NEJM AI

NEJM AI 2026;3(2)
DOI: 10.1056/AIp2500808

PERSPECTIVE

Designing Clinically Useful AI: A Blueprint for Impact

Shyon Parsa, M.D.,[1,2] Timothy Keyes, Ph.D.,[3] Dev Dash, M.D., M.P.H.,[1,4] Danton Char, M.D.,[5] Michelle M. Mello, J.D., Ph.D.,[6,7] Alison Callahan, Ph.D.,[8] Margaret Ann Smith, M.B.A.,[9] Sinjin Lee, M.D.,[10] Thomas Wang, Ph.D.,[8] Heidi Salisbury, C.N.S.,[11] Shinichi Goto, M.D., Ph.D.,[12] Vicki Parikh, M.D.,[11] Kenneth W. Mahaffey, M.D.,[1,13,14] Michael Salerno, M.D., Ph.D.,[15] Euan A. Ashley, D.Phil.,[1,11,13] Nigam H. Shah, Ph.D.,[1,3] and Sneha S. Jain, M.D., M.B.A.[2,13]

Received: July 19, 2025; Revised: October 29, 2025; Accepted: November 20, 2025; Published: January 22, 2026

Abstract

Most artificial intelligence (AI) tools in health care are evaluated on statistical performance for diagnostic accuracy alone, which often fails to account for the realities of the clinical systems into which they may be deployed. This disconnect has contributed to a proliferation of AI tools that perform well in development but fail to gain traction or generate meaningful impact in clinical use. We propose the use of health AI target product profiles, which specify the performance thresholds an AI tool must meet to produce benefit within a specific care setting, accounting for workflow, capacity, and utility trade-offs. Using hypertrophic cardiomyopathy (HCM) detection as an example, we simulate the performance of an AI-augmented clinical program across a range of AI tool characteristics and health care resource constraints to identify the conditions under which clinical value could be realized. Health AI target product profiles can guide AI tool development, inform AI tool selection if multiple AI tools have already been developed, guide implementation strategies for AI-augmented programs, and prevent investment in AI tools that are unlikely to create value. Ultimately, this approach offers a proactive and context-driven pathway for designing clinically useful AI that can empower health systems, patients, and providers as active members of the AI design process. (Funded by the American Heart Association.)

Introduction

Artificial intelligence (AI) in health care is often developed and evaluated in ways that overlook the realities of clinical practice. These tools are often assessed under idealized assumptions and in contexts disconnected from the constraints of real-world workflows.[1] Most evaluations emphasize diagnostic accuracy, using metrics such as sensitivity, specificity, and the area under the receiver operating characteristic curve,[2-4] which are essential, setting-independent properties of any diagnostic tool. However, even AI technologies with high diagnostic accuracy may fail to deliver real-world value unless they also perform well under local conditions. Predictive accuracy (positive predictive value and negative predictive value)[5] and clinical usefulness — defined as the model's impact on

Shyon Parsa and Timothy Keyes contributed equally to this article.

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

Sneha S. Jain can be contacted at snehashahjain@stanford.edu or at the Center for Academic Medicine, Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, 453 Quarry Road, Palo Alto, CA 94304.

outcomes, workflow, and cost — are equally important. Overemphasis on diagnostic accuracy without assessing predictive accuracy or usefulness has contributed to the gap between development and meaningful deployment of AI tools in practice.[6,7] A growing body of work now emphasizes the importance of evaluating AI tools in the context of real-world constraints, including workflow integration, resource availability, financial sustainability, and benefit–harm trade-offs. These considerations have shaped frameworks such as the Fair, Usable, and Reliable AI Model assessment process.[8-10] However, to truly close the gap between model development and clinical value, real-world factors must guide AI design from the outset, not just in its evaluation. An important question needs to be answered early in the planning and implementation process: How well *should* an AI-augmented clinical program work for this problem, in this setting, under real-world constraints, in order to create value?

We propose the use of health AI target product profiles (TPPs), inspired by pharmaceutical and diagnostic test development pipelines, to align design with clinical needs, as a framework to guide model development.[11-14] A health AI TPP defines the minimum and optimal performance characteristics needed for an AI-augmented clinical program to meaningfully improve clinical practice. By systematically simulating real-world clinical outcomes across a broad range of hypothetical AI model performance characteristics — and explicitly incorporating constraints such as clinical capacity and workflow limitations — we aim to reveal the specific performance thresholds required for meaningful clinical benefit. This structured approach facilitates head-to-head comparison of tools, prioritizes investment in the AI tools most likely to generate impact, and embeds stakeholder needs into the design process. If health systems proactively simulate these real-world conditions, we could shift the focus from evaluating one-off AI tools for specific use cases to defining up front how well any model and AI-augmented workflow need to perform in context.[15] Health AI TPPs can empower clinicians, patients, and health systems to proactively guide and codevelop impactful AI tools that are fit for purpose.

Health AI TPPs consider three critical components that determine whether an AI model can deliver value in a proposed clinical context: workflow design, capacity constraints, and outcome utility trade-offs. Workflow design examines the baseline pathway, or non–AI-guided current state, for diagnosis or treatment, including how the workflow may be modified by AI integration. Capacity constraints consider the current availability of clinicians and services vital to the baseline and AI-augmented diagnostic workflow. Outcome utility is a quantification of the benefit of accurate early diagnosis and potential misclassification harm. Health systems can incorporate financial projection assessments to assess for sustainability and ethical evaluations to guide workflow design and postdeployment performance monitoring as needed.[6] Together, these components determine whether an AI model's theoretical performance might translate into practical clinical impact.

Hypertrophic Cardiomyopathy: A Use Case

We developed a TPP for hypertrophic cardiomyopathy (HCM) detection to guide the design and implementation of HCM AI tools (Fig. 1). HCM is a clinical condition in which early diagnosis substantially alters management and outcomes.[16] Currently, HCM is underdiagnosed, but a number of AI tools have been developed with the goal of identifying patients with HCM who are missed by usual care.[17-20] The diagnosis of HCM depends on confirmatory imaging and cardiology subspecialists, both of which are limited in many health systems.[21,22] AI tools have demonstrated impressive area under the receiver operating characteristic curve (>0.95) in identifying HCM from screening electrocardiogram data.[23,24] In practice, there are significant concerns regarding operationalizing HCM AI tools, including determining who is responsible for responding to AI-guided flags for high-risk patients and the capacity for diagnostic assessment and specialized care.[25,26] The value of an AI model is not intrinsic; rather, it is a function of multiple components within the clinical environment into which it is deployed.

To address these real-world uncertainties, we used A Python Library for Usefulness Simulations of Machine Learning Models in Healthcare (APLUS), a discrete-event simulation tool, to generate a TPP for AI-guided HCM screening under realistic clinical constraints.[27] We used Stanford Health Care (SHC), a multihospital health system in California, as the setting for the simulation. Initially, we used real-world capacity constraints in SHC's diagnostic workflow for HCM patients — specifically, the weekly appointment capacities of SHC's HCM triage and HCM clinic (see Fig. 1A in the Supplementary Appendix, and code available at https://github.com/som-guideai/health-ai-tpp) — to simulate the effects of using an AI-guided workflow across multiple combinations of the model's sensitivity and specificity. Sensitivity and specificity values ranged from modest (sensitivity, ~0.5; specificity, ~0.8) to very high (sensitivity,

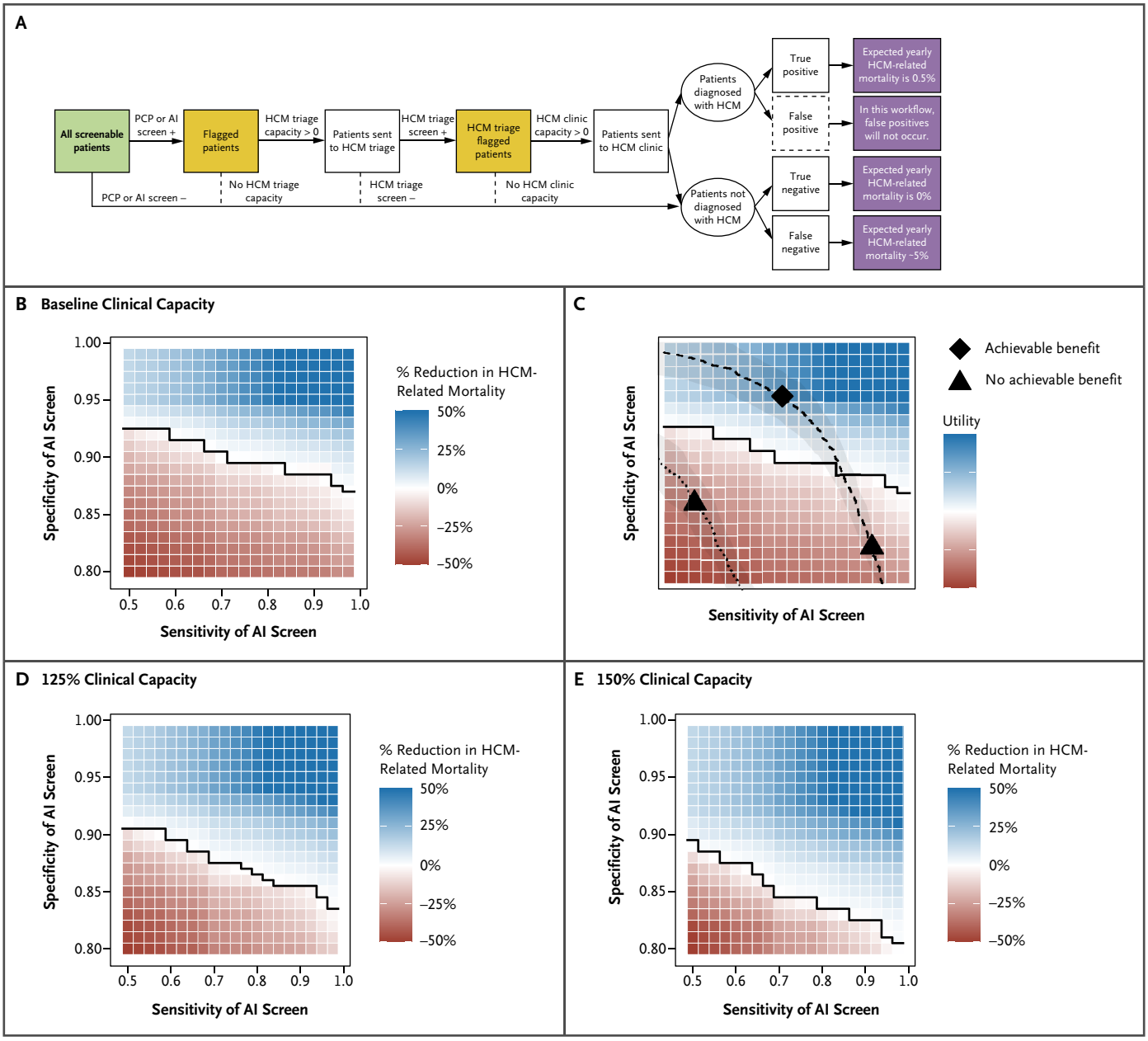

Figure 1. Health AI Target Product Profiles: Simulated Impact of AI-Guided Screening on Hypertrophic Cardiomyopathy–Related Mortality under Varying Workflow Capacities.

Panel A shows a workflow diagram illustrating the simulated hypertrophic cardiomyopathy (HCM) screening and management pathway. The process begins with all screenable patients (light green), who may be flagged by a primary care provider or artificial intelligence (AI)–guided screening, and then progress through stages constrained by triage and clinic capacity (bright yellow). Final outcomes (violet) reflect the expected usefulness of each diagnostic category: True positives receive active management (0.5% annual mortality), true negatives require no treatment (0% mortality), and false negatives remain untreated (up to 5% mortality), while false positives are assumed not to occur in this workflow. Dashed lines indicate pathways where patients are not advanced due to insufficient capacity at triage or clinic stages. Panel B shows a heat map illustrating the relative (percentage) reduction in simulated HCM-related mortality with AI-guided screening, relative to the primary care provider referral–driven baseline, across a grid of simulated AI screening sensitivities and specificities using Stanford Health Care's stated (baseline) HCM triage and HCM clinic capacities. The grid's color scale represents the relative change in HCM-related mortality in each simulation, with brick red indicating an increase in mortality, blue indicating a decrease in mortality, and white indicating minimal change. The solid black line represents the zero-difference curve along which simulated HCM-related mortality is unchanged relative to the current-state, non–AI-guided workflow. Panel C is an illustrative example of using a health AI target product profile (TPP) to evaluate whether two proposed AI models would yield an achievable patient benefit in a specific clinical context.

0.975; specificity, ~0.99), allowing us to profile the performance characteristics that an AI model would require to achieve real-world benefit compared with the current standard of care within SHC.

Our simulations revealed that only AI tools with exceptionally high specificity (~0.925) achieved reductions in HCM-related mortality relative to the clinical current state under the known capacity constraints of SHC's HCM workflow (Fig. 1B and 1C), as lower-specificity AI tools overwhelmed clinical capacity with excessive false positives. This meant that, to ultimately decrease HCM-related mortality, we would require a series of AI and/or human screens to improve overall specificity (e.g., an electrocardiogram screen, an echo screen, and possibly a large language model–augmented HCM triage), investment in additional capacity, or a combination of both actions.[28]

We performed simulations after expanding clinical capacity incrementally by 25% and 50%. These simulations demonstrated a notable improvement in the tool's clinical value. At an increased capacity of 50%, AI tools with a combined specificity as low as 0.85 were able to yield reductions in HCM-related mortality (Fig. 1D and 1E). These findings underscore the dynamic nature of TPPs and their adaptability to different resource availability and workflow constraints within a proposed deployment.

A Blueprint before the Build

Target product profiles serve as a guide for clinical systems deploying AI tools, specifying the diagnostic accuracy required of an AI-augmented clinical program under specific workflow constraints to produce a net benefit. For example, a system with limited triage resources might demand high specificity to avoid overwhelming capacity. However, one with ample diagnostic throughput might prioritize sensitivity to catch more early-stage disease. Furthermore, TPPs can guide decision-making around the impact of addressing capacity constraints. As certain capacity bottlenecks are addressed at the system level, such as hiring more triage staff, incorporating AI or non-AI triage tools, or increasing available imaging modalities, the model specifications needed can be easily updated.[28]

Beyond HCM, this framework generalizes to other use cases. In AI-augmented colon cancer screening, for example, the burden of false positives must be weighed against colonoscopy or biopsy availability as well as the economic toll of false positives.[29,30] Although our HCM workflow represents an ambulatory workflow, TPPs are flexible to account for varying clinical environments. They can incorporate estimates for additional behavioral and process measures, such as end-user uptake, to fully capture possible clinical usefulness. In AI tools that aim to predict inpatient decompensation and a potentially proactive transfer to the intensive care unit, the TPP must include considerations of critical care bed and nursing staff limitations, as well as the likelihood that clinicians would react to a notification, to understand the true possible impact.[31,32] Ultimately, many AI tools may never be successfully deployed, not because the algorithms are flawed, but because they fail to address a real need or that the ecosystem cannot support the interventions proposed. In this way, TPPs are not only tools for deployment, but are also safeguards against investing in the creation of AI tools that, while able to achieve high diagnostic accuracy, are not useful.

Meaningful evaluation of AI in health care must move beyond only diagnostic accuracy performance metrics and must adopt a proactive approach to identify the true needs in systems where AI can advance care. By developing health AI TPPs, health systems can define the characteristics and performance thresholds necessary for AI to deliver meaningful, real-world benefits. Health AI TPPs position health

Figure 1. (*Continued*) Using the health AI TPP from Panel B, we projected each model's operating curve onto the simulated usefulness grid to visualize how its trade-off between sensitivity and specificity translates into achievable or unachievable clinical benefit under the modeled capacity constraints. Overlaid curves represent the operating characteristics of two example simulated models across probability thresholds. One model (dashed line; area under the receiver operating characteristic curve=0.95) falls largely within the (blue) region of achievable clinical benefit under realistic capacity constraints, depending on the probability threshold chosen for the deployment. The other model (dotted line; area under the receiver operating characteristic curve=0.85) falls entirely in the (brick red) region of unachievable clinical benefit, indicating that no probability threshold is likely to result in a successful deployment. Markers along each curve denote example probability thresholds for each model that would allow for an achievable benefit or no achievable benefit, respectively. Gray ribbons represent the simulated 95% confidence intervals around each point along the curve. Panels D and E show additional heat maps, similar to those in Panel B, but in simulations in which the HCM clinic and triage capacity was increased to 125% of baseline (Panel D) and 150% of baseline (Panel E). These plots illustrate how health AI TPPs can capture the dynamic interplay between a model's diagnostic accuracy and the clinical characteristics of the specific environment into which the model is deployed. AI denotes artificial intelligence; HCM, hypertrophic cardiomyopathy; and PCP, primary care provider.

systems, clinicians, and patients as active participants in shaping AI's future — helping to ensure AI tools are not merely impressive in theory, but transformative in practice.

Disclosures

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

This work is supported by funding from the American Heart Association.

Artificial intelligence tools were used to aid in the editing of the manuscript for this article. These tools were used under human oversight, and all scientific content, analysis, and conclusions reflect the authors' original work and judgment. All final content was determined and approved by the authors.

Author Affiliations

[1] Department of Medicine, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

[2] Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

[3] Department of Biomedical Data Science, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

[4] Department of Emergency Medicine, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

[5] Division of Pediatric Cardiac Anesthesia, Department of Anesthesiology, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

[6] Department of Health Policy, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

[7] Stanford Law School, Stanford University, Stanford, CA, USA

[8] Center for Biomedical Informatics Research, Stanford Medicine, Stanford University, Stanford, CA, USA

[9] Stanford Healthcare AI Applied Research Team, Division of Primary Care and Population Health, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

[10] Department of Cardiology, Kaiser Permanente, San Diego, CA, USA

[11] Stanford Center for Inherited Cardiovascular Disease, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

[12] Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, MA, USA

[13] Division of Cardiovascular Medicine, Department of Medicine, Stanford University, Stanford, CA, USA

[14] Stanford Center for Clinical Research, Department of Medicine, Stanford University, Stanford, CA, USA

[15] Division of Cardiovascular Medicine, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA

References

1. Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. JAMA 2019;322:1351-1352. DOI: 10.1001/jama.2019.10306.

2. Kleppe A. Area under the curve may hide poor generalisation to external datasets. ESMO Open 2022;7:100429. DOI: 10.1016/j.esmoop.2022.100429.

3. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur Radiol 2015;25:932-939. DOI: 10.1007/s00330-014-3487-0.

4. Huang C, Li S-X, Caraballo C, et al. Performance metrics for the comparative analysis of clinical risk prediction models employing machine learning. Circ Cardiovasc Qual Outcomes 2021;14:e007526. DOI: 10.1161/CIRCOUTCOMES.120.007526.

5. Lampert J, Bhatt DL, Vaid A, et al. Calibration of ECG-based deep-learning algorithm scores for patients flagged as high risk for hypertrophic cardiomyopathy. NEJM AI 2025;2(5). DOI: 10.1056/AIoa2400421.

6. Callahan A, McElfresh D, Banda JM, et al. Standing on FURM ground: a framework for evaluating fair, useful, and reliable AI models in health care systems. NEJM Catal 2024;5(10). DOI: 10.1056/CAT.24.0131.

7. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. Science 2024;383:164-167. DOI: 10.1126/science.adg8538.

8. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. NPJ Digit Med 2020;3:107. DOI: 10.1038/s41746-020-00318-y.

9. Singh K, Shah NH, Vickers AJ. Assessing the net benefit of machine learning models in the presence of resource constraints. J Am Med Inform Assoc JAMIA 2023;30:668-673. DOI: 10.1093/jamia/ocad006.

10. Ko M, Chen E, Agrawal A, et al. Improving hospital readmission prediction using individualized utility analysis. J Biomed Inform 2021;119:103826. DOI: 10.1016/j.jbi.2021.103826.

11. Ibnidris A, Liaskos N, Eldem E, et al. Facilitating the use of the target product profile in academic research: a systematic review. J Transl Med 2024;22:693. DOI: 10.1186/s12967-024-05476-1.

12. Gates Foundation. Integrated development: accelerating global health solutions. September 9, 2025 (https://www.gatesfoundation.org/our-work/programs/global-health/integrated-development).

13. World Health Organization. Target product profiles. September 9, 2025 (https://www.who.int/observatories/global-observatory-on-health-research-and-development/analyses-and-syntheses/target-product-profile/who-target-product-profiles).

14. Federal Register. Draft guidance for industry and review staff on target product profile-a strategic development process tool; availability. March 30, 2007 (https://www.federalregister.gov/documents/2007/03/30/E7-5949/draft-guidance-for-industry-and-review-staff-on-target-product-profile-a-strategic-development).

15. Shah NH, Jain SS. From better models to better care. JACC Heart Fail 2025;13:88-90. DOI: 10.1016/j.jchf.2024.09.021.

16. Maron BJ, Desai MY, Nishimura RA, et al. Diagnosis and evaluation of hypertrophic cardiomyopathy: JACC state-of-the-art

review. J Am Coll Cardiol 2022;79:372-389. DOI: 10.1016/j.jacc.2021.12.002.

17. Karra N, Klempfner Y, Fiman M, et al. Diagnosis of hypertrophic cardiomyopathy by artificial intelligence using standard transthoracic echocardiography. Eur Heart J 2024;45:ehae666.1003. DOI: 10.1093/eurheartj/ehae666.1003.

18. Siontis KC, Wieczorek MA, Maanja M, et al. Hypertrophic cardiomyopathy detection with artificial intelligence electrocardiography in international cohorts: an external validation study. Eur Heart J Digit Health 2024;5:416-426. DOI: 10.1093/ehjdh/ztae029.

19. Hillis JM, Bizzo BC, Mercaldo SF, et al. Detection of hypertrophic cardiomyopathy on electrocardiogram using artificial intelligence. Circ Heart Fail 2025;18:e012667. DOI: 10.1161/CIRCHEARTFAILURE.124.012667.

20. Goto S, Solanki D, John JE, et al. Multinational federated learning approach to train ECG and echocardiogram models for hypertrophic cardiomyopathy detection. Circulation 2022;146:755-769. DOI: 10.1161/CIRCULATIONAHA.121.058696.

21. Khalili A, Drummond J, Ramjattan N, Zeltser R, Makaryus AN. Diagnostic and treatment utility of echocardiography in the management of the cardiac patient. World J Cardiol 2020;12:262-268. DOI: 10.4330/wjc.v12.i6.262.

22. Narang A, Sinha SS, Rajagopalan B, et al. The supply and demand of the cardiovascular workforce. J Am Coll Cardiol 2016;68:1680-1689. DOI: 10.1016/j.jacc.2016.06.070.

23. Ko W-Y, Siontis KC, Attia ZI, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. J Am Coll Cardiol 2020;75:722-733. DOI: 10.1016/j.jacc.2019.12.030.

24. Desai MY, Jadam S, Abusafia M, et al. Real-world artificial intelligence-based electrocardiographic analysis to diagnose hypertrophic cardiomyopathy. JACC Clin Electrophysiol 2025;11:1324-1333. DOI: 10.1016/j.jacep.2025.02.024.

25. Mahmood U, Shukla-Dave A, Chan H-P, et al. Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing. Bjr Artif Intell 2024;1:ubae003. DOI: 10.1093/bjrai/ubae003.

26. Alves M, Seringa J, Silvestre T, Magalhães T. Use of artificial intelligence tools in supporting decision-making in hospital management. BMC Health Serv Res 2024;24:1282. DOI: 10.1186/s12913-024-11602-y.

27. Wornow M, Gyang Ross E, Callahan A, Shah NH. APLUS: a Python library for usefulness simulations of machine learning models in healthcare. J Biomed Inform 2023;139:104319. DOI: 10.1016/j.jbi.2023.104319.

28. O'Sullivan JW, Palepu A, Saab K, et al. Towards democratization of subspeciality medical expertise. October 1, 2024 (https://doi.org/10.48550/arXiv.2410.03741). Preprint.

29. Kim S-S, Seo H, Choi K, et al. Artificial intelligence model for detection of colorectal cancer on routine abdominopelvic CT examinations: a training and external-testing study. Am J Roentgenol 2025;224:e2432396. DOI: 10.2214/AJR.24.32396.

30. Lubell J. How AI is improving the rate of colon cancer risk detection. American Medical Association, February 10, 2025 (https://www.ama-assn.org/practice-management/digital-health/how-ai-improving-rate-colon-cancer-risk-detection).

31. Sutherland ME, Yarmis SJ, Lemkin DL, Winters ME, Dezman ZDW. National early warning score is modestly predictive of care escalation after emergency department-to-floor admission. J Emerg Med 2020;58:882-891. DOI: 10.1016/j.jemermed.2020.03.032.

32. Sundrani S, Chen J, Jin BT, Abad ZSH, Rajpurkar P, Kim D. Predicting patient decompensation from continuous physiologic monitoring in the emergency department. NPJ Digit Med 2023;6:60. DOI: 10.1038/s41746-023-00803-0.