

PERSPECTIVE

Clinical AI in 2025–2026 — Six Themes from the SAIL Year in Review

Eileen Kim , M.D.,^{1,2} Marianna V. Elia , M.S.E.,^{3,4} Michelle M. Li , Ph.D.,⁵ Elizabeth Healey , Ph.D.,⁶ Lauren Erdman , Ph.D.,⁷ Payal Chandak , Ph.D.,⁸ Peniel Argaw , Ph.D.,⁹ Ben Li , M.D.,¹⁰ Kameron C. Black , M.D.,¹¹ Jayson Marwaha , M.D., M.Sc.,¹² Hugo Morales , M.D.,¹³ Anna Taylor , M.S., C.I.P.C.T.,¹⁴ Joseph Kvedar , M.D.,¹⁵ Peter Brodeur , M.D., M.A.,¹⁶ Brett Beaulieu-Jones , Ph.D.,¹⁷ Emily Alsentzer , Ph.D.,¹⁸ Julian C. Hong , M.D., M.S.,^{3,4,19} and Jean Feng , Ph.D., M.S.^{3,4,20}

Received: June 16, 2026; Revised: August 21, 2026; Accepted: August 28, 2026; Published: October 16, 2026

Abstract

In the past year, the medical field's relationship to clinical AI has progressed from rapid and often speculative adoption toward evaluation of its impact on patient outcomes, clinical workflows, and society. Drawing from the Symposium on AI for Learning Health Systems (SAIL) Year in Review (May 2026), this Perspective examines six key areas that dominated the domain of clinical AI in the past year: clinical AI evaluation, workflow integration, direct-to-consumer applications, regulation and societal impacts, AI in medical education, and foundational advancements in technology. AI's successes across these areas have been uneven, reflecting its jagged frontier. We highlight milestones that illustrate what AI has already achieved in health care, its potential, and major challenges that remain ahead. (Funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health [grant number T32HD040128].)

Introduction

Applications of AI in health care have rapidly expanded in the past year and continue to grow, with reports that over 80% of health care organizations are now using it^{1,2} and 32% of adults are turning to it for health information and advice.³ At the same time, rapid expansion must be weighed against the current limits and risks of AI, as well as its cost. Notably, its successes across clinical applications have been uneven, reflecting what has been termed a “jagged frontier”: AI models can excel at certain tasks while unexpectedly underperforming on others that appear similarly or even less complex, making its capabilities difficult to predict.⁴

This Perspective summarizes insights from the Year in Review at the Symposium on AI for Learning Health Systems (SAIL) held in May 2026, an annual meeting focusing on the opportunities and barriers to deploying AI in clinical settings. Ahead of the conference, the SAIL Year in Review committee emailed a survey to all SAIL 2025 attendees to identify the past 12 months' key themes, priorities for the near future, influential recent publications,

Eileen Kim and Marianna V. Elia and Julian C. Hong and Jean Feng contributed equally to this article.

The author affiliations are listed at the end of the article.

Jean Feng can be contacted at jean.feng@ucsf.edu.

Table 1. Themes in Clinical AI for 2025–2026 from SAIL Year in Review.*		
Theme	Examples	Looking Ahead
Evaluation: moving up the hierarchy of evidence	- AI and agentic benchmarks^{5,6} - Randomized studies on clinical vignettes^{7,8} - Prospective, and increasingly randomized, studies of deployed systems^{9,12}	More RCTs powered on patient outcomes; stronger construct validity for benchmarks; methodological rigor (blinding, assignment concealment) as trial volume grows
Clinical workflow integration	- Ambient AI scribes^{13,16} - Chart/discharge summarization^{17,18} - Patient communication and perceptions of AI use^{19,20}	Life cycle management of AI systems; outcome-level evidence; understanding human–AI interactions
Rise of direct-to-consumer AI	- Scale of consumers using AI for health-related queries²¹ - Purpose-built tools^{22,23} - Risks of general-purpose use^{24,25}	Safety guardrails and regulatory clarity for consumer-facing products
Regulation and societal impacts	- Contrasting regulatory approaches²⁶ - Evolving regulation^{27,28} - New reimbursement models²⁹ - Environmental footprint³⁰ - Labor impacts^{31–33}	Establishing regulatory frameworks for AI systems; evaluation of alternative reimbursement pathways; transparency in environmental footprint; addressing labor market effects
AI in medical education	- AI-enabled pedagogy^{34,35} - Effects of AI on medical knowledge and skill acquisition^{36,37} - Deskilling and never-skilling risks^{38,39}	Studying and addressing skill retention and development; developing curricula that build AI literacy
Foundational technology advances	- Large-scale foundation models^{40,41} - Multimodal clinical models, including wearables^{42–44}	Moving from prediction to causal understanding; extending AI beyond episodic, in-clinic encounters; open-source ecosystems for collaborative advancement

*RCT denotes randomized controlled trial; and SAIL, Symposium on AI for Learning Health Systems.

and major gaps in the field (Supplementary Appendix). After reviewing and discussing the 27 responses, our team of clinical AI researchers distilled six themes reflecting the progress and persistent challenges in the field (Table 1^{5–44}). Both respondent and review groups represented a wide breadth of clinical, data science, and operational backgrounds across both academia and industry.

Evaluation of Clinical AI Systems: Moving up the Hierarchy of Evidence

Demand for rigorous evaluation has increased in the past year, now that clinical AI systems have saturated simpler benchmarks based on medical licensing examinations.⁴⁵ The past year has focused on moving up in the hierarchy of evidence, starting with better benchmarks at the lowest level, moving up to randomized studies on vignettes and case reports, and finally to randomized clinical trials that measure the impact on patient outcomes.

Clinical AI benchmarks have grown increasingly sophisticated in an effort to more faithfully represent real-world clinical applications. Health AI–centric benchmarks such as MedHELM (Holistic Evaluation of Large Language Models for Medical Tasks) now aim to capture more fully real tasks

encountered in clinical practice, including documentation, patient communication, and administrative tasks.⁵ At the time of MedHELM’s release, DeepSeek R1 achieved top overall scores (66%), while other models such as Claude 3.5 Sonnet offered favorable cost–accuracy trade-offs, achieving comparable results at 15% lower cost. Agentic benchmarks have also emerged alongside the increasing prevalence of AI agents: MedAgentBench evaluated agents in their ability to interact with a synthetic electronic health record, noting that early models struggled not with medical reasoning, but with basic instruction-following and application programming interface requests.⁶ Subsequent modifications to MedAgentBench’s harness to improve output consistency led to improved performance.⁴⁶ Despite this progress, concerns around the “construct validity”⁴⁷ of these benchmarks remain: Recent works have highlighted that the benchmarks do not accurately measure the skill they claim to,⁴⁸ as these systems still struggle on real patient records and real-world clinical tasks that are often not only more complex, but also require navigating ambiguous or underspecified instructions.

Consequently, many studies have pushed progress up the hierarchy of evidence in the past year. A significant number of these were randomized studies that assessed the impact of AI on clinical reasoning based on clinical vignettes. Because earlier studies had found that human–AI teams can often demonstrate worse performance than AI alone,⁴⁹

and do not necessarily outperform clinical teams with access to only conventional resources,⁵⁰ these randomized studies placed much heavier emphasis on improving human-computer interaction. These studies showed that thoughtfully designed collaborative workflows largely mitigated this phenomenon, such that no significant differences in diagnostic accuracy could be detected between human-AI teams and AI-only systems.⁷ Improving clinical reasoning with AI also requires proper AI literacy training: A randomized study in Pakistan showed that a mandatory 20-hour AI literacy curriculum improved physicians' diagnostic reasoning scores with AI assistance to 71.4%, compared with 42.6% for those using conventional resources alone.⁸

Finally, there are now a growing number of prospective randomized clinical trials (RCTs) of deployed AI systems, though they generally remain few in number.⁵¹ A multisite RCT enrolling 368 adults at increased risk for diabetes found that a fully automated AI-led diabetes prevention program was noninferior to a human-led program at much lower cost.⁹ The Mammography Screening with Artificial Intelligence (MASAI) trial in Sweden, enrolling 105,934 women, demonstrated that AI-supported mammography screening achieved 9% higher sensitivity without sacrificing specificity, yielding a 29% increase in cancer detection and 12% reduction in interval cancer rates, while reducing radiology workload by 44%.¹⁰ A pragmatic cluster-randomized trial conducted across 74 units in two health systems found a 35% decrease in instantaneous risk of death and an 11% decrease in length of stay with an early warning system based on machine learning using nurse surveillance documentation patterns.¹¹ While these trials represent positive results from the past year, other studies have provided more mixed findings. For instance, the Impact of Immediate AI-Enabled Patient Triage to Chest CT on the Lung Cancer Pathway (LungIMPACT) RCT showed that AI prioritization of suspicious chest x-rays decreased time to obtaining a report but did not improve time to cancer diagnosis, highlighting the need to consider clinical pathways more holistically.¹² Furthermore, a review of RCTs for clinical AI systems showed that, even among those that report favorable outcomes, methodological concerns were common, particularly with regard to assignment concealment and blinding.⁵¹ Consequently, these results illustrate the importance of proper experimental design, multidisciplinary collaboration, and reporting transparency to assess the clinical impact of distinct AI applications.

Clinical Workflow Integration

Administrative tasks such as documentation continue to dominate the clinical AI landscape. This progress has been led by the widespread deployment of AI scribes, which constituted one of the most significant clinical integration efforts of the last year. To understand the utility of these systems, leading health systems have created structured quality assurance frameworks as well as postdeployment monitoring systems.⁵² However, early evidence reveals a continuing gap between enthusiasm and demonstrated benefit.⁵³ Multiple large RCTs of ambient AI scribes found only modest documentation time savings and identified clinically significant inaccuracies across multiple platforms.¹³⁻¹⁶ Furthermore, these evaluations should be viewed as providing a snapshot of these tools' behaviors, as AI systems are prone to shifts and drifts in their performance and behavior.⁵⁴ In one notable case, a routine note template change caused ICD-10 documentation accuracy to plummet from 79% to 35% before user retraining restored it.⁵⁵ While initial frameworks have been proposed for the monitoring and maintenance of clinical AI systems,^{55,56} more research is needed to ensure that these systems have sufficient power to detect and remediate harmful data shifts before they affect care quality.⁵⁷

Another AI application that has seen rapid adoption across hospitals is in summarizing patient charts, such as with the release of the Draft Hospital Course Notes and Inpatient Insight tools by Epic Systems. Studies on AI-drafted discharge summaries have generally been promising, showing that AI-generated summaries were of comparable quality to those generated by physicians.^{17,18} Epic also reports hospitals achieving significant gains in productivity using these tools.² Nevertheless, reports on patient outcomes generally remain more limited. A recent report showed that designing these tools to interface with patients can help bridge this gap: When AI was used to translate complex hospital discharge summaries into plain language, patient comprehension significantly improved, particularly among those with lower health literacy and from historically marginalized racial and ethnic groups.¹⁹

Finally, the potential of AI to improve clinical workflows must be weighed against the current landscape of evolving patient attitudes toward physician AI usage. A recent study found that patients evaluated clinicians who marketed their use of AI as being less competent and empathetic, and that this decreased their willingness to schedule appointments

with them.²⁰ However, this may represent an evolving target as AI technology continues to develop and meaningful use cases are established.

The Rise of Direct-to-Consumer AI

Direct-to-consumer (DTC) AI in health care expanded dramatically this past year, driven in particular by frontier AI companies such as OpenAI, Microsoft, and Amazon.⁵⁸ For example, more than 230 million people reportedly used ChatGPT alone on a weekly basis for health advice.²¹ The growth in the number of such applications is, in no small part, due to the more limited regulatory framework, as these tools are stated not to give medical advice. The implications of these products in this largely unregulated landscape are considerable, with examples of these applications providing both benefit and harm.

Generally speaking, DTC applications built for purpose have demonstrated more meaningful promise. For instance, in the field of mental health, an RCT of the Therabot chatbot found significant reductions in symptoms of depression and anxiety compared with patients waitlisted for therapy, achieving a therapeutic alliance comparable to human health care providers.²² The explicit codesign of AI tools further supports this concept: A chatbot codesigned by community stakeholders and researchers was shown in an RCT to significantly increase health awareness among patients across a diverse set of 11 Chinese provinces.²³

On the other hand, recent studies have shown the immense risks that accompany trust in inaccurate tools. In an imitation game study, patients were unable to distinguish AI responses from those of physicians, rating AI responses as more trustworthy than human physicians, even when inaccurate.²⁴ Recent findings have also shown how consumer tools such as ChatGPT Health frequently undertriaged patients in critical condition, missing severe cases such as diabetic ketoacidosis over 50% of the time, and overwhelmingly failed to trigger crisis intervention resources when evaluating patients with mental health concerns.²⁵

Regulation and Societal Impacts

Medical device regulators globally are still developing strategies to best manage the ever-evolving landscape of AI-based technologies. A decade-long review highlighted how regulators in the top five medical device markets — the United States, the European Union, China, South Korea,

and Japan — share some similarities, such as the use of risk-stratified oversight and postmarket surveillance. However, they also vary in how they balance the trade-off between the speed at which products can be brought to market and the strength of clinical evidence required.²⁶ While the United States and China prioritize time to market and rely more heavily on postmarket surveillance, the European Union prioritizes clinical evidence, requiring stronger research findings such as those generated by RCTs; meanwhile, the approach of South Korea and Japan sits somewhere in between. In the past year, the United States has continued to push for faster release of products to the market by clarifying when software is exempt from regulatory oversight, including specific examples of generative AI applications.^{27,28} There are also growing efforts, both within and outside the U.S. Food and Drug Administration, to leverage AI to accelerate regulatory process and research.^{59,60}

Beyond regulation, the Centers for Medicare and Medicaid Services launched the Advancing Chronic Care with Effective, Scalable Solutions (ACCESS) model, which establishes a framework for outcome-aligned payments designed to reward vendors and providers who utilize technology to achieve measurable improvements in chronic disease management.²⁹ This will provide many AI-enabled technologies with a path for reimbursements, which has been a challenge historically.^{30,61} Health care systems currently have limited ability to track the environmental impacts of their generative AI workloads, which affects policies such as the English National Health Service's goal of achieving net zero,⁶² and labor market disruptions due to AI are emerging in the clinical world, as anticipated in early analyses by Anthropic.⁶³ Across the United States, a growing number of protests and strikes in health care have revolved around the implementation and impact of AI.³¹⁻³³

AI in Medical Education

The integration of AI into medical training was also a theme in the period 2025–2026, with early evidence suggesting potential applications to facilitate both skill acquisition and structured assessment. One broadly considered educational path is the ability for AI to personalize education. Medical students who trained with AI-tailored feedback reported greater preparedness and demonstrated superior performance in objective structured clinical examinations compared with peers using conventional methods.³⁴ This personalization can also extend to the ability to evaluate trainees; in a separate study, a multimodal AI system successfully evaluated anesthesia residents during simulated

clinical scenarios, demonstrating strong agreement with expert human examiners in the competency domains.³⁵ The results of this single trial, however, should be interpreted in the context of a broader body of literature, with a recent meta-analysis suggesting no difference in general knowledge acquisition, though with specific advantages over an extended period of time and practical skill development.³⁶

The potential of AI for upskilling or deskilling in medicine remains a subject of significant debate. While a long-appreciated phenomenon, the risks of automation bias remain relevant in the use of AI. The general body of literature suggests a high risk of deskilling, with a recent scoping review highlighting detriments across diagnostic specialties such as radiology, gastroenterology, and pathology.³⁸ Researchers even warn of “never skilling,” referring to the concern that trainees may never acquire certain skills.³⁹ In contrast, there are some data to suggest that AI can upskill trainees,^{7,8} and in a recent study, medical students using ChatGPT-4o outperformed both control groups and the AI model alone in diagnostic reasoning tasks.³⁷

Advancements in AI Capabilities

Many advancements have pushed the frontier of clinical AI this past year and are likely to shape the field going forward. One major area of advancement was training substantially larger foundation models and incorporating longitudinal data. Examples include Epic’s foundation model, which was trained on over 16 billion encounters in Cosmos⁴⁰; this model matched or exceeded task-specific models across 78 prediction tasks. Similarly, a transformer model trained on data in the U.K. Biobank demonstrated disease risk prediction on par with or better than task-specific approaches.⁴¹ Nevertheless, substantial technical challenges remain, as current models are optimized for prediction and not causal understanding.⁴²

Multimodal capabilities also saw significant advances. Google’s MedGemma 1.5 incorporates high-dimensional imaging modalities and demonstrates improved interpretation of clinical documents, with substantial gains across MRI and pathology classification, anatomical localization, and longitudinal x-ray analysis.⁴³ Open-source models will help the field collaboratively advance capabilities of clinical AI models.

Finally, a growing area of interest is to expand clinical AI beyond hospital or clinic walls. Multiple multimodal foundation models were developed to synthesize complex sensor data and behavioral data.^{44,64} Integrating these data

enables continuous, longitudinal phenotyping beyond episodic care, with particular consequence for conditions whose progression can be subclinical and therefore invisible to traditional clinical encounters.

Outlook

The medical field’s relationship to clinical AI shifted over the period 2025–2026 from speculative adoption toward rigorous evaluation and accountability. Research from this period has demonstrated that AI’s success in health care does indeed follow a jagged frontier: While carefully designed AI systems have achieved measurable improvements in patient outcomes, other studies have unmasked real-world challenges when deploying AI into clinical workflows and consumer products. We highlight several urgent priorities for the coming year. First, more widespread prospective evaluation of AI’s effect on clinical outcomes is needed. Second, the field requires robust tooling, guidelines, and regulation for life cycle management of clinical AI tools, particularly in the rapidly expanding consumer-facing space. Finally, we must better understand and address the systemic impacts of clinical AI systems, on medical skill retention and development, labor markets, and the environment. Nonetheless, we remain optimistic about AI’s prospects, given the deep engagement from the clinical AI community in advancing and validating the capabilities of these systems both inside and outside hospital walls.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Dr. Healey is supported by a grant (T32HD040128) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank those who responded to the survey from the Symposium on AI for Learning Health Systems (SAIL) Year in Review committee, including Hyung L. Kim, Monica M. Bertagnolli, Vivian Lee, Ricky Bloomfield, Yun Liu, Orvil Martinez, Prabhjot Singh, Abiel Roche-Lima, Shilpa Kobren, Alexandre DaSilva, Xiaomeng Li, Pierre Elias, Zhandong Liu, Simon Lee, Ming Tai-Seale, Duncan McElfresh, Marzyeh Ghassemi, Irene Chen, Faraz Ahmad, Shreya Johri, Roy Perlis, João Monteiro, Mary Beth Hamel, and Rupa Sarkar.

Author Affiliations

¹Division of General Internal Medicine, Department of Medicine, University of California, San Diego, La Jolla, USA

- ²Joan & Irwin Jacobs Center for Health Innovation, University of California, San Diego Health, La Jolla, USA
- ³Computational Precision Health Program, University of California, San Francisco and University of California, Berkeley, San Francisco and Berkeley, USA
- ⁴Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, USA
- ⁵Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
- ⁶Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA
- ⁷Department of Pediatrics, Cincinnati Children's Hospital, Cincinnati, OH, USA
- ⁸Harvard-MIT Program in Health Sciences and Technology, Harvard University and Massachusetts Institute of Technology, Boston and Cambridge, MA, USA
- ⁹Microsoft Research, Redmond, WA, USA
- ¹⁰Division of Vascular Surgery, University of Toronto, Toronto, ON
- ¹¹Stanford Division of Hospital Medicine, Stanford University, Stanford, CA, USA
- ¹²Department of Surgery, University of Michigan, Ann Arbor, USA
- ¹³Munai Health, São Paulo, Brazil
- ¹⁴MultiCare Connected Care, Tacoma, WA, USA
- ¹⁵Department of Dermatology, Mass General Brigham, Boston, MA, USA
- ¹⁶Department of Internal Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
- ¹⁷Department of Medicine, University of Chicago, Chicago, IL, USA
- ¹⁸Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
- ¹⁹Department of Radiation Oncology, University of California, San Francisco, San Francisco, USA
- ²⁰Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, USA

References

1. Lamb J. Generative AI in healthcare: adoption matures as agentic AI emerges. April 16, 2026 (<https://www.mckinsey.com/industries/healthcare/our-insights/generative-ai-in-healthcare-current-trends-and-future-outlook>).
2. Epic. Real results, right now: how Epic AI is reducing costs, improving care, and helping patients. March 10, 2026 (<https://www.epic.com/epic/post/real-results-right-now-how-epic-ai-is-reducing-costs-improving-care-and-helping-patients/>).
3. Montero A, Montalvo III J, Kearney A, Valdes I, Kirzinger A, Hamel L. KFF tracking poll on health information and trust: use of AI for health information and advice. KFF, March 25, 2026 (<https://www.kff.org/public-opinion/kff-tracking-poll-on-health-information-and-trust-use-of-ai-for-health-information-and-advice/>).
4. Dell'Acqua F, McFowland III E, Mollick E, et al. Navigating the jagged technological frontier: field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Harvard Business School Working Paper (Paper No. 24-013). September 2023 (<https://www.hbs.edu/faculty/Pages/item.aspx?num=64700>).
5. Bedi S, Cui H, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med* 2026;32:943-951. DOI: [10.1038/s41591-025-04151-2](https://doi.org/10.1038/s41591-025-04151-2).
6. Jiang Y, Black KC, Geng G, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. *NEJM AI* 2025;2(9). DOI: [10.1056/AIdbp2500144](https://doi.org/10.1056/AIdbp2500144).
7. Everett SS, Bunning BJ, Jain P, et al. From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. *NPJ Digit Med* 2026;9:409. DOI: [10.1038/s41746-026-02545-1](https://doi.org/10.1038/s41746-026-02545-1).
8. Qazi IA, Ali A, Khawaja AU, Akhtar MJ, Sheikh AZ, Alizai MH. Large language model diagnostic assistance for physicians in a lower-middle-income country: a randomized controlled trial. *Nat Health* 2026;1:198-205. DOI: [10.1038/s44360-025-00007-8](https://doi.org/10.1038/s44360-025-00007-8).
9. Mathioudakis N, Lalani B, Abusamaan MS, et al. An AI-powered lifestyle intervention versus human coaching in the diabetes prevention program: a randomized clinical trial. *JAMA* 2025;334:2079-2089. DOI: [10.1001/jama.2025.19563](https://doi.org/10.1001/jama.2025.19563).
10. Gommers J, Hernström V, Josefsson V, 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;407:505-514. DOI: [10.1016/S0140-6736\(25\)02464-X](https://doi.org/10.1016/S0140-6736(25)02464-X).
11. Rossetti SC, Dykes PC, Knaplund C, et al. Real-time surveillance system for patient deterioration: a pragmatic cluster-randomized controlled trial. *Nat Med* 2025;31:1895-1902. DOI: [10.1038/s41591-025-03609-7](https://doi.org/10.1038/s41591-025-03609-7).
12. Woznitza N, Smith L, Rawlinson J, et al. AI-based chest x-ray prioritization in the lung cancer diagnostic pathway: the LungIMPACT randomized controlled trial. *Nat Med* 2026;32:1737-1744. DOI: [10.1038/s41591-026-04253-5](https://doi.org/10.1038/s41591-026-04253-5).
13. Lukac PJ, Turner W, Vangala S, et al. Ambient AI scribes in clinical practice: a randomized trial. *NEJM AI* 2025;2(12). DOI: [10.1056/AIoa2501000](https://doi.org/10.1056/AIoa2501000).
14. Afshar M, Baumann MR, Resnik F, et al. A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI* 2025;2(12). DOI: [10.1056/AIoa2500945](https://doi.org/10.1056/AIoa2500945).
15. Shin HS, Williams H, Braykov N, Jahan A, Meller J, Orenstein EW. The influence of artificial intelligence scribes on clinician experience and efficiency among pediatric subspecialists: a rapid, randomized quality improvement trial. *Appl Clin Inform* 2025;16:1041-1052. DOI: [10.1055/a-2657-8087](https://doi.org/10.1055/a-2657-8087).
16. Chowdhury A, Casey M, Wilson J, et al. Comparing ambient scribes: a randomized crossover clinical trial addressing ambient scribe technologies' impact on physician burnout. *J Am Med Inform Assoc* 2026;33:990-999. DOI: [10.1093/jamia/ocag018](https://doi.org/10.1093/jamia/ocag018).
17. Williams CYK, Subramanian CR, Ali SS, et al. Physician- and large language model-generated hospital discharge summaries. *JAMA*

- Intern Med 2025;185:818-825. DOI: [10.1001/jamainternmed.2025.0821](https://doi.org/10.1001/jamainternmed.2025.0821).
18. Mehri T, Nadalini T, Hoekman AH, et al. Assessing the quality of AI-generated and physician-written discharge summaries: evaluation of an EHR-integrated tool in a Dutch academic hospital. *EBioMedicine* 2026;127:106247. DOI: [10.1016/j.ebiom.2026.106247](https://doi.org/10.1016/j.ebiom.2026.106247).
 19. Kumar A, Wang H, Muir KW, Mishra V, Engelhard M. 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(2). DOI: [10.1056/AIoa2400402](https://doi.org/10.1056/AIoa2400402).
 20. Reis M, Reis F, Kunde W. Public perception of physicians who use artificial intelligence. *JAMA Netw Open* 2025;8:e2521643. DOI: [10.1001/jamanetworkopen.2025.21643](https://doi.org/10.1001/jamanetworkopen.2025.21643).
 21. Open AI. Introducing ChatGPT health. January 7, 2026 (<https://openai.com/index/introducing-chatgpt-health/>).
 22. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI* 2025;2(4). DOI: [10.1056/AIoa2400802](https://doi.org/10.1056/AIoa2400802).
 23. Li S, Li Y, Zhou S, et al. A community-codesigned LLM-powered chatbot for primary care: a randomized controlled trial. *Nat Health* 2026;1:238-250. DOI: [10.1038/s44360-025-00021-w](https://doi.org/10.1038/s44360-025-00021-w).
 24. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. *NEJM AI* 2025;2(6). DOI: [10.1056/AIoa2300015](https://doi.org/10.1056/AIoa2300015).
 25. Ramaswamy A, Tyagi A, Hugo H, et al. ChatGPT health performance in a structured test of triage recommendations. *Nat Med* 2026;32:1671-1675. DOI: [10.1038/s41591-026-04297-7](https://doi.org/10.1038/s41591-026-04297-7).
 26. Zhang S, Li Y, Liu W, et al. A decade of review in global regulation and research of artificial intelligence medical devices (2015–2025). *Front Med (Lausanne)* 2025;12:1630408. DOI: [10.3389/fmed.2025.1630408](https://doi.org/10.3389/fmed.2025.1630408).
 27. U.S. Food and Drug Administration. Clinical decision support software: guidance for industry and food and drug administration staff. January 29, 2026 (<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>).
 28. Lawrence L, Aguilar M, Palmer K, Trang B. FDA announces sweeping changes to oversight of wearables, AI-enabled devices. *STAT*, January 6, 2026 (<https://www.statnews.com/2026/01/06/fda-pulls-back-oversight-ai-enabled-devices-wearables/>).
 29. Centers for Medicare and Medicaid Services. ACCESS (advancing chronic care with effective, scalable solutions) model. September 18, 2026 (<https://www.cms.gov/priorities/innovation/innovation-models/access>).
 30. Reynolds DJ. Carbon reporting practices in the NHS: emissions and omissions relating to artificial intelligence. *J Med Internet Res* 2025;27:e79174. DOI: [10.2196/79174](https://doi.org/10.2196/79174).
 31. Marcelo P, Peltz J, The Associated Press. New York nurses union wins 12% raise, AI safeguards in a tentative deal to end monthlong strike. *Fortune*, February 20, 2026 (<https://fortune.com/2026/02/20/new-york-nurses-union-raise-ai-safeguards-deal-longest-strike/>).
 32. Wong Q. One of California's first labor fights over AI is playing out at Kaiser. *Los Angeles Times*, February 6, 2026 (<https://www.latimes.com/business/story/2026-02-06/kaiser-workers-launch-war-against-ai-protesting-potential-job-losses-patient-harm>).
 33. Carroll J. Hundreds of Kaiser Permanente healthcare professionals stage informational picket. *KPBS*, August 1, 2025 (<https://www.kpbs.org/news/quality-of-life/2025/08/01/hundreds-of-kaiser-permanente-healthcare-professionals-stage-informational-picket>).
 34. Lavigne E, Lopez A, Frandon J, et al. AI-standardized clinical examination training on OSCE performance. *NEJM AI* 2025;2(8). DOI: [10.1056/AIoa2500066](https://doi.org/10.1056/AIoa2500066).
 35. Gershov S, Mahameed F, Raz A, Laufer S. Towards accurate and interpretable competency-based assessment: enhancing clinical competency assessment through multimodal AI and anomaly detection. *NPJ Digit Med* 2026;9:219. DOI: [10.1038/s41746-025-02299-2](https://doi.org/10.1038/s41746-025-02299-2).
 36. Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: a meta-analysis of randomized controlled trials. *BMC Med Educ* 2025;25:1175. DOI: [10.1186/s12909-025-07750-2](https://doi.org/10.1186/s12909-025-07750-2).
 37. Roemer A, Schlicker N, Kernder A, et al. Large language models enhance diagnostic reasoning of medical students in rheumatology: a randomized controlled trial. *BMC Med Educ* 2026;26:579. DOI: [10.1186/s12909-026-09079-w](https://doi.org/10.1186/s12909-026-09079-w).
 38. Heudel PE, Crochet H, Filori Q, Bachelot T, Blay JY. Artificial intelligence in medicine: a scoping review of the risk of deskilling and loss of expertise among physicians. *ESMO Real World Data Digit Oncol* 2026;12:100693. DOI: [10.1016/j.esmorw.2026.100693](https://doi.org/10.1016/j.esmorw.2026.100693).
 39. Ke Y, Jin L, Ong JCL, et al. AI-induced never-skilling in medical education. *Nat Med* 2026;32:1997-2006. DOI: [10.1038/s41591-026-04438-y](https://doi.org/10.1038/s41591-026-04438-y).
 40. Waxler S, Blazek P, White D, et al. Generative medical event models improve with scale. November 7, 2025 (<https://arxiv.org/abs/2508.12104>). Preprint.
 41. Shmatko A, Jung AW, Gaurav K, et al. Learning the natural history of human disease with generative transformers. *Nature* 2025;647:248-256. DOI: [10.1038/s41586-025-09529-3](https://doi.org/10.1038/s41586-025-09529-3).
 42. Bareinboim E. Causal artificial intelligence: a roadmap for building causally intelligent systems. October 11, 2025 (<https://causalai-book.net/>).
 43. Sellergren A, Kazemzadeh S, Jaroensri T, et al. MedGemma technical report. April 6, 2026 (<https://arxiv.org/abs/2507.05201>). Preprint.
 44. Erturk E, Kamran F, Abbaspourazad S, et al. Beyond sensor data: foundation models of behavioral data from wearables improve

- health predictions. In: Proceedings of the 43rd International Conference on Machine Learning. Seoul, South Korea: PMLR, June 23, 2025 (<https://openreview.net/pdf?id=DtVVltU1ak>).
45. Gong EJ, Bang CS, Lee JJ, Baik GH. Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks. *J Med Internet Res* 2025;27:e84120. DOI: [10.2196/84120](https://doi.org/10.2196/84120).
 46. Chen E, Postelnik S, Black K, Jiang Y, Chen JH. MedAgentBench v2: improving medical LLM agent design. *Pac Symp Biocomput* 2026;31:354-371. DOI: [10.1142/9789819824755_0025](https://doi.org/10.1142/9789819824755_0025).
 47. Bean AM, Othniel Kearn R, Romanou A, et al. Measuring what matters: construct validity in large language model benchmarks. In: Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks. San Diego, CA, November 3, 2025 (<https://openreview.net/pdf?id=mdA5lVvNcU>).
 48. Alaa A, Hartvigsen T, Golchini N, et al. Position: medical large language model benchmarks should prioritize construct validity. In: Proceedings of the 43rd International Conference on Machine Learning. Seoul, South Korea: PMLR, June 23, 2025 (<https://openreview.net/pdf?id=YuMEUNNpeb>).
 49. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *Nature* 2025;642:451-457. DOI: [10.1038/s41586-025-08869-4](https://doi.org/10.1038/s41586-025-08869-4).
 50. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open* 2024;7:e2440969. DOI: [10.1001/jamanetworkopen.2024.40969](https://doi.org/10.1001/jamanetworkopen.2024.40969).
 51. Wang Q, Yang Y, Qiu L, et al. A quantitative analysis of global AI medical studies: gaps in randomized controlled trials. *NPJ Digit Med* 2026;9:403. DOI: [10.1038/s41746-026-02698-z](https://doi.org/10.1038/s41746-026-02698-z).
 52. Cain CH, Davis AC, Broder B, et al. Quality assurance during the rapid implementation of an AI-assisted clinical documentation support tool. *NEJM AI* 2025;2(4). DOI: [10.1056/AIcs2400977](https://doi.org/10.1056/AIcs2400977).
 53. Kim E, Liu VX, Singh K. AI scribes are not productivity tools (yet). *NEJM AI* 2025;2(12). DOI: [10.1056/AIe2501051](https://doi.org/10.1056/AIe2501051).
 54. van Amsterdam WAC, Oberst M, Feng J, et al. Clinical trials for continuously monitored and updated AI systems. *Nat Med* 2026;32:1972-1974. DOI: [10.1038/s41591-026-04368-9](https://doi.org/10.1038/s41591-026-04368-9).
 55. Afshar M, Resnik F, Baumann MR, et al. A novel playbook for pragmatic trial operations to monitor and evaluate ambient artificial intelligence in clinical practice. *NEJM AI* 2025;2(9). DOI: [10.1056/AIdbp2401267](https://doi.org/10.1056/AIdbp2401267).
 56. Subasri V, Krishnan A, Kore A, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Netw Open* 2025;8:e2513685. DOI: [10.1001/jamanetworkopen.2025.13685](https://doi.org/10.1001/jamanetworkopen.2025.13685).
 57. Feng J, Xia F, Singh K, Pirracchio R. Not all clinical AI monitoring systems are created equal: review and recommendations. *NEJM AI* 2025;2(2). DOI: [10.1056/AIra2400657](https://doi.org/10.1056/AIra2400657).
 58. Luther J, Yilma H, Washington I. Companies expand AI health offerings, even as accuracy questions remain. KFF, April 23, 2026 (<https://www.kff.org/health-information-trust/companies-expand-ai-health-offerings-even-as-accuracy-questions-remain/>).
 59. U.S. Food and Drug Administration. FDA launches agency-wide AI tool to optimize performance for the American people. June 2, 2025 (<https://www.fda.gov/news-events/press-announcements/fda-launches-agency-wide-ai-tool-optimize-performance-american-people>).
 60. Li H, He X, Subbaswamy A, et al. Scaling medical device regulatory science using large language models. *NPJ Digit Med* 2026;9:221. DOI: [10.1038/s41746-026-02353-7](https://doi.org/10.1038/s41746-026-02353-7).
 61. Abramoff MD, Dai T, Zou J. Scaling adoption of medical AI — reimbursement from value-based care and fee-for-service perspectives. *NEJM AI* 2024;1(5). DOI: [10.1056/AIpc2400083](https://doi.org/10.1056/AIpc2400083).
 62. NHS England. Five years of a greener NHS: progress and forward look. September 29, 2025 (<https://www.england.nhs.uk/long-read/five-years-greener-nhs-progress-forward-look/>).
 63. Anthropic. Labor market impacts of AI: a new measure and early evidence. March 5, 2026 (<https://www.anthropic.com/research/labor-market-impacts>).
 64. Zhang Y, Ayush K, Qiao S, et al. SensorLM: learning the language of wearable sensors. June 10, 2025 (<https://arxiv.org/abs/2506.09108>). Preprint.