DIAGNOSTIC EXCELLENCE

Decoding Artificial Intelligence to Achieve Diagnostic Excellence
Learning From Experts, Examples, and Experience

Jonathan H. Chen, MD, PhD
Stanford Center for Biomedical Informatics Research, Division of Hospital Medicine, Stanford University, Stanford, California.

Gurpreet Dhaliwal, MD
Department of Medicine, University of California, San Francisco; and Medical Service, San Francisco Veteran Affairs Medical Center, San Francisco, California.

Daniel Yang, MD
Gordon and Betty Moore Foundation, Palo Alto, California.

Clinical decision support systems that use artificial intelligence (AI) to improve diagnostic accuracy, efficiency, and safety have long been aspirational goals for computer scientists and clinicians. Yet diagnostic AI development has seen multiple cycles of inflated peaks of expectations followed by troughs of disillusionment. Clinicians are understandably wary of embracing new diagnostic AI solutions without understanding how they work and relate to their existing practice.

Diagnostic AI refers to a broad range of applications that use learning strategies that mimic human approaches to learning. When clinicians understand the underlying mechanisms of diagnostic AI, they can become informed users of these tools, appreciating both their advantages and limitations. This Viewpoint outlines 3 learning methods that form the basis of many diagnostic AI systems—learning from experts, examples, and experience—and their parallels to clinicians' existing approaches to learning.

Learning From Experts

Some of the earliest examples of AI programs were rules-based expert systems for clinical diagnosis in which human knowledge was encoded into computer-executable rules. For example, the program MYCIN was developed in the 1970s to assist clinicians in the diagnosis and treatment of bacterial infections.[1] The program queried cli-

> As clinicians understand the ways in which diagnostic AI systems develop "intelligence," they may…envision how to combine human and artificial intelligence to achieve diagnostic excellence better than either can alone.

nicians about their patient's case (eg, any growth in culture? Recent urologic procedure?) and used the answers to generate a differential diagnosis of likely bacteria.

An advantage of such rule-based AI systems is the relative ease of understanding the logic of how a system arrives at its diagnostic conclusions (eg, *Staphylococcus aureus* is not the cause of the infection *because* a rule indicates that *Staphylococcus* is a Gram-positive organism, which does not match the provided case data).

Despite the accuracy of rules-based systems, clinicians often found that they were impractical to use and had very narrow capabilities. The promise of general expert AI systems that assist clinicians across a broad range of conditions has fallen out of favor because the time-intensive

effort of human experts to manually encode thousands of rules is poorly suited for a complex adaptive field like medicine in which rules can contradict each other and regularly become obsolete in the face of new knowledge.[2]

Training these early AI systems by encoding knowledge rules resembles the learning process for physicians early in their careers. Medical students learn from their teachers (experts) and mimic their diagnostic thinking and rules (eg, "if fever, cough, and pulmonary infiltrate, then diagnose pneumonia").

Learning From Examples

Most recent popular applications of diagnostic AI rely on supervised machine learning, which discerns patterns from example cases labeled by humans with the "correct" answer. Rather than being programmed by experts, these algorithms can write their own rules. For example, by exposing machine learning algorithms to thousands of retinal images that include cases of diabetic retinopathy labeled by ophthalmologists, these systems can make the diagnosis in future images without being told what to look for.[3]

Advantages of AI approaches that learn by example include the ability to achieve expert-level performance on a wide variety of diagnostic tasks by learning from thousands of case examples more rapidly and consistently than humans have the capacity for. Such algorithms can estimate risks of future events (eg, myocardial infarction) using subtle diagnostic features (eg, retinal image patterns) that may be invisible or undiscovered by human experts.[4]

The disadvantage of such systems is that their diagnostic logic is often inscrutable to clinicians, which has introduced concerns about the "black box" nature of these diagnostic AI tools. Algorithms can now examine an electrocardiogram (ECG) and identify the signatures of previous atrial fibrillation episodes in a currently normal sinus rhythm tracing.[5] If the algorithm cannot explain how it makes this diagnosis, will physicians be willing to act on it and prescribe anticoagulation?

A broader limitation for many applications of supervised learning is the need for large amounts of manually labeled training data. This is not only resource intensive; it relies on the accuracy of human-designated labels, which may be biased or have poor interrater reliability (eg, diagnosing urinary tract infection vs asymptomatic bacteriuria).

This learning by example approach mirrors physicians' immersion during residency, a phase of extensive patient encounters and supervised learning. After

Corresponding Author: Jonathan H. Chen, MD, PhD, Stanford Center for Biomedical Informatics Research, Division of Hospital Medicine, Stanford University, 3180 Porter Dr, B132, Palo Alto, CA 94304 (jonc101@stanford.edu).

clinicians have seen enough patients with a given condition, their reliance on textbook rules is replaced by recognition based on multiple examples. A resident may learn descriptions of ischemic ST segment changes and benign early repolarization on the ECG, but will more effectively learn the difference by seeing many examples of ECG tracings that are labeled by a supervising clinician with the correct diagnosis. Through this process, the trainee will accurately recognize the difference in future instances of ST segment changes, even if the textbook rules are forgotten.

Learning From Experience

Game-playing AI algorithms that have defeated world champion human players in complex strategy games, including poker, chess, Go, and StarCraft, demonstrate the potential of reinforcement learning. These game-playing AI algorithms can learn from experience, playing millions of simulated games against themselves to experiment with different strategies.[6] This allows the algorithms to explore and exploit unconventional choices against objective win conditions, offering the potential to discover novel moves that advance the state-of-the-art.

Analogously, given a series of steps along the diagnostic pathway, a reinforcement learning approach could try different testing sequences to determine which lead to the most timely, accurate, and efficient diagnosis. For example, such an approach could simulate the thousands of different workups of a solitary pulmonary nodule, assessing the influence of different diagnostic moves both within and beyond the conventional sequencing of history, examination, laboratory testing, imaging, pathology, and genetic testing.

The key limitation to reinforcement learning approaches is that their best results depend on the ability to reliably simulate cases and safely conduct unlimited experiments with clear outcomes. Clinical practice lacks the key properties of games with their static rules, verifiable outcomes (win-lose), and instant feedback.[7] In the absence of reliable clinical simulation, embedded comparative effectiveness experiments between existing standards of care in a learning health system may offer the closest framework for explicit reinforcement learning in medicine.[8,9]

This experiential approach reflects the learning process of clinicians as they become independent practitioners and can

---

Box. Key Points for Diagnostic Excellence

- Diagnostic artificial intelligence (AI) technologies represent a heterogeneous set of learning approaches that mimic human learning from experts, examples, and experience.
- Understanding this analogy can help clinicians demystify the "black box" of diagnostic AI and enable them to become informed users of these systems.
- AI systems will eventually become important tools that augment the diagnostic process. Ideally, these tools will offload computational and data-intensive work while enabling clinicians to focus on tasks they are uniquely well-suited for, including history taking, communicating uncertainty, and understanding the patient's context.

---

challenge conventional wisdom around diagnostic pathways by exploring variations and tracking their outcomes. For instance, a clinician with many years of experience may understand that joint aspiration and detection of urate crystals in synovial fluid is the formal way to diagnose gout, but may have also learned through experimentation that physical examination, x-ray findings, and measurement of serum uric acid and C-reactive protein levels can be integrated to make the same diagnosis with greater efficiency and similar accuracy.

Conclusions

AI-trained systems cannot completely execute the diagnostic process, which involves human interactions, judgments, and social systems that are beyond what computers can model. AI systems will nonetheless become important tools—just like laboratory tests or imaging studies—in the increasingly complex quest to diagnose patients' health problems (**Box**).

Diagnostic AI systems learn by mimicking experts, acquiring examples, or conducting experiments, much as clinicians do throughout their careers. As clinicians understand the ways in which diagnostic AI systems develop "intelligence," they may recognize the strengths and limitations of their own learning practices and envision how to combine human and artificial intelligence to achieve diagnostic excellence better than either can alone.

REFERENCES

1. van Melle W. MYCIN: a knowledge-based consultation program for infectious disease diagnosis. *Int J Man Mach Stud*. 1978;10:313-322. doi:10.1016/S0020-7373(78)80049-2

2. Martínez García L, Sanabria AJ, García Alvarez E, et al; Updating Guidelines Working Group. The validity of recommendations from clinical guidelines. *CMAJ*. 2014;186(16):1211-1219.

3. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA*. 2018;320(11):1107-1108.

4. Diaz-Pinto A, et al. Predicting myocardial infarction through retinal scans and minimal personal information. *Nat Mach Intell*. 2022;4:55-61. doi:10.1038/s42256-021-00427-7

5. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm. *Lancet*. 2019;394(10201):861-867.

6. Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*. 2018;362(6419):1140-1144.

7. Gottesman O, Johansson F, Komorowski M, et al. Guidelines for reinforcement learning in healthcare. *Nat Med*. 2019;25(1):16-18.

8. Angus DC. Fusing randomized trials with big data: the key to self-learning health care systems? *JAMA*. 2015;314(8):767-768.

9. Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med*. 2019;381(12):1175-1179.