents with excess body weight: a randomised controlled trial	China	Adaptive AI VR sports system using deep reinforcement learning and transformer agents.	To evaluate the efficacy of REVERIE (Real-World Exercise and Virtual Reality Exercise Research in Education), a virtual reality (VR)-based exercise system, in reducing fat mass and improving cognitive and physical health in adolescents
Qazi <i>et al</i> ⁵¹	AI-literacy training enhances physician-LLM diagnostic collaboration in a resource-limited setting: a randomised controlled trial	Pakistan	LLMs used as diagnostic decision support.	To determine if a structured AI-literacy curriculum improves physician-LLM diagnostic reasoning and reduces bias in healthcare settings.
Qazi <i>et al</i> ⁷⁹	Automation Bias in Large Language Model Assisted Diagnostic Reasoning Among AI-Trained Physicians: A Randomised Clinical Trial	Pakistan	LLMs providing diagnostic recommendations.	To assess whether AI-trained physicians remain vulnerable to automation bias when exposed to erroneous LLM recommendations.
Goh <i>et al</i> ⁴⁰	GPT-4 assistance for improvement of physician performance on patient care tasks: a randomised controlled trial	USA	GPT-4 with conventional resources	To evaluate the impact of GPT-4 on physician performance regarding open-ended patient management reasoning tasks.
Wu <i>et al</i> ³⁵	An eyecare foundation model for clinical assistance: a randomised controlled trial	China	Multimodal vision-language foundation model	To evaluate a multimodal foundation model as a clinical copilot for retinal disease screening.
Abdallah <i>et al</i> ⁸⁰	TrialMatchAI: an end-to-end AI-powered clinical trial recommendation system to streamline patient-to-trial matching	France	Retrieval-Augmented Generation (RAG) framework using Gemma-2-2B	To develop and validate a RAG-based framework for automated clinical trial matching using entire unstructured patient note histories.
Afshar <i>et al</i> ⁴⁶	A Pragmatic Randomised Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being	USA	Abridge (Abridge AI)	To evaluate the impact of ambient AI scribes on healthcare practitioners' well-being (burnout and fulfilment) and documentation efficiency in a real-world deployment.
Lukac <i>et al</i> ⁴⁵	Ambient AI Scribes in Clinical Practice: A Randomised Trial	USA	Microsoft Dragon Ambient eXperience Copilot and Nabla	To investigate the effects of two AI scribes on documentation time, physician psychometrics, usability, accuracy and safety.

This list of randomised controlled trials was selected from the State of Clinical AI Report 2026. AI, artificial intelligence.

for every clinical question in the domain. With this shift come challenges such as the need for large, high-quality data, the risk of limited transferability across clinical contexts and increased validation complexity.

AI-enabled risk stratification and triage

Alongside imaging, specialties are increasingly repurposing abundant, non-invasive data to improve risk stratification. In the emergency department, ROMIAE achieved accuracy comparable to established clinical

tools for ruling out acute myocardial infarction, and augmenting the History, ECG, Age, Risk Factors, and Troponin (HEART) score with AI-ECG improved net reclassification by 20% across 18 university-level teaching hospitals in South Korea.⁶⁷ Further prospective research suggested earlier patient identification in primary care diabetic kidney disease screening and fewer complications in colorectal cancer surgery.⁶⁸ Operational use cases are also broadening in parallel, including triaging

patients/referrals, clinical trial eligibility and adjudicating adverse events in clinical trials.^{69–71}

Prospective evaluation and real-world deployment

While frontier models strive for field-wide transformation, real paths to clinical deployment are prospectively studying models within context and task-specific domains. In the largest real-world evaluation of stroke AI to date, England's National Health Service deployed Brainomix 360 across 26 hospitals and observed a 100% increase in endovascular thrombectomy rates at AI-enabled sites, significantly greater than trends in hospitals without AI. Transfer time was also reduced and AI-reviewed patients had higher odds of good functional outcomes (OR 1.16, 95% CI 1.07 to 1.25).⁷² With gains in clinical accuracy, the focus should now be on prospective clinical trials to assess patient outcomes and the optimal integration of workflows.

DISCUSSION

Several themes emerge from this synthesis that merit explicit consideration:

The benchmark-to-bedside gap is a structural problem

A consistent finding across domains is that strong performance on controlled evaluations does not reliably predict clinical utility. Models that achieve near-perfect scores on medical licensing examinations fail to recognise when the correct answer is absent.²⁰ Systems that produce expert-level diagnoses from structured vignettes lose substantial accuracy when the same cases are presented as unstructured dialogue.²⁵ Multiagent architectures with individually superior components can underperform less-optimised configurations due to failures in information flow.³³ These studies reflect a fundamental mismatch between how models are evaluated and how medicine is practised. Closing this gap will require evaluation frameworks that highlight uncertainty handling, longitudinal reasoning and performance under incomplete information rather than accuracy on static, well-defined tasks.

Human–AI collaboration is the critical unsolved problem

Across multiple randomised trials, clinician–AI teams failed to outperform the AI model alone, despite consistently outperforming clinicians working without AI.⁴⁰ This ceiling effect suggests that current interaction paradigms do not enable clinicians to effectively leverage AI capabilities. Compounding this, emerging evidence of automation bias and deskilling suggests that poorly designed AI integration could erode the very clinical expertise it is meant to augment.^{51 52} Designing human–AI workflows that achieve genuine complementarity, where each corrects the other's errors, remains an open research challenge and arguably the highest-priority prerequisite for safe deployment.

Safety evaluation lags behind capability development

The NOHARM framework's finding that frontier models produce potentially harmful outputs in approximately one in five primary care consultations is striking, given these same

models' strong performance on standard benchmarks.²⁷ Similarly, the persistent inability of models to calibrate their own confidence portends a concerning possibility that the systems most likely to cause harm are also the least likely to signal uncertainty.²⁸ This concern is compounded in patient-facing contexts, where patients cannot reliably distinguish AI-generated advice from that authored by clinicians and frequently rate low-accuracy AI outputs as equally or more trustworthy, suggesting that safety assurance cannot rely on patient-level discernment and must instead be built upstream through evaluation, transparency and accountability structures.⁵³ As that report argues, a coordinated national infrastructure for clinical AI evaluation, analogous to the federal investment that enabled EHR adoption, may be necessary to close this governance gap.

Where to go from here

Three priorities should guide the field. First, research investment should shift toward prospective, multisite trials that evaluate AI systems (and claimed capabilities) within authentic clinical workflows, measuring not only diagnostic accuracy but also downstream patient outcomes, clinician workflow impact and safety events. Second, evaluation must evolve beyond accuracy-centric benchmarks to frameworks that assess metacognitive calibration, performance under uncertainty, equity across populations and behaviour over longitudinal patient trajectories. Lastly, and most importantly, patient and user safety should be placed at the forefront. In addition to ensuring that these models minimise harm to patients, a more pernicious danger is harm to the physicians or healthcare staff using them; care must be taken to avoid physician deskilling or automation bias; hence, future work on safe human–computer interaction is critical.

Limitations

This narrative review has several limitations. It draws primarily on the ARISE's State of Clinical AI Report 2026 and a targeted literature search rather than a systematic review with predefined search protocols. Study selection was supplemented by expert nomination, which may introduce selection bias towards studies from well-networked academic institutions. The review reflects a temporal snapshot of a rapidly evolving field; findings may shift as preprints undergo peer review and new models are released. This review also has limitations of scope. By concentrating on the six domains where recent evidence clustered, we do not address equity and algorithmic bias, regulation and governance or implementation science in depth. Equity in particular remains a first-order concern as model performance depends on the representativeness of training data, and few studies reviewed report outcomes stratified by population, meriting future dedicated scrutiny as the evidence base matures.

CONCLUSION

What changed in 2025–2026? As expected, models have increased in size, scope and ambition. For example, medical-event foundation models matured from proof-of-concept to systems that tokenise longitudinal records to forecast

clinical trajectories (Delphi-2M, CoMET).^{29 30} Perhaps even more importantly, we see progress in how we evaluate these models, prospectively and *in mundo*: for the first time, a large-scale prospective trial tested LLM decision support in routine care.⁴¹ This past year, we have also seen some of the dangers of overreliance on these tools: the first RCT-grade and observational-grade evidence of automation bias and clinician deskilling emerged^{51 52}; multiagent systems revealed an unexpected ‘optimisation paradox’ in which added components can degrade rather than improve performance.³³ Together these mark a shift from asking ‘what can models do’ to asking, ‘can we safely, durably and effectively deploy them?’ This is the question the next year of inquiry must answer.

Clinical AI in 2026 presents a key challenge: model capabilities are advancing faster than the evidence, evaluation infrastructure and governance structures needed to deploy them safely. Across six domains, we find that the most credible translational progress occurs on narrowly defined tasks, whether tackled by specialty-tuned models or general-purpose LLMs, supported by prospective evaluation and clear workflow integration, while broader clinical autonomy remains premature. The gap between what clinical AI can do in controlled settings and what it can be trusted to do at the bedside is one of the defining tensions of this new field. Bridging it will require not only better models but also better trials, better benchmarks and a better understanding of how best to partner human and machine intelligence.

Author affiliations

¹The University of Arizona College of Medicine Phoenix, Phoenix, Arizona, USA

²Stanford Department of Medicine, Stanford, Stanford, California, USA

³Radiation Oncology, Stanford University, Stanford, California, USA

⁴Department of Internal Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

⁵Department of Medicine, New York-Presbyterian/Columbia University Irving Medical Center, New York, New York, USA

⁶Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

⁷Department of Dermatology, Harvard Medical School, Boston, Massachusetts, USA

⁸Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

⁹Medical University of South Carolina, Charleston, South Carolina, USA

¹⁰Massachusetts Institute of Technology Institute for Medical Engineering & Science, Cambridge, Massachusetts, USA

¹¹Division of Neurology, University of Alberta, Edmonton, Alberta, Canada

¹²Department of Internal Medicine, Cambridge Health Alliance, Cambridge, Massachusetts, USA

¹³Hospital Medicine, Stanford University School of Medicine, Stanford, California, USA

¹⁴The Institute of Medical Education Research Rotterdam, Rotterdam, The Netherlands

¹⁵Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

¹⁶Stanford Division of Computational Medicine, Stanford University, Stanford, California, USA

¹⁷Stanford Clinical Excellence Research Center, Stanford University, Stanford, California, USA

¹⁸Division of Hospital Medicine, Department of Medicine, Stanford University, Stanford, California, USA

Contributors PB, DJHW, and EG conceived the review. PB led the literature search and synthesis of evidence. JEW, AP, and DJHW drafted the manuscript and figures. EG, AR, and JHC contributed to interpretation and critical revisions. All

authors reviewed and approved the final manuscript. The corresponding author, EG, attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. AI was used for minor grammatical edits. Content drafting was manually done by authors first, with subsequent review by AI, with content suggestions considered or rejected at the discretion of the authors. All final content was reviewed by authors. EG is the guarantor and accepts full responsibility for the work and/or conduct of the study.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Disclaimer This review is presented in written narrative form the ARISE network’s ‘State of Clinical AI Report 2026’ presentation, published online January 2026, with which it shares authorship, the six-domain organisational frameworks and the study selection. This article extends that report by synthesising key insights through formal narrative analysis, subjecting findings to rigorous peer review and broadening accessibility of healthcare AI evidence across academic and public audiences.

Competing interests Dr EG receives funding from the Gordon and Betty Moore Foundation, Macy Foundation, Stanford Artificial Intelligence in Medicine and Imaging—Human-Centered Artificial Intelligence Partnership Grant, Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) [R12]. Dr EG reports consulting fees or other compensation from Google, OpenAI, Samsung Research America, Roche Diagnostics, Novartis, Hello Heart, Grow Care and Faculty Connection. Dr Rodman reports funding from the Moore Foundation, Macy Foundation, NIH, ARPA-H, Google and Google DeepMind. Dr Chen reports cofounding Reaction Explorer, which develops and licenses organic chemistry education software, as well as paid consulting fees from Sutton Pierce, Younker Hyde Macfarlane and Sykes McAllister as a medical expert witness. He receives funding from the National Institutes of Health (NIH)/National Institute of Allergy and Infectious Diseases (1R01AI17812101), the NIH—National Center for Advancing Translational Sciences Clinical & Translational Science Award (UM1TR004921), the Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) [R12] [JHC], the NIH/Center for Undiagnosed Diseases at Stanford (U01 NS134358), the Stanford RAISE Health Seed Grant (2024), the Josiah Macy Jr. Foundation (AI in Medical Education) and the Stanford CARE AI Scholar Fellowship.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages) and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See <https://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Anastasia Perez <https://orcid.org/0009-0009-0836-4056>

Ethan Goh <https://orcid.org/0000-0003-1965-025X>

REFERENCES

- Google Slides. State of clinical AI report 2026. Available: https://docs.google.com/presentation/d/1A-TchQb5Hg3-0MoiUFV199FI DBgmGvYhh9DyEU9C_Po/edit?slide=id.g3a1c1566c13_0_0#slide=id.g3a1c1566c13_0_0 [Accessed 02 Jul 2026].
- Brodeur PG, Buckley TA, Kanjee Z, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv* 2025.

- 3 Preiksaitis C, Ashenburg N, Bunney G, *et al*. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. *JMIR Med Inform* 2024;12:e53787.
- 4 Li H, Zhang R, Lee Y-C, *et al*. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit Med* 2023;6:236.
- 5 Wang L, Bhanushali T, Huang Z, *et al*. Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations. *JMIR Ment Health* 2025;12:e70014.
- 6 Gong EJ, Bang CS, Lee JJ, *et al*. Large Language Models in Gastroenterology: Systematic Review. *J Med Internet Res* 2024;26:e66648.
- 7 Chen SF, Alyakin A, Seas A, *et al*. LLM-assisted systematic review of large language models in clinical medicine. *Nat Med* 2026;32:1152–9.
- 8 OpenAI. Introducing ChatGPT health. 2026. Available: <https://openai.com/index/introducing-chatgpt-health/> [Accessed 12 Jun 2026].
- 9 Introducing Perplexity Health. 2026. Available: <https://www.perplexity.ai/hub/blog/introducing-perplexity-health> [Accessed 17 Apr 2026].
- 10 Microsoft AI. Introducing copilot health. Available: <https://microsoft.ai/story/introducing-copilot-health/> [Accessed 12 Jun 2026].
- 11 Advancing claude in healthcare and the life sciences. Available: <https://www.anthropic.com/news/healthcare-life-sciences> [Accessed 12 Jun 2026].
- 12 The 2026 AI index report | stanford HAI. Available: <https://hai.stanford.edu/ai-index/2026-ai-index-report> [Accessed 12 Jun 2026].
- 13 Angus DC, Khera R, Lieu T, *et al*. AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. *JAMA* 2025;334:1650–64.
- 14 Buckley TA, Conci R, Brodeur PG, *et al*. Teaching large language models to reason like expert diagnosticians. *arXiv* 2026.
- 15 Tu T, Schaeckermann M, Palepu A, *et al*. Towards conversational diagnostic artificial intelligence. *Nature* 2025;642:442–50.
- 16 Jin Q, Wang Z, Yang Y, *et al*. AgentMD: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nat Commun* 2025;16:9377.
- 17 Metwally AA, Heydari AA, McDuff D, *et al*. Insulin resistance prediction from wearables and routine blood biomarkers. *Nature* 2026;652:451–61.
- 18 Scheid MR, Friedmann B, Oppenheim M, *et al*. Development and validation of a clinical wearable deep learning based continuous in-hospital deterioration prediction model. *Nat Commun* 2025;16:9513.
- 19 Li Y, Huang Q, Jiang J, *et al*. Large language model-based biological age prediction in large-scale populations. *Nat Med* 2025;31:2977–90.
- 20 Bedi S, Jiang Y, Chung P, *et al*. Fidelity of Medical Reasoning in Large Language Models. *JAMA Netw Open* 2025;8:e2526021.
- 21 McCoy LG, Swamy R, Sagar N, *et al*. Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study. *NEJM AI* 2025;2:Aldbp2500120.
- 22 Bedi S, Liu Y, Orr-Ewing L, *et al*. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* 2025;333:319–28.
- 23 Bedi S, Cui H, Fuentes M, *et al*. Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med* 2026;32:943–51.
- 24 Arora RK, Wei J, Hicks RS, *et al*. HealthBench: evaluating large language models towards improved human health. *arXiv* 2025.
- 25 Johri S, Jeong J, Tran BA, *et al*. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med* 2025;31:77–86.
- 26 Jiang Y, Black KC, Geng G, *et al*. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. *NEJM AI* 2025;2:Aldbp2500144.
- 27 Wu D, Haredasht FN, Maharaj SK, *et al*. First, do NOHARM: towards clinically safe large language models. *ArXiv* 2025:arXiv:2512.01241v2.
- 28 Griot M, Hemptinne C, Vanderdonck J, *et al*. Large Language Models lack essential metacognition for reliable medical reasoning. *Nat Commun* 2025;16:642.
- 29 Shmatko A, Jung AW, Gaurav K, *et al*. Learning the natural history of human disease with generative transformers. *Nature* 2025;647:248–56.
- 30 Waxler S, Blazek P, White D, *et al*. Generative medical event models improve with scale. *arXiv* 2025.
- 31 Nori H, Daswani M, Kelly C, *et al*. Sequential diagnosis with language models. *arXiv* 2025.
- 32 Chen X, Yi H, You M, *et al*. Enhancing diagnostic capability with multi-agents conversational large language models. *Npj Digit Med* 2025;8:159.
- 33 Bedi S, Mlaoui I, Shin D, *et al*. The optimization paradox in clinical AI multi-agent systems. *arXiv* 2025.
- 34 Xiang J, Wang X, Zhang X, *et al*. A vision-language foundation model for precision oncology. *Nature* 2025;638:769–78.
- 35 Wu Y, Qian B, Li T, *et al*. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nat Med* 2025;31:3404–13.
- 36 Gu Y, Fu J, Liu X, *et al*. The illusion of readiness in health AI. *arXiv* 2025.
- 37 Thapa R, Wu Q, Wu K, *et al*. Disentangling reasoning and knowledge in medical large language models. *arXiv* 2025.
- 38 Yun J, Sohn J, Park J, *et al*. Med-prm: medical reasoning models with stepwise, guideline-verified process rewards. *arXiv* 2025.
- 39 Zeltzer D, Kugler Z, Hayat L, *et al*. Comparison of Initial Artificial Intelligence (AI) and Final Physician Recommendations in AI-Assisted Virtual Urgent Care Visits. *Ann Intern Med* 2025;178:498–506.
- 40 Goh E, Gallo RJ, Strong E, *et al*. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat Med* 2025;31:1233–8.
- 41 Korom R, Kiptinness S, Adan N, *et al*. AI-based clinical decision support for primary care: a real-world study. *arXiv* 2025.
- 42 Eisemann N, Bunk S, Mukama T, *et al*. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat Med* 2025;31:917–24.
- 43 Doe G, Banya W, Edwards G, *et al*. AI-Assisted Spirometry Interpretation in Primary Care: A Randomized Controlled Trial. *NEJM AI* 2025;2:Aloa2400804.
- 44 Olson KD, Meeker D, Troup M, *et al*. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. *JAMA Netw Open* 2025;8:e2534976.
- 45 Lukac PJ, Turner W, Vangala S, *et al*. Ambient AI Scribes in Clinical Practice: A Randomized Trial. *NEJM AI* 2025;2:Aloa2501000.
- 46 Afshar M, Ryan Baumann M, Resnik F, *et al*. A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being. *NEJM AI* 2025;2:Aloa2500945.
- 47 Williams CYK, Subramanian CR, Ali SS, *et al*. Physician- and Large Language Model-Generated Hospital Discharge Summaries. *JAMA Intern Med* 2025;185:818–25.
- 48 Asgari E, Montaña-Brown N, Dubois M, *et al*. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med* 2025;8:274.
- 49 Liu P, Zhang J, Chen S, *et al*. Human-AI teaming in healthcare: 1 + 1 > 2? *Npj Artif Intell* 2025;1:47.
- 50 Everett SS, Bunning BJ, Jain P, *et al*. From tool to teammate: a randomized controlled trial of clinician-ai collaborative workflows for diagnosis. *Health Informatics* [Preprint] 2025.
- 51 Qazi IA, Ali A, Khawaja AU, *et al*. AI-literacy training enhances physician-llm diagnostic collaboration in a resource-limited setting: a randomized controlled trial. *medRxiv* [Preprint] 2025.
- 52 Budzyń K, Romańczyk M, Kitala D, *et al*. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol Hepatol* 2025;10:896–903.
- 53 Khasentino J, Belyaeva A, Liu X, *et al*. A personal health large language model for sleep and fitness coaching. *Nat Med* 2025;31:3394–403.
- 54 Mantena S, Johnson A, Oppezzo M, *et al*. Fine-tuning LLMs in behavioral psychology for scalable health coaching. *NPJ Cardiovasc Health* 2025;2:48.
- 55 Shekar S, Pataranutaporn P, Sarabu C, *et al*. People Overtrust AI-Generated Medical Advice despite Low Accuracy. *NEJM AI* 2025;2:Aloa2300015.
- 56 Mathioudakis N, Lalani B, Abusamaan MS, *et al*. An AI-Powered Lifestyle Intervention vs Human Coaching in the Diabetes Prevention Program: A Randomized Clinical Trial. *JAMA* 2025;334:2079–89.
- 57 Ray M, Kats DJ, Moorkens J, *et al*. Evaluating a Large Language Model in Translating Patient Instructions to Spanish Using a Standardized Framework. *JAMA Pediatr* 2025;179:1026–33.
- 58 Kumar A, Wang H, Muir KW, *et al*. A Cross-Sectional Study of GPT-4-Based Plain Language Translation of Clinical Notes to Improve Patient Comprehension of Disease Course and Management. *NEJM AI* 2025;2:Aloa2400402.
- 59 Luo L, Wu M, Li M, *et al*. A large model for non-invasive and personalized management of breast cancer from multiparametric MRI. *Nat Commun* 2025;16:3647.
- 60 Poterucha TJ, Jing L, Ricart RP, *et al*. Detecting structural heart disease from electrocardiograms using AI. *Nature* 2025;644:221–30.

- 61 Jiang H, Zhao A, Yang Q, *et al.* Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system. *Nat Commun* 2025;16:7893.
- 62 Jaeckle F, Denholm J, Schreiber B, *et al.* Machine Learning Achieves Pathologist-Level Coeliac Disease Diagnosis. *NEJM AI* 2025;2:aioa2400738.
- 63 Gao Z, Zhang G, Liang H, *et al.* A lung CT vision foundation model facilitating disease diagnosis and medical imaging. *Nat Commun* 2025;17:35.
- 64 Vukadinovic M, Chiu I-M, Tang X, *et al.* Comprehensive echocardiogram evaluation with view primed vision language AI. *Nature* 2026;650:970–7.
- 65 Zhu M, Lin H, Jiang J, *et al.* Large language model trained on clinical oncology data predicts cancer progression. *NPJ Digit Med* 2025;8:397.
- 66 Blankemeier L, Kumar A, Cohen JP, *et al.* Merlin: a computed tomography vision-language foundation model and dataset. *Nature* 2026;652:1318–28.
- 67 Lee MS, Shin TG, Lee Y, *et al.* Artificial intelligence applied to electrocardiogram to rule out acute myocardial infarction: the ROMIAE multicentre study. *Eur Heart J* 2025;46:1917–29.
- 68 Rosen AW, Ose I, Gögenur M, *et al.* Clinical implementation of an AI-based prediction model for decision support for patients undergoing colorectal cancer surgery. *Nat Med* 2025;31:3737–48.
- 69 Gaber F, Shaik M, Allega F, *et al.* Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *NPJ Digit Med* 2025;8:263.
- 70 Unlu O, Varugheese M, Shin J, *et al.* Manual vs AI-Assisted Prescreening for Trial Eligibility Using Large Language Models-A Randomized Clinical Trial. *JAMA* 2025;333:1084–7.
- 71 Marti-Castellote PM, Badrouchi S, Claggett B, *et al.* Using Artificial Intelligence to Adjudicate Major Adverse Cardiovascular Events in Clinical Trials. *J Am Coll Cardiol* 2026;87:1856–66.
- 72 Nagaratnam K, Neuhaus AA, Fensome L, *et al.* Artificial intelligence imaging decision support for acute stroke treatment in England: a prospective observational study. *Lancet Digit Health* 2025;7:100927.
- 73 Aali A, Bikia V, Varma M, *et al.* 2026 Toward expert-level medical text validation with language models. *Npj Digital Medicine*.
- 74 Dai W, Chen P, Lu M, *et al.* CLIMB: data foundations for large scale multimodal clinical foundation models. *arXiv [Preprint]* 2025.
- 75 Diego FV, Proniakin O, Gruber V-E, *et al.* 2026 MedPI: Evaluating AI Systems in Medical Patient-facing Interactions. *Health Informatics*.
- 76 Mehndru N, Golchini N, Garg N, *et al.* ER-Reason: a benchmark dataset for LLM clinical reasoning in the emergency room. *arXiv [Preprint]* 2025.
- 77 Zuo Y, Qu S, Li Y, *et al.* MedXpertQA: benchmarking expert-level medical reasoning and understanding. *arXiv [Preprint]* 2025.
- 78 Wang J, Qin Y, Wu Q, *et al.* An adaptive AI-based virtual reality sports system for adolescents with excess body weight: a randomized controlled trial. *Nat Med* 2025;31:2255–68.
- 79 Qazi IA, Ali A, Khawaja AU, *et al.* Automation Bias in Large Language Model-Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy — A Randomized Clinical Trial. *NEJM AI* 2026;3.
- 80 Abdallah M, Nakken S, Georges M, *et al.* 2026 TrialMatchAI: an end-to-end AI-powered clinical trial recommendation system to streamline patient-to-trial matching. *Nat Commun* 17.