Review

Edward P. Hoffer*, Cornelius A. James, Andrew Wong and Sumant Ranji

Artificial intelligence and medical diagnosis: past, present and future

https://doi.org/10.1515/dx-2025-0111
Received August 1, 2025; accepted August 18, 2025;
published online September 10, 2025

Abstract: The NASEM report suggested that health information technology could reduce diagnostic error if carefully implemented. Computer-based diagnostic decision support systems have a long history, but to date have not had major impact on clinical practice. Current research suggests that AI-enabled decision support systems, properly integrated into clinical workflows, will have a growing role in reducing diagnostic error. The history, current landscape and anticipated future of AI in diagnosis are discussed in this paper.

Keywords: diagnosis; artificial intelligence; medical informatics

Introduction

The NASEM Report recognized that health IT could improve diagnosis and reduce diagnostic errors by enabling easy and timely access to information, facilitating communication among providers and between providers and patients and supporting clinical reasoning. Looking at the state of health IT at the time of the report, they noted that it more often hindered than helped. They recommended that health IT vendors work with the Office of the National Coordinator for Health IT and users to see that systems use good user interface, integrate with clinical workflow, offer clinical decision support and facilitate timely information flow.

Where do we stand 10 years later? Has health IT lived up to its promise?

The history of artificial intelligence in diagnosis

Despite the perception that the use of artificial intelligence (AI) in medicine began with the release of ChatGPT in November 2022, neither AI nor its application to medical diagnosis are new. Ledley and Lusted laid the foundation for computer-assisted diagnosis in their seminal 1959 paper that attempted to develop a mathematical model of how to best diagnose patients from their findings and were among the first to suggest the use of Bayes' theorem in this endeavor [1].

What is new is the enormous power of today's computers. IBM's first commercial computer, the 701, introduced in 1952, used vacuum tubes and could perform 16,000 operations per second. Hewlett Packard's FRONTIER supercomputer can perform 1.1 quintillion (that is 1.1 billion X 1 billion) operations per second. It is this power that makes current neural networks and generative AI possible.

Medical AI-enabled diagnostic decision support commonly takes one of three forms. The first diagnostic decision support systems (DDSSs) developed were "expert mimics," rule-based systems that tried to emulate the performance of skilled practitioners. This form of AI has the advantage of being transparent and easily explained but the disadvantage of needing to anticipate all possible situations.

Machine learning (ML) is the subset of artificial intelligence which uses algorithms that learn from data to make predictions.

A neural network (NN) is a type of ML that uses interconnected nodes in a layered structure that mimics the way brain neurons are connected. Neural networks are ideal for pattern recognition tasks on large samples of data, and have been widely used to examine images, such as skin lesions or retinal photographs.

***Corresponding author: Edward P. Hoffer**, MD, Laboratory of Computer Science, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA, E-mail: ehoffer@gmail.com
Cornelius A. James, Departments of Internal Medicine, Pediatrics and Learning Health Systems, University of Michigan Medical School, Ann Arbor, MI, USA
Andrew Wong, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA
Sumant Ranji, Division of Hospital Medicine, Department of Medicine, UCSF at San Francisco General Hospital, San Francisco, CA, USA; and Division of Clinical Informatics and Digital Transformation, Department of Medicine, UCSF, San Francisco, CA, USA

Large language models (LLMs), also referred to as generative AI, such as ChatGPT from OpenAI and Gemini from Google have come to dominate discussion about most medical applications. This form of ML is trained on vast amounts of textual data and can extract meaning from text and understand the relationships between words and phrases in human language.

"Expert mimic" DDSSs initially focused on limited domains, where essentially all possible relevant history, physical exam, laboratory and imaging findings were known, and the set of possible diagnoses was limited. Thyroid disease was a common target and multiple systems out-performed non-expert clinicians [2]. In 1972, De Dombal developed a DDSS focused on abdominal pain and reported that the system outperformed the most senior member of the clinical team to see the patient with an overall accuracy of 91.8 % compared to the clinicians' 79.6 % [3].

The narrow focus of these early efforts limited their value but paved the way for more general-purpose decision support.

Iliad, developed at the University of Utah, covered 650 diseases in 10 subspecialties of internal medicine in the mid-1980s [4]. It used values to represent the frequency of findings in patients with and without each disease and how prevalent each disease was in the community to produce a rank-ordered list of diseases. It was widely accepted in its own institution as a teaching tool but had little impact on practicing clinicians [5].

INTERNIST-1 was developed at the University of Pittsburgh beginning in 1972. Like Iliad, the system used the frequency with which a finding occurred in a disease and added an "evoking strength" indicating how strongly the presence of a finding suggested a disease [6, 7]. INTERNIST-1 and its successor, Quick Medical Reference (QMR) had some success outside the developers' institution. QMR was commercially distributed by First Databank between about 1989 and the early 2000s, when support and sales ceased.

Work on DXplain began at the Massachusetts General Hospital in the mid-1980s [8]. It used a similar paradigm as INTERNIST-1. An added function gave higher weight to less specific findings when findings from different organ systems co-existed. The system also guided the user by suggesting clinical or laboratory findings that would support or rule out leading diseases. The system has grown steadily and now contains descriptions of 2,690 diseases and 6,175 findings. Distributed on a subscription basis to medical schools, hospitals and group practices, DXplain has been shown to improve the diagnostic accuracy of medical residents [9] and to shorten the length of stay of patients admitted to hospital with complex conditions [10]. In early 2025, it out-performed leading large language models in diagnosing previously unpublished challenging clinical cases [11].

Work on Isabel began in 1999, focusing on pediatric diseases. Adult diseases were added in 2006. In addition to its own database of diseases and common presentations, Isabel performs a word search of medical texts to find good matches of findings entered and diseases. The user can copy a clinical scenario and paste it into Isabel, thus speeding data entry. It has been found to include the correct diagnosis somewhere on its list in over 90 % of test cases, and most closely mirrors modern LLMs [12, 13].

While rule-based systems can improve diagnostic accuracy, a factor limiting the impact of DDSSs on patient outcomes has been the need for clinicians to recognize they need help and seek consultation. Today's busy clinical environment emphasizes rapid patient turnover, and any tool that requires additional time from physicians faces a major hurdle. Added to that is the known over-confidence of most physicians, meaning they rarely see the need for diagnostic decision help, and it is not surprising that despite decades of development, no DDSS has yet demonstrated a major impact on reducing diagnostic error.

Current use of artificial intelligence for diagnosis – machine learning, neural networks, and large language models

Following the 1980s, interest in AI waned, resulting in decreased funding and limited growth in the field of diagnostic AI. The field was reinvigorated during the early 2000s with progress in ML. These advances led hospitals across the US to adopt ML models integrated into electronic health records (EHR). However, this adoption was not without controversy. For instance, the Epic sepsis model (ESM), was widely adopted by US hospitals, but several studies showed that the ESM performed significantly worse than reported [14, 15]. This raised important questions about the risks of inaccurate predictions (e.g., missed opportunities, clinician alert fatigue) and the need for local and national AI governance structures and more robust training methods.

Deep learning (DL), a subdomain of ML that includes two or more hidden processing layers within multilayered neural networks, ushered in a new paradigm of AI capable of learning patterns in very large data sets containing diverse data types. DL models have proven to be particularly useful for advanced image recognition and performing various natural language tasks. An early example of success was the

development of a DL model that detected diabetic retinopathy in retinal images with groundbreaking accuracy [16]. Additionally, DL models have shown the ability to mine clinical notes and other EHR data to accurately detect early heart failure [17]. IDx-DR, a DL model designed to diagnose diabetic retinopathy, became the first US Food and Drug Administration (FDA) approved autonomous AI system [18].

While DL models offered improved accuracy, their increasing complexity made it impossible to understand how a model arrived at its output in some cases (i.e., the AI "black box"). DL models may identify features in datasets associated with a diagnosis that either can't be identified by humans or that were not known to be associated with a diagnosis. Thus, explainability and transparency are ongoing concerns related to using AI for diagnostic purposes.

Despite these concerns, nearly 1,000 AI-enabled medical devices have been approved by the FDA, and many of these models play a role in the diagnostic process. FDA approvals are largely based on studies of model development and validation that report performance metrics, with limited relevance to real-world clinical practice, and studies that compare a model's performance of a diagnostic task to clinicians' performance of the same task (e.g., diagnosing breast cancer on mammogram images) [19]. However, it has become clear that a model's safety and effectiveness should not be determined by FDA approval alone. As an example, MelaFind, a model approved by both the FDA and the European Union to identify lesions suspicious for melanoma, was discontinued because of unnecessary biopsies, difficulty integrating the technology into dermatologists' workflows, and challenges with insurance coverage [20]. Furthermore, there are few randomized controlled trials (RCT) of AI-based interventions demonstrating efficacy in real-world clinical settings [21]. Despite the dearth of RCT-level evidence, AI continues to be deployed in clinical settings, often with limited impact on diagnostic safety.

Each medical specialty has been at least narrowly impacted by task-specific DL models (e.g., identifying skin cancer on smartphone images). While task-specific DL models offer increased flexibility and accuracy relative to their rules-based predecessors, they are still in a nascent phase. For example, a 2023 study using CPT codes from insurance claims data to quantify the adoption and usage of AI-based medical devices showed limited overall use in health care [22]. This study showed that clinicians billed most frequently for models designed to assist with the diagnosis of coronary artery disease, diabetic retinopathy, and liver conditions. Additionally, use of these tools was concentrated in specific geographic regions in the US, which highlights issues around availability and accessibility.

Foundation models such as OpenAI's GPT series have led to unprecedented optimism about the potential for diagnostic AI. Foundation models are based on deep neural networks that are trained on an extremely large corpus of data (e.g., the Internet, books) [22] and then adapted (or fine-tuned) for a broad range of downstream tasks These general-purpose models can generate new content based on their training data, hence the term generative AI (GenAI), and are far more flexible than task-specific DL models. GenAI, particularly large language models (LLM), are transforming every industry in society, including health care. OpenAI made headlines by gaining 1 million users in the first 5 days and 100 million users in the first two months of ChatGPT's launch. Other big tech companies, including Google, Apple, and Amazon have invested billions in GenAI, including health care-related GenAI. Additionally, startup companies like Open Evidence have developed LLMs designed to answer clinical questions at the point of care.

LLMs can interpret structured data in the EHR (such as vital signs and laboratory tests) as well as unstructured, text-based data to provide the necessary real-world context to arrive at a more accurate clinical diagnosis.

To understand the important of context, consider the "broken leg problem," first described by Meehl and colleagues in 1954 [23]. Consider a highly accurate AI model designed to predict attendance at a weekly movie, and a devoted attendee who attends weekly. One day, this individual suffers a fractured femur. While a human would immediately recognize that this would lead to her absence, a model not trained to evaluate for broken limbs would fail to adjust its prediction. Many critics cite "broken leg" events as a reason why even highly accurate automated diagnostic models underperform in real-world clinical settings. Since the range of such possible events is limitless, encoding every possibility into a structured algorithm is infeasible.

Unlike previous diagnostic systems, LLMs can handle such events by integrating unstructured language-based data from medical and clinical vignettes to inform the automated diagnostic process. LLMs have been shown to understand a wide array of human concepts and are much better suited to navigate these events and their impact on clinical diagnosis and prediction [24].

This clinical contextualization is not limited to rare events. Clinicians understand that structured EHR data (e.g., vital signs and laboratory results) should not be analyzed in a vacuum; instead, these findings are interpreted within a broader clinical context that includes the patient's tempo of disease progression, accompanying symptoms, baseline risk, and socioeconomic background. LLMs computationally integrate this same information from the patient's history to help guide clinical diagnosis.

Finally, LLMs can quickly integrate large quantities of text-based data from clinical notes and outside hospital records, a task that physicians often struggle with due to time constraints [25]. LLMs have demonstrated effectiveness at data processing tasks critical to clinical diagnosis, including electronic health record summarization, medical database search, and large-scale data analysis [26, 27]. They have also shown excellent performance with administrative tasks such as generating responses to patient portal messages [28] and clinical documentation [29].

Studies have shown that LLMs are as good or better than clinicians when generating an accurate differential diagnosis or identifying the correct diagnosis in simulated complex clinical cases [30–32]. Another study showed that GPT-4 alone performed better on diagnostic tasks compared to physicians using GPT-4 or traditional point-of-care resources (e.g., UpToDate or Google) [32]. However, LLMs have not performed as well on pediatric diagnostic challenges. A recent study showed that GPT-4 had an error rate of 83 % on *New England Journal of Medicine* and *Journal of the American Medical Association Pediatrics* case challenges [33]. While the discrepancy between GPT-4's diagnostic accuracy on adult vs. pediatric cases may be due to factors such as representativeness of pediatric data in training datasets, this shows the potential performance limitations of these models in specific populations.

A common critique of LLMs is that they were originally developed to be next-word predictors, or chatbots whose primary purpose is to identify the next word in a conversation. While this is true for traditional LLMs (e.g. OpenAI's GPT, Anthropic's Claude, Meta's Llama), a branch of LLMs known as reasoning models (e.g. OpenAI's o1/o3-mini, Deepseek's R1) are specifically trained to handle complex reasoning tasks through reinforcement learning. Unlike traditional LLMs which are designed to quickly interpret and generate human language akin to "system 1" thinking in humans, reasoning models are designed to perform logical, stepwise thinking to solve reasoning tasks, more like "system 2" thinking. These models have far outperformed previous LLMs in multiple benchmarks for logical deduction, multi-step reasoning, and pattern recognition [34, 35]. Current reasoning models have primarily been trained to excel in math and computer science problems, but the development of reasoning models specific to clinical diagnosis are already underway and may prove to be the final push that enables AI to surpass humans in diagnostic ability.

Future directions for clinical diagnosis in the era of AI

AI is already being widely implemented in clinical care, particularly in fields such as radiology and pathology. A recent survey reporting that about 65 % of physicians have used GenAI for clinical purposes [36]. Depending upon the diagnostic task, future clinicians will likely interact with AI across a continuum of diagnostic capacities. At one end of the spectrum AI may play an assistive role, and at the other end AI may act as a "copilot," generating a differential diagnosis for complex clinical cases, or identifying and classifying abnormalities on pathology slides. Though real-world data is still lacking, recent technological advancements have addressed many of the diagnostic performance deficiencies of earlier models.

Despite these improvements, we have yet to capitalize on the potential benefits of diagnostic AI. Few studies have shown that AI-based clinical decision support systems improve clinician diagnostic performance [37]. While it is unlikely that AI models will achieve diagnostic superiority to human experts in their next iteration, successive improvements may achieve this goal over the next several years. Barriers to the widespread adoption of AI in clinical medicine will probably not be purely technical given the rapidity with which LLMs have already improved. Instead, factors that have affected the uptake of earlier iterations of diagnostic decision support systems may also limit more widespread use of AI. To reach their potential, AI applications will need to support the patient's and clinician's journey through the diagnostic process. This will involve addressing several key questions relating to how AI is implemented, evaluated, and optimized in real-world clinical practice.

First, is AI *decision support* or a *decision maker*? Diagnostic AI applications are generally configured as decision support systems, offering a "virtual second opinion" or additional guidance after a clinician has determined the initial working diagnosis. When the AI application should prompt the clinician to consider other diagnoses remains unknown. Relying on clinicians to "opt in" to using AI decision support may result in underutilization, as clinicians would need to recognize when they have a diagnostic challenge and be willing to take additional steps to generate and integrate AI recommendations. However, an AI that pushes diagnostic prompts to clinicians risks contributing to alert fatigue or creating unnecessary interruptions – both of which are known patient safety risks [38, 39].

Current diagnostic AI applications should not be used to generate stand-alone diagnoses that lead directly to treatment plans. That said, AI is now being used to assist in diagnosis of time-sensitive, highly morbid conditions such as acute ischemic stroke [40]. In these situations, emerging evidence suggests that clinicians may treat the AI output as an initial working diagnosis – placing the AI closer to the center of the diagnostic process. This represents an inversion of the usual diagnostic workflow, where the clinician would gather data and formulate an initial hypothesis before seeking additional information through testing and decision support. Further research is required to determine which settings are best suited for AI to achieve superior diagnostic accuracy and to identify the potential patient safety risks associated with such a profound change.

Second, how should AI applications be evaluated before they are integrated into clinical practice? Studies of AI using simulated cases demonstrate tremendous potential, but it remains unclear whether AI systems alone will ever truly outperform AI-augmented humans in real-world clinical diagnosis, or when this may occur. Evaluation of AI tools must focus on clinical outcomes, rather than simply measuring diagnostic reasoning or diagnostic accuracy [41, 42]. Though it has been reported that LLMs can outperform clinicians at diagnosis, the example of computer chess engines may provide an informative analogy for how AI in clinical practice will evolve. When the supercomputer Deep Blue defeated the world chess champion Garry Kasparov in 1997, many in the chess community declared the death of the game. For the next 15 years, however, the field was dominated by computer-assisted human players rather than computers alone. Human experts continued to discover flaws in computer-based play and develop novel ways to use chess engines to push the boundaries of old and new strategies. The same will likely apply to clinical diagnosis in actual practice, even after AI surpasses the threshold of human-level performance in simulated scenarios.

Finally, what are the risks of AI-based diagnostic applications? Regardless of the overall diagnostic performance of AI algorithms, these systems will inevitably have weaknesses requiring human oversight and intervention. This is particularly true for rare diseases with few verified cases that can be used to train AI models and for diseases that have abnormal geographic or demographic distributions. Recent research also indicates that LLMs have similar biases in their clinical reasoning as human diagnosticians. A crucial question for more advanced reasoning models will be whether they can be trained to mitigate these biases. As AI is increasingly integrated into clinical practice, AI developers and health care systems will need to monitor the performance of these applications and assess for model drift and deterioration in performance over time.

We can be sure that AI systems will continue to improve, making it likely that human or system level factors will be the most significant barriers to realizing the diagnostic potential of AI. To realize the full potential of AI-based diagnostic technologies, human, system, and technology level barriers to and facilitators of effective human-AI teaming must be identified and addressed. Thus, an essential next step will be implementation studies that ensure that diagnostic AI is user-centered and compatible with clinical workflows, and that health care providers and patients are empowered to successfully leverage these technologies to achieve diagnostic excellence. The true value of diagnostic decision support help will only be realized when such systems are embedded in the EHR and work in the background, telling physicians when a serious disease may have been overlooked.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission and shared equally in conception, writing and revision.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: Dr. Hoffer is involved in the maintenance of DXplain, which is mentioned in the paper. He derives no income related to how well or poorly revenue is received by the project. All other authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

1. Ledley R, Lusted L. Reasoning foundations of medical diagnosis. Science 1959;130:9–21.
2. Taylor TR, Shields S, Black R. Study of cost-conscious computer-assisted diagnosis in thyroid disease. Lancet 1972;2:79–83.
3. de Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J 1972;2: 9–13.
4. Warner HR. The iliad program: an expert computer diagnostic program. Med Pract Manag 1992;8:123–8.
5. Lincoln MJ, Turner CW, Haug PJ, Warner HR, Williamson JW, Bouhaddou O, et al. Iliad training enhances medical students' diagnostic skills. J Med Syst 1991;15:93–110.
6. Miller RA. A history of the INTERNIST-1 and quick medical reference (QMR) computer-assisted diagnosis projects, with lessons learned. Yearb Med Inform 2010:121–36. https://doi.org/10.1055/s-0038-1638702.

7. Miller RA, Pople HEJ, Myers JD. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 2010;307:468–76.
8. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. JAMA 1987;258:67–74.
9. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, Hernandez-Torres I, Rivera-Lopez I, Spicer T, et al. Diagnostic accuracy in family medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. Diagnosis (Berl) 2018;5:71–6.
10. Elkin PL, Liebow M, Bauer BA, Chaliki S, Wahner-Roedler D, Bundrick J, et al. The introduction of a diagnostic decision support system (DXplain^{TM}) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging diagnostic related groups (DRGs). Int J Med Inf 2010;79:772–7.
11. Feldman MJ, Hoffer EP, Conley JJ, Chang J, Chung JA, Jernigan MC, et al. Dedicated AI expert system vs generative AI with large language model for clinical diagnoses. JAMA Netw Open 2025;8:e2512994.
12. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation. Arch Dis Child 2003;88:408–13.
13. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 2008;23:37–40.
14. Wong A, Otles E, Donnelly JP, Krumin A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021; 181:1065–70.
15. Lyons PG, Hofford MR, Yu SC, Michelson AP, Payne PRO, Hough CI, et al. Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. JAMA Intern Med 2023;183:611–2.
16. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Navayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.
17. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inf Assoc 2017;24:361–70.
18. Chang MF. Artificial intelligence getting smarter: innovations from the vision field. NIH Director's Blog; 2022. Available from: https://directorsblog.nih.gov/tag/idx-dr/ [Accessed 27 May 2025].
19. Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. JAMA 2025;333: 241–7.
20. Venkatesh KP, Kadakia KT, Gilbert S. Learnings from the first AI-enabled skin cancer device for primary care authorized by FDA. NPJ Digit Med 2024;7:1–4.
21. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open 2022;5:e2233946.
22. Wu K, Wu E, Theodorou B, Liang W, Mack C, Glass L, et al. Characterizing the clinical adoption of medical AI devices through U.S. insurance claims. NEJM AI 2024;1:AIoa2300030.
23. Bommasani R, Hudson DA, Adeli E, Altman R, Aroras S, von Ary S, et al. On the opportunities and risks of foundation models. 2022. https://doi.org/10.48550/arXiv.2108.07258.
24. Dawes RM, Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. Science 1989;243:1668–74.
25. Patel R, Pavlick E. Mapping language models to grounded conceptual spaces. In: Proceedings of the international conference on learning representations; 2022. Available from: https://openreview.net/forum?id=gJcEM8sxHK [Accessed 24 Mar 2025].
26. Holmgren AJ, Adler-Milstein J, Apathy NC. Electronic health record documentation burden crowds out health information exchange use by primary care physicians. Health Aff (Millwood) 2024;43:1538–45.
27. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30: 1134–42.
28. Guevara M, Chen S, Thomas S, Chaunzwa TI, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med 2024;7:1–14.
29. Garcia P, Ma SP, Shah S, Smith M, Jeong Y, Devon-Sand A, et al. Artificial intelligence–generated draft replies to patient inbox messages. JAMA Netw Open 2024;7:e243201.
30. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal Innov Care Deliv 2024;5:CAT–23.0404.
31. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Intern Med 2023;183:1028–30.
32. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330: 78–80.
33. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024;7:e2440969.
34. Barile J, Margolis A, Cason G, Kim R, Kalash S, Tchacomas A, et al. Diagnostic accuracy of a large language model in pediatric case studies. JAMA Pediatr 2024;178:313–15.
35. Li ZZ, Zhang D, Zhang ML, hang H, Song X, Zhang R, et al. From system 1 to system 2: a survey of reasoning large language models. 2025. https://doi.org/10.48550/arXiv.2502.17419.
36. DeepSeek AI, Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. 2025. https://doi.org/10.48550/arXiv.2501.12948.
37. Augmented intelligence research – physician sentiments around the use of AI in health care: motivations, opportunities, risks, and use cases. Shifts from 2023–2024. American Medical Association; 2025. Available from: https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf
38. Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open 2021;4:e211276.
39. Sloane JF, Donkin C, Newell BR, Singh H, Meyer AND. Managing interruptions to improve diagnostic decision-making: strategies and recommended research agenda. J Gen Intern Med 2023;38:1526–31.
40. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inf Assoc 2006;13: 138–47.
41. D'Adderio L, Bates DW. Transforming diagnosis through artificial intelligence. NPJ Digit Med 2025;8:54.
42. Rodman A, Zwaan L, Olson A, Manrai AK. When it comes to benchmarks, humans are the only way. NEJM AI 2025;2:AIe2500143.