JAMA | Special Communication

A Nationwide Network of Health AI Assurance Laboratories

Nigam H. Shah, MBBS, PhD; John D. Halamka, MD, MS; Suchi Saria, PhD; Michael Pencina, PhD; Troy Tazbaz, BS; Micky Tripathi, PhD, MPP; Alison Callahan, PhD; Hailey Hildahl, BS; Brian Anderson, MD

IMPORTANCE Given the importance of rigorous development and evaluation standards needed of artificial intelligence (AI) models used in health care, nationwide accepted procedures to provide assurance that the use of AI is fair, appropriate, valid, effective, and safe are urgently needed.

OBSERVATIONS While there are several efforts to develop standards and best practices to evaluate AI, there is a gap between having such guidance and the application of such guidance to both existing and new AI models being developed. As of now, there is no publicly available, nationwide mechanism that enables objective evaluation and ongoing assessment of the consequences of using health AI models in clinical care settings.

CONCLUSION AND RELEVANCE The need to create a public-private partnership to support a nationwide health AI assurance labs network is outlined here. In this network, community best practices could be applied for testing health AI models to produce reports on their performance that can be widely shared for managing the lifecycle of AI models over time and across populations and sites where these models are deployed.

JAMA. doi:10.1001/jama.2023.26930
Published online December 20, 2023.

Multimedia

CME at jamacmelookup.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Nigam H. Shah, MBBS, PhD, Center for Biomedical Informatics Research, 3180 Porter Dr, 112B, Palo Alto, CA 94305 (nigam@stanford.edu).

In March 2022, rigorous evaluation and guardrails for health care–related artificial intelligence (AI) were called for, which led to the creation of the Coalition for Health AI (CHAI) in December 2022 committed to developing guidelines for the responsible use of AI in health care.[1-3] CHAI is a community of health systems, public and private organizations, academia, patient advocacy groups, and expert practitioners of AI and data science that came together to harmonize standards and reporting for health AI and educate end users on how to evaluate efficacy and safe integration of these technologies into health care settings before adoption. In April 2023, members of the CHAI community released a draft blueprint for trustworthy AI implementation guidance and assurance for health care,[4] which as a next step envisioned assurance laboratories as a place to evaluate AI models via an agreed-on set of principles. Additionally, the labs would provide a sandbox environment for the development community that would enable ongoing innovation and future development, testing, and validation of safe and effective AI algorithms.

Over this same period following the launch of the viral app ChatGPT in November 2022, discussions of generative AI (genAI) have entered the mainstream media and dominated the narrative around AI more broadly, fueling hype in both the promise and perils that AI poses.[5-7] AI has dominated scholarly publications in science and health as well over this period; the number of publications in PubMed with "ChatGPT" in the title or abstract went from a mere 4 in December 2022 to 1456 as of October 2023, translating to roughly 5 articles being added every day since January 1, 2023, as of this writing.[5]

In reviewing existing community best practices for trustworthy AI, Lu et al[8] found more than 200 recommendations for reporting performance of models or describing characteristics of the source data via "model cards" and "data cards." While many randomized clinical trials or other types of scientific studies have evaluated the performance of AI models, each uses a different set of evaluation criteria, making it difficult to compare algorithms. This issue is compounded when applied to the wide variety of predictive AI models from disease detection to clinical intervention[9-11] that need performance validation and ongoing monitoring for algorithmic effectiveness across demographic and social determinants such as race and ethnicity, gender, age, geography, and income.[12,13] In areas where AI models fall within regulatory oversight, a framework for establishing safety, reliability, and efficacy exists.[14] However, any AI model falling outside of regulation, such as models for early detection of disease, automating billing procedures, facilitating scheduling, supporting public health disease surveillance, and other uses beyond traditional clinical decision support, should still follow similar rigor in its development, testing, and validation, as well as performance monitoring, when considering development and integration of decision support and/or administrative capabilities. For this discussion, health AI models according to the proposed rule 88 FR 23746, dated April 18, 2023, were scoped.[15]

Given Executive Order 14110 by President Biden,[16] which in section (G)(ii) calls for the development of AI assurance policy and infrastructure for measuring premarket and postmarket performance of AI models against real-world data, there is an urgent need for (1) development of standards, guidelines, and best practices to harness the capabilities of using AI guidance, while minimizing risk associated with it; (2) concrete guidance on procedures to ensure that the use of AI, including genAI, in health care is fair, appropriate, valid, effective, and safe (FAVES)[15,16]; (3) a place—an assurance

lab—where standards and validation procedures can be applied to produce reports on model performance that can be widely shared; and (4) processes for managing the lifecycle of AI models to ensure they maintain their performance over time, populations, and sites. While there are several bodies focused on the first item, there is a long road from enumeration to practical application of standards and best practices. Although the CHAI draft blueprint envisioned independent assurance labs, as of now, there is no publicly available, nationwide approach that enables objective assessment of health AI models and the consequences of their use.

Therefore, there is a rapidly growing need for a nationwide network of health AI assurance labs, whose purpose would be to evaluate models using nationwide standards and best practices. These labs could leverage an agreed-on set of community best practices for the development of trustworthy health AI, such as those developed by the CHAI,[4] and those from efforts like the National Academy of Medicine AI Code of Conduct.[17] Such a network of labs could be based on a set of patient privacy–respecting sources of data, collected, curated, and maintained by health care systems, payers, research organizations, and life science companies for the purpose of enabling transparent and localized testing of new AI models.[18] Specifically, a nationwide network for assurance labs could achieve the following critical goals for evaluation and development of AI in health care.

Shared Resource for Development and Validation

Assurance labs could serve as a shared resource for the industry to validate AI models, thus accelerating the pace of development and innovation, responsible and safe AI deployment, and successful market adoption.[19] A network of assurance labs could comprise both private and public entities, rather than one national organization, given the number and diversity of emerging models, the need for localized testing,[18] and the increasing recognition of the need for ongoing monitoring as well as reporting.[16] Such a network could fill a critical gap in an ecosystem dominated by well-meaning but often overexuberant and inexperienced developers who lack the depth of understanding of health care delivery. Given that health AI more broadly, including genAI, is subject to existing liability regulation for health care systems and physicians,[20,21] it is imperative that mechanisms are developed that use nationwide standards and best practices for testing and evaluation to ensure that the AI models developed for use in health care are trustworthy.

Comprehensive Evaluation of AI Models

Such labs could provide different levels of evaluation, ranging from a technical evaluation of model performance and bias for a specific use case,[22] to an interpretation of its performance for stratified subgroups of patients,[23-25] to a prospective evaluation of usability and adoption via human-machine teaming[26-29] and predeployment simulation of the consequences of using the model's output in light of specific policies and work capacity constraints.[30-32] The **Figure** shows an example evaluation report that might be generated in such assurance labs, for instance, a hypothetical scenario of using a prediction model to guide care interventions, such as one that predicts sepsis risk to guide patient care in the intensive care unit, with the report summarizing performance and achievable benefit in light of work capacity constraints.[31,32] Additionally, these labs could partner with model developers to help remediate specific areas (eg, bias) for improved performance and adherence to best practices. Such labs might also collaborate with the broader community to develop an agreed-on framework for evaluating genAI models.[33]

Transparent Reporting

The results of such evaluations could be published openly to a nationwide registry of AI tools that would include the model as an integral part. This registry would promote transparency by sharing plain language summaries of the evaluation with the general public, including patient stakeholders. A precedence exists for this exact approach in the Electronic Health Records Meaningful Use program created by the Health Information Technology for Economic and Clinical Health Act and the resulting Certified Health Product List.[34]

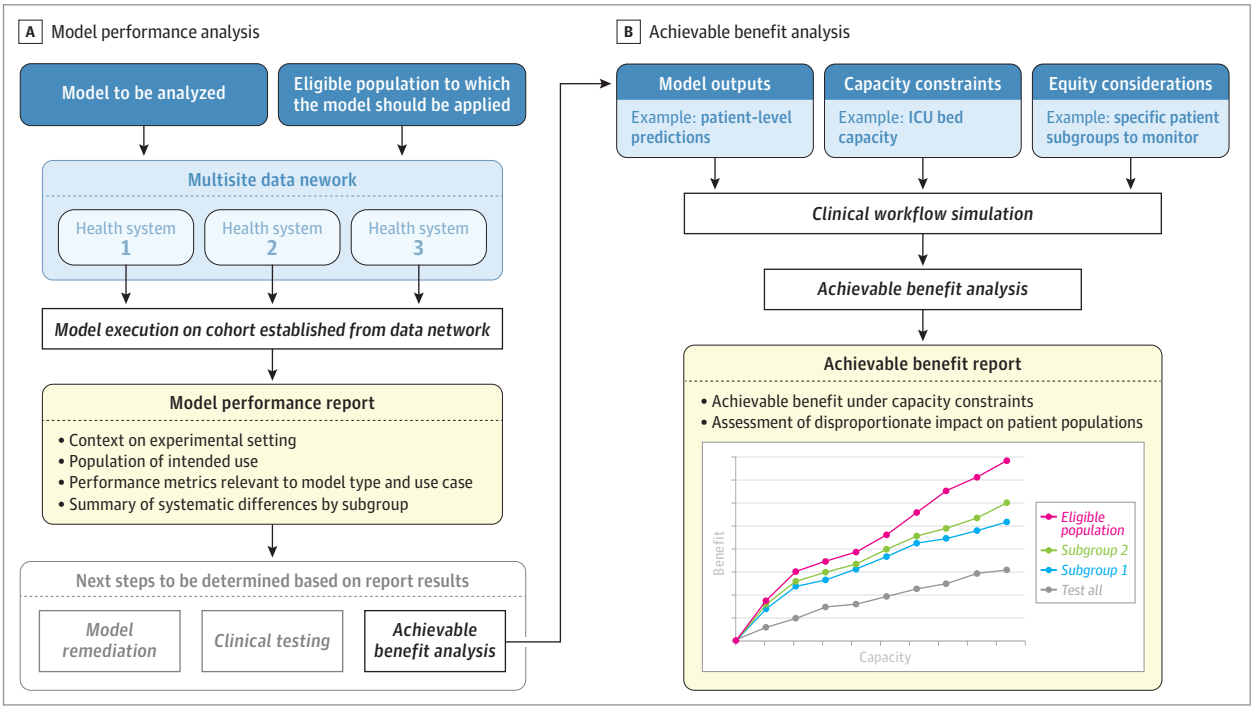
Promoting Regulatory Guidance

Further, these labs could be leveraged in implementing guidance set forth by regulatory agencies to generate a set of metrics and testing scripts for evaluating an AI model's performance. For example, currently, the US Food and Drug Administration is tasked with evaluating and approving models that are software as medical devices and are commercially marketed.[14] While this approach does provide an existing set of guardrails for evaluation, given the expected volume of submissions, as well as the need for "local validation,"[18] there may be value in partnering with qualified labs to produce the required metrics for validating the quality, safety, and efficacy of a model prior to premarket submission—analogous to CE (Conformité Européenne) marking of devices by notified bodies in Europe.[35] An example of a partnered approach is the Office of the National Coordinator for Health Information Technology's Certification Program, which is a voluntary program composed of functional and technical requirements known as "certification criteria" to which conformance is demonstrated using test procedures approved by the Office of the National Coordinator for Health Information Technology and National Institute of Standards and Technology, and performed by designated testing labs accredited by standards bodies based on the principles of the International Organization for Standardization and International Electrotechnical Commission framework.[36]

Ongoing Monitoring

A network of assurance labs could also provide monitoring of ongoing performance of AI models to ensure their intended objectives are achieved, in addition to offering services supporting federal regulation, such as the Predetermined Change Control Plan[37] and others, as they emerge. Such a network would help clinicians verify the appropriateness of AI models developed for use in health care delivery, whether those models are embedded in systems offered by electronic health record vendors or offered separately by third-party

Figure. Example Reports That Can Be Generated by an Assurance Laboratory



The consequences of using a model's output to guide decisions are a complex interplay of the properties of the model (its performance, inputs, and outputs), the workflow in which the model is deployed, and the action triggered by the model including the capacity of the team executing the action, the potential benefits and costs of acting, and whether those benefits and costs differ for different types of patients. The figure illustrates the different facets of a comprehensive assessment of a hypothetical scenario of using a prediction model to guide care interventions, such as one that predicts sepsis risk to guide patient care in the intensive care unit (ICU). A, The panel illustrates how an assurance lab can evaluate a model in silico using data from a multisite data network as well as details on the patient population for which it would be deployed to produce a report of the model's performance and bias by subgroups. B, The panel illustrates how an assurance lab can use the model's output, workflow, team capacity constraints, information about specific subgroups to monitor to perform a simulation (such as in Wornow et al[31]) estimating achievable benefit at different capacity constraints.

developers or created by the health care organization itself. There is a need to provide credible verification of information to clinicians for the use of health care–specific medical and nonmedical algorithms, including verifying health equity risks of models before they are integrated into the care delivery process. Independent third-party testing of AI models—irrespective of the source of the model—provides a path for adhering to assurance standards agreed on via a community consensus and would greatly facilitate governance decisions at health systems about which algorithms are trustworthy.

While the concept of a nationwide network of assurance labs might be the most direct path toward trustworthy health AI, it is not without limitations. First, applications of AI tools are inherently local and any evaluation needs to account for local context[18]; a network of assurance labs needs to develop an approach that takes local context into account. An alternative would be to enable health systems to create their own local assurance labs. While possible for the larger health systems and academic medical centers, this alternative would not scale, even if all were to rely on consensus standards developed by CHAI. Having such local labs would also exacerbate health system level inequity, with better-resourced systems able to provide stronger protections. Our proposed approach also must ensure system-level equity. Specifically, the assurance lab network needs to develop a revenue model in a manner that does not further disadvantage less well-resourced health systems. Another alternative would be to create a national-level, government-operated assurance lab. However, this would require an enormous effort and investment and run into similar problems of poor local connectivity. Yet another approach would encourage commercial assurance labs, either connected to large AI developers or private for-profit assurers. Such a setup raises ethical problems—large AI developers assuring their own products can be likened to a fox guarding the hen house. While we believe that incentives would be better aligned if such entities were nonprofits focused on development and scaling of assurance-enabling technologies, we would not preclude consideration of for-profit assurance labs that adhere to community standards. Yet another approach is to empower solution developers to do local testing and validation and share results in a manner that is verifiable by the assurance labs. Given the diversity of choices available, we propose a modest start with a small number of assurance labs that experiment with these diverse approaches and gather evidence that the creation of such labs can meet the goals laid out in the Executive Order's section (G)(ii).

A focus on AI testing and a structure for doing so that uses open, consensus-based nationwide standards applied to datasets specific to the use case of the model and examines the implications of using a model's output for the use case at hand is critically needed. An assurance labs network enhances the possibility of delivering on the high expectations of AI in health, and may mitigate against potential disappointments as has happened with AI adoption in other sectors (eg, self-driving cars).[38]

Conclusions

As the technology for building models becomes widely available and community consensus on how to evaluate their performance emerges, the rationale for "a lab for testing" to ensure model credibility as well as accountability is increasing. A public-private partnership to launch a nationwide network of health AI assurance labs could promote transparent, reliable, and credible health AI. CHAI, which includes ex-officio government members (US Food and Drug Administration, Office of the National Coordinator for Health Information Technology, Centers for Medicare & Medicaid Services, National Institutes of Health, Veterans Affairs, White House Office of Science and Technology Policy, and others) as observers and works in close partnership with the National Academy of Medicine's AI Code of Conduct initiative,[17] looks forward to fostering a nationwide conversation to help shape the creation, implementation, and operation of an assurance labs network that can help fulfill the promise that responsible health AI offers for the health care system.

ARTICLE INFORMATION

Accepted for Publication: December 10, 2023.

Published Online: December 20, 2023. doi:10.1001/jama.2023.26930

Author Affiliations: Stanford Medicine, Palo Alto, California (Shah, Callahan); Coalition for Health AI, Dover, Delaware (Shah, Halamka, Saria, Pencina, Anderson); Mayo Clinic Platform, Mayo Clinic, Rochester, Minnesota (Halamka, Hildahl); Bayesian Health, New York, New York (Saria); Johns Hopkins University, Baltimore, Maryland (Saria); Johns Hopkins Medicine, Baltimore, Maryland (Saria); Duke AI Health, Duke University School of Medicine, Durham, North Carolina (Pencina); US Food and Drug Administration, Silver Spring, Maryland (Tazbaz); US Office of the National Coordinator for Health IT, Washington, DC (Tripathi); MITRE Corporation, Bedford, Massachusetts (Anderson).

Conflict of Interest Disclosures: Dr Shah reported being a cofounder of Prealize Health (a predictive analytics company) and Atropos Health (an on-demand evidence generation company); receiving funding from the Gordon and Betty Moore Foundation for developing virtual model deployments; and being a member of working groups of the Coalition for Healthcare AI (CHAI), a consensus-building organization providing guidelines for the responsible use of artificial intelligence in health care. Dr Saria reported receiving equity from Bayesian Health. Dr Pencina reported receiving grants from the Gordon and Betty Moore Foundation; personal fees from McGill University Health Centre, Cleerly Inc, Eli Lilly, and Janssen; and stock options from Azra Care; in addition, Dr Pencina had a patent for copyright/trademark pending for algorithmic governance. Ms Hildahl reported being employed by Mayo Clinic Platform's Validate, which offers objective, third-party validation for a fee. No other disclosures were reported.

Disclaimer: Dr Pencina is a Statistical Reviewer for *JAMA* but was not involved in any of the decisions regarding review of the manuscript or its acceptance.

REFERENCES

1. Ross C. A new coalition aims to close AI's credibility gap in medicine with testing and oversight. STAT. December 7, 2022. Accessed November 14, 2023. https://www.statnews.com/2022/12/07/artificial-intelligence-hospitals-coalition-health-ai/

2. Halamka JD, Saria S, Shah NH. Health-related artificial intelligence needs rigorous evaluation and guardrails. STAT. March 17, 2022. Accessed November 14, 2023. https://www.statnews.com/2022/03/17/health-related-ai-needs-rigorous-evaluation-and-guardrails/

3. Saria S. Not all AI is created equal: strategies for safe and effective adoption. *NEJM Catal*. Published March 23, 2022. doi:10.1056/CAT.22.0075

4. Coalition for Health AI. Blueprint for trustworthy AI implementation guidance and assurance for healthcare: version 1.0_April 04, 2023. Accessed November 14, 2023. https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

5. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388:1233-1239. doi:10.1056/NEJMsr2214184

6. Drazen JM, Kohane IS, Leong TY. Artificial intelligence in US health care delivery. *N Engl J Med*. 2023. doi:10.1056/NEJMra2204673

7. Sanders NE, Schneier B. How ChatGPT hijacks democracy. *New York Times*. January 15, 2023. Accessed November 14, 2023. https://www.nytimes.com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy.html

8. Lu JH, Callahan A, Patel BS, et al. Assessment of adherence to reporting guidelines by commonly used clinical prediction models from a single vendor: a systematic review. *JAMA Netw Open*. 2022;5(8):e2227779. doi:10.1001/jamanetworkopen.2022.27779

9. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*. 2020;369:m1328. doi:10.1136/bmj.m1328

10. van der Vegt AH, Scott IA, Dermawan K, Schnetler RJ, Kalke VR, Lane PJ. Deployment of machine learning algorithms to predict sepsis: systematic review and application of the SALIENT clinical AI implementation framework. *J Am Med Inform Assoc*. 2023;30(7):1349-1361. doi:10.1093/jamia/ocad075

11. Lee S, Chu Y, Ryu J, Park YJ, Yang S, Koh SB. Artificial intelligence for detection of cardiovascular-related diseases from wearable devices: a systematic review and meta-analysis. *Yonsei Med J*. 2022;63(suppl):S93-S107. doi:10.3349/ymj.2022.63.S93

12. Celi LA, Cellini J, Charpignon ML, et al; for MIT Critical Data. Sources of bias in artificial intelligence that perpetuate healthcare disparities: a global review. *PLOS Digit Health*. 2022;1(3):e0000022. doi:10.1371/journal.pdig.0000022

13. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283-286. doi:10.1056/NEJMc2104626

14. Center for Devices and Radiological Health, US Food and Drug Administration. Software as a medical device (SaMD). Accessed November 14, 2023. https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd

15. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing. *Federal Register*. Accessed November 14, 2023. https://www.federalregister.gov/d/2023-07229/p-562

16. Safe, secure, and trustworthy development and use of artificial intelligence. *Federal Register*. Accessed November 14, 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

17. National Academy of Medicine. Health Care Artificial Intelligence Code of Conduct. Accessed November 14, 2023. https://nam.edu/programs/value-science-driven-health-care/health-care-artificial-intelligence-code-of-conduct/

18. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. *Nat Med*. 2023;29(11):2686-2687. doi:10.1038/s41591-023-02540-z

19. The White House. Fact Sheet: Biden-Harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI. July 21, 2023. Accessed November 14, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

20. Price WN II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA*. 2019;322(18):1765-1766. doi:10.1001/jama.2019.15064

21. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. *JAMA Health Forum*. 2023;4(5):e231938. doi:10.1001/jamahealthforum.2023.1938

22. Wang HE, Landers M, Adams R, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. *J Am Med Inform Assoc*. 2022;29(8):1323-1333. doi:10.1093/jamia/ocac065

23. Yang Y, Zhang H, Katabi D, Ghassemi M. Change is hard: a closer look at subpopulation shift. *arXiv*. Preprint posted August 17, 2023. doi:10.48550/arXiv.2302.12254

24. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical

risk prediction. *J Biomed Inform*. 2021;113:103621. doi:10.1016/j.jbi.2020.103621

25. Cary MP Jr, Zink A, Wei S, et al. Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review. *Health Aff (Millwood)*. 2023;42(10):1359-1368. doi:10.1377/hlthaff.2023.00553

26. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. 2021;27(2):186-187. doi:10.1038/s41591-021-01229-5

27. Henry KE, et al. Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. npj. *NPJ Digit Med*. 2022;5:1-6. doi:10.1038/s41746-022-00597-7

28. Henry KE, Adams R, Parent C, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nat Med*. 2022;28(7):1447-1454. doi:10.1038/s41591-022-01895-z

29. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med*. 2022;28(7):1455-1460. doi:10.1038/s41591-022-01894-0

30. Chohlas-Wood A, Coots M, Zhu H, Brunskill E, Goel S. Learning to be fair: a consequentialist approach to equitable decision-making. *arXiv*. Preprint posted February 1, 2023. doi:10.48550/arXiv.2109.08792

31. Wornow M, Ross EG, 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

32. 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*. 2023;30(4):668-673. doi:10.1093/jamia/ocad006

33. Fleming SL, Lozano A, Haberkorn WJ, et al. MedAlign: a clinician-generated dataset for instruction following with electronic medical records. *arXiv*. Preprint posted August 27, 2023. doi:10.48550/arXiv.2308.14089

34. Certified Health IT Product List. Accessed November 14, 2023. https://chpl.healthit.gov/#/search

35. European Commission. Internal market, industry, entrepreneurship and SMEs: CE marking. Accessed November 14, 2023. https://single-market-economy.ec.europa.eu/single-market/ce-marking_en

36. HealthIT.gov. Certification of health IT. Accessed November 14, 2023. https://www.healthit.gov/topic/certification-ehrs/certification-health-it

37. Center for Devices and Radiological Health, Food and Drug Administration. CDRH issues draft guidance on predetermined change control plans for artificial intelligence/machine learning-enabled medical devices. March 30, 2023. Accessed November 14, 2023. https://www.fda.gov/medical-devices/medical-devices-news-and-events/cdrh-issues-draft-guidance-predetermined-change-control-plans-artificial-intelligencemachine

38. Angwin J. Autonomous vehicles are driving blind. *New York Times*. October 11, 2023. Accessed November 14, 2023. https://www.nytimes.com/2023/10/11/opinion/driverless-cars-san-francisco.html