NEJM AI

NEJM AI 2024; 1 (8)
DOI: 10.1056/AIp2400223

PERSPECTIVE

A Call for Artificial Intelligence Implementation Science Centers to Evaluate Clinical Effectiveness

Christopher A. Longhurst, M.D., M.S.,[1] Karandeep Singh, M.D., M.M.Sc.,[1] Aneesh Chopra, M.P.P.,[2] Ashish Atreja, M.D., M.P.H.,[3] and John S. Brownstein, Ph.D.[4,5]

Received: March 1, 2024; Revised: May 7, 2024; Accepted: May 16, 2024; Published: July 10, 2024

Abstract

Artificial intelligence (AI) holds significant promise for revolutionizing health care by enhancing diagnosis, treatment, and patients' safety. However, the current disparity between the abundance of AI research and the scarcity of evidence on real-world impact underscores the urgent need for comprehensive clinical effectiveness evaluations. These evaluations must go beyond model validation to explore the real-world effectiveness of AI models in clinical settings, especially because so few have gone on to show any meaningful impact. The importance of local context in AI model validation and impact assessment cannot be overstated. We call for increased recognition of implementation science principles and their adoption through development of a network of health care delivery organizations to focus on the clinical effectiveness of AI models in real-world settings to help achieve the shared goal of safer, more effective, and equitable care for all patients.

Introduction

Artificial intelligence (AI) has become a promising tool in health care, with increasing evidence that these tools can support a transformational improvement in diagnosis, treatment, and patients' safety.[1] We must proceed carefully, because AI algorithms have the potential to amplify biases intrinsic to training data and unintentionally exacerbate underlying health inequities.[2] In response to this risk, a group including representatives from the U.S. Food and Drug Administration (FDA) and the Office of the National Coordinator for Health Information Technology (ONC) recently called for the establishment of a national network of health AI assurance laboratories to provide a standardized approach to evaluating these algorithms before FDA approval.[3] We believe that validating algorithms through computer simulation may be helpful but is insufficient for demonstrating improved patient outcomes, which will require real-world evaluations similar to clinical trials for a new drug. In fact, in vivo and real-world evaluation is the most critical part of approval for a novel drug or device, especially in settings in which retrospective data are sparse but

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

Dr. Longhurst can be contacted at clonghurst@health.ucsd.edu or at 9560 Towne Center Dr., La Jolla, CA 92122.

in vivo interventions are common; colleagues have recently called for a similar, patient-centered, outcomes-focused approach to regulating AI.[4] We propose that in vivo validation of clinical effectiveness should become a standard part of assessing the value of health AI, encourage funding agencies to resource these efforts, and suggest that regulators require this type of outcomes data.

This gap becomes particularly evident when we consider the vast amount of research on AI in specific areas such as sepsis. A recent analysis of more than 500 articles published on machine learning for sepsis diagnosis and management found just two evaluating the benefit of a model to improve patient-centered outcomes.[5] Another recent study evaluating the discriminative performance of AI sepsis predictions made relative to the time of treatment found that many models were built using features that simply capture clinical suspicion; this illustrates one of many reasons why computer simulation evaluations could produce overly optimistic estimates of model performance that would not be as favorable once implemented.[6] This stark contrast between the volume of research and the scarcity of evidence for real-world impact underscores the urgent need for implementation of science frameworks and evaluations that go beyond model validation and delve into the practicalities and clinical effectiveness of AI in real-world settings.[7] Although AI models may exhibit impressive performance in controlled settings, their true worth is determined by their integration into the complex, dynamic, and often unpredictable health care ecosystem (Table 1).

Health delivery organizations must play a critical role in this real-world validation, and mitigation of health inequities will require that public health systems serving diverse populations be well-represented. For example, recent work at UC San Diego Health, which developed a bespoke AI tool for sepsis detection, reported a 17% increase in the relative likelihood of patient survival in the emergency department, with the greatest impact at the location serving the highest proportion of underserved patients.[8] However, implementation of this same algorithm in a different health system would be unlikely to have the same outcome without local tuning, iteration, and ongoing monitoring.[9]

As with any clinical decision support tool, implementation of AI-based decision support necessitates a profound understanding of the local health care context, rigorous and ethical validation within these settings, workflow integration and evolution, and continuous monitoring after implementation, as suggested by the White House Office of Science and Technology Policy in the AI Bill of Rights.[10] By focusing on implementation evaluations, we can gain crucial insights into how AI tools integrate into existing workflows, how health care professionals interact with them, and, ultimately, how these tools can best translate into tangible improvements in patient outcomes. The following sections detail some key insights that a network of health care delivery organizations focused on health AI implementation assurance and demonstration of clinical effectiveness could uniquely provide.

Local Data and Real-World Insights

Health systems possess extensive repositories of local patient data, encompassing diverse populations, diseases, and clinical practices. These data repositories are a valuable resource for AI model training and validation. AI models trained on global datasets may not fully capture the intricacies of local patient populations and health care practices. Health systems can provide the necessary context to refine and adapt AI models to suit their specific patient demographic and clinical workflows.

Ethical AI Development

Ensuring the ethical use of AI in health care is paramount. Health systems can contribute to ethical AI development by actively participating in the validation process, identifying potential biases, and ensuring that outcomes associated with the use of AI models align with the FAVES (Fair, Appropriate, Valid, Effective, and Safe) principles developed by the ONC, amplified by the White House, and endorsed by many health systems.[11] Local context-aware validation helps uncover and rectify biases that may disproportionately affect certain patient groups. UC San Diego Health, UC Davis Health, and Boston Children's Hospital have health AI committees that evaluate machine learning models from ethical and health equity perspectives to ensure safety and trust.

Workflow Integration

Successful AI implementation requires seamless integration into clinical workflows. Health systems are best positioned to design and adapt workflows to incorporate AI

Table 1. Key Aspects of AI Implementation and Related Frameworks.*

Health System Process	Key Points	High-Level Examples	Relevant Frameworks
Local data and real-world insights	Rich datasets reflecting local patient demographic characteristics, disease prevalence, and clinical practices. AI models trained on generic data may miss important nuances specific to a particular region or population. Local data can be used to fine-tune AI models for optimal performance in specific clinical settings.	A rural hospital uses local data to train an AI model for early detection of diabetic foot ulcers, leading to improved outcomes for patients with limited access to specialists. An AI model for predicting hospital readmissions performs poorly in a state with a high prevalence of chronic illnesses, highlighting the need for local adjustments. A general AI model for identifying pneumonia is adapted using local data to differentiate between bacterial and viral pneumonia, allowing for more precise treatment decisions.	Determinant frameworks can help identify barriers and facilitators to interventions such that they can be tailored to account for the local context. For example, the TDF identifies 14 domains that can influence behavior change, including knowledge, skills, beliefs about capabilities, and environmental context and resources.
Ethical AI development	Proactive identification and mitigation of potential biases and ensuring adherence to ethical principles. Local context helps identify and address biases that may disproportionately affect certain groups. AI models should be developed and implemented in accordance with ethical guidelines, such as fairness, transparency, and accountability.	A health system identifies and addresses potential racial bias in an AI model for predicting disease risk, ensuring fair and equitable outcomes for all patients. A hospital develops an AI-based language translator specifically tailored to the needs of their multilingual patient population, promoting better communication and patient engagement. An AI system designed to automate patient appointment scheduling is programmed with clear ethical guidelines to ensure patient privacy and prevent discrimination.	The Theory of Change Framework maps out desired outcomes and steps needed to achieve them, which can ensure that evaluations of AI fairness are measured against the ability to achieve the desired outcomes.
Workflow integration	Seamless integration of AI tools into existing clinical workflows to optimize efficiency and minimize disruption. Health systems can adapt their workflows to leverage the capabilities of AI tools effectively. Ongoing monitoring and evaluation are crucial to ensure smooth integration and optimal performance of AI within workflows.	An AI-based system for medication dosage recommendations is integrated directly into the electronic health record, allowing clinicians to access and use information efficiently during patient consultations. A hospital develops a new workflow for managing appointments based on an AI-powered patient risk assessment tool, resulting in reduced wait times and improved patient satisfaction. A health system regularly evaluates the impact of an AI-based diagnostic tool on its workflow, making adjustments as needed to maintain efficiency and patient safety.	Process models can help map out the iterative process required to integrate AI into clinical workflows. For example, the CFIR can diagnose potential issues before they become barriers in the implementation process and helps in adjusting strategies for better outcomes.
Continuous monitoring and adaptation	Ongoing assessment of AI performance in real-world settings and adaptability to evolving circumstances. Health systems can identify and address discrepancies between controlled testing and real-world performance through continuous monitoring and impact assessment. Continuous monitoring and adaptation ensure that AI models remain safe, effective, affordable, and relevant in the dynamic health care environment.	An AI model for predicting patient mortality initially performs well, but its accuracy declines with changes in treatment protocols. The model is adapted and retrained to maintain its effectiveness. A hospital discovers that an AI-based tool for identifying potential drug interactions performs differently in specific patient cohorts. This insight leads to the development of tailored interventions for those groups. A health care system regularly updates its AI models based on the latest data and clinical research, ensuring they remain at the forefront of technological advancement and provide the best possible care for patients.	The PDSA cycles are iterative and can ensure that AI is continuously optimized based on real-world use and feedback. Evaluation frameworks can help plan for larger impact. For example, the RE-AIM framework evaluates the health impact of an intervention by assessing five key components: reach to target population, effectiveness in achieving desired outcomes, adoption by target settings or institutions, implementation fidelity, and maintenance of the intervention in practice over time.

* AI denotes artificial intelligence; CFIR, Consolidated Framework for Implementation Research; PDSA, Plan-Do-Study-Act; RE-AIM, Reach, Efficacy, Adoption, Implementation, Maintenance; and TDF, Theoretical Domains Framework.

tools effectively. They can assess the impact of AI on existing processes and make necessary adjustments to ensure the optimal use of AI to drive impact. For example, Boston Children's Hospital successfully integrated the POPP (Prediction of Patient Placement) system into their admissions workflow. POPP uses clinical data to predict incoming admissions from the emergency department, allowing for proactive coordination of downstream operations.[12] This reduces emergency department boarding times for patients and can generate substantial cost savings. To ensure rapid iteration and user-centric design, the multidisciplinary project team collaborated with the end users (operational staff coordinating patient placement) throughout the design, development, and deployment processes. By predicting inpatient capacity needs, POPP supports broader hospital capacity planning efforts and helps facilitate timely patient care.

Continuous Monitoring and Adaptation

The dynamic nature of health care necessitates continuous monitoring and adaptation of AI models. Health systems can play a vital role in postimplementation evaluation, identifying any discrepancies between in situ model performance and its impact in real-world scenarios. This feedback loop will help ensure that AI remains safe, effective, and of high value over time. As an example, UC San Diego Health successfully deployed an AI tool to identify Covid-19 pneumonia on chest radiographs early in the pandemic, influencing clinical decisions for a significant portion of emergency department patients.[13] However, the AI tool was no longer needed when Covid-19 testing became ubiquitous, highlighting the importance of continuous monitoring and adaptation of AI solutions. Of note, a recent systematic review of publications about Covid-19 and AI during a 2-year period found that this study was just one of four studies among more than 9000 to evaluate clinical impact, highlighting again the need to focus on outcomes.[14] Fortunately, some electronic health record vendors are now beginning to provide open source tools to support testing and monitoring of models in situ.[15]

Local Context

The importance of local context in AI model validation and comprehensive real-world impact assessment cannot be overstated. Although AI models developed on a global scale provide a foundation, they must be deployed along with meaningful interventions in a socio-technical environment in which workflow considerations may far outweigh subtleties of algorithm performance. For example, a recent pilot at UC San Diego Health integrating large language models to generate draft replies to patient messages was expected to reduce time spent answering messages; however, the randomized quality improvement study showed that the AI-enabled workflow did not save time, although it did decrease cognitive burden and burnout among primary care clinicians.[16]

The nuances of patient populations, clinical practices, and data availability vary from one health system to another. Recently, dozens of health care systems have formed collaboratives (e.g., https://validai.health/) to adopt implementation science standards and measure AI impact; efforts such as these will require broader support and funding for sustainability. We are heartened by similar calls to action as noted in the 2023 Report on Patient Safety from the President's Council of Advisors on Science and Technology; this report encourages federal agencies to launch an "AI for Patient Safety" collaborative effort, with emphasis on insights gained when integrating technologies "within current workflows."[17]

The field of implementation science has developed various frameworks that can be applied to AI in health care. Determinant frameworks (e.g., TDF [Theoretical Domains Framework], NASSS [Non-Adoption, Abandonment, Scale-up, Spread, and Sustainability], and CFIR [Consolidated Framework for Implementation Research]) focus on identifying factors, including local context, that can influence outcomes; process models (e.g., PRISM [Practical, Robust Implementation and Sustainability Model], EPIS [Exploration, Adoption/Preparation, Implementation, Sustainment], and PDSA [Plan-Do-Study-Act]) provide a step-by-step guide to the implementation process; and evaluation frameworks (e.g., Implementation Outcomes Framework and the RE-AIM [Reach, Efficacy, Adoption, Implementation, Maintenance] framework) can be used to plan and evaluate complex local environments and larger health impacts, respectively. By focusing on implementation science and evaluations, we can gain crucial insights into how AI tools integrate into existing workflows, how health care professionals interact with them, and, ultimately, how these tools can best translate into tangible improvements in patient outcomes.

As with the role of regional extension centers funded by the ONC to help physicians meaningfully use electronic health record systems in practice, we call on public and private stakeholders to fund and resource a network of health care delivery organizations to function as AI implementation science centers focused on the clinical effectiveness of AI models in real-world implementations. Such networks will help ensure we achieve outcomes that matter to patients and deliver on their promise of safer, more effective, and more equitable care for all patients.

Disclosures

Author disclosures are available at ai.nejm.org.

Author Affiliations

[1] Department of Biomedical Informatics, UC San Diego Health, San Diego, CA

[2] CareJourney, Arlington, VA

[3] Department of Medicine, UC Davis Health, Sacramento, CA

[4] Department of Pediatrics, Harvard Medical School, Boston

[5] Innovation and Digital Health Accelerator, Boston Children's Hospital, Boston

References

1 Classen DC, Longhurst C, Thomas EJ. Bending the patient safety curve: how much can AI help? NPJ Digit Med 2023;6:2. DOI: 10.1038/s41746-022-00731-5.

2. Goldberg CB, Adams L, Blumenthal D, et al. To do no harm — and the most good — with AI in health care. NEJM AI 2024;1(3). DOI: 10.1056/AIp2400036.

3. Shah NH, Halamka JD, Saria S, et al. A nationwide network of health AI assurance laboratories. JAMA 2024;331:245-249. DOI: 10.1001/jama.2023.26930.

4. Ayers JW, Desai N, Smith DM. Regulate artificial intelligence in health care by prioritizing patient outcomes. JAMA 2024;331:639-640. DOI: 10.1001/jama.2024.0549.

5. Wardi G, Owens R, Josef C, Malhotra A, Longhurst C, Nemati S. Bringing the promise of artificial intelligence to critical care: what the experience with sepsis analytics can teach us. Crit Care Med 2023;51:985-991. DOI: 10.1097/CCM.0000000000005894.

6. Kamran F, Tjandra D, Heiler A, et al. Evaluation of sepsis prediction models before onset of treatment. NEJM AI 2024;1(3). DOI: 10.1056/AIoa2300032.

7. Moullin JC, Dickson KS, Stadnick NA, et al. Ten recommendations for using implementation frameworks in research and practice. Implement Sci Commun 2020;1:42. DOI: 10.1186/s43058-020-00023-7.

8 Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. NPJ Digit Med 2024;7:14. DOI: 10.1038/s41746-023-00986-6.

9 Kwong JCC, Nickel GC, Wang SCY, Kvedar JC. Integrating artificial intelligence into healthcare systems: more than just the algorithm. NPJ Digit Med 2024;7:52. DOI: 10.1038/s41746-024-01066-z.

10. The White House. Blueprint for an AI Bill of Rights: making automated systems work for the American people. October 2022 (https://www.whitehouse.gov/ostp/ai-bill-of-rights/).

11. The White House. Delivering on the promise of AI to improve health outcomes. December 13, 2023 (https://www.whitehouse.gov/briefing-room/blog/2023/12/14/delivering-on-the-promise-of-ai-to-improve-health-outcomes/).

12. Barak-Corren Y, Fine AM, Reis BY. Early prediction model of patient hospitalization from the pediatric emergency department. Pediatrics 2017;139:e20162785. DOI: 10.1542/peds.2016-2785.

13. Carlile M, Hurt B, Hsiao A, Hogarth M, Longhurst CA, Dameff C. Deployment of artificial intelligence for radiographic diagnosis of COVID-19 pneumonia in the emergency department. J Am Coll Emerg Physicians Open 2020;1:1459-1464. DOI: 10.1002/emp2.12297.

14 Ahmadi Marzaleh M, Peyravi M, Mousavi S, Sarpourian F, Seyedi M, Shalyari N. Artificial intelligence functionalities during the COVID-19 pandemic. Disaster Med Public Health Prep 2023;17:e336. DOI: 10.1017/dmp.2023.3.

15. Landi H. Epic plans to launch AI validation software for healthcare organizations to test, monitor models. Fierce Healthcare. April 3, 2024 (https://www.fiercehealthcare.com/ai-and-machine-learning/epic-plans-launch-ai-validation-software-healthcare-organizations-test).

16 Tai-Seale M, Baxter SL, Vaida F, et al. AI-generated draft replies integrated into health records and physicians' electronic communication. JAMA Netw Open 2024;7:e246565. DOI: 10.1001/jamanetworkopen.2024.6565.

17. President's Council of Advisors on Science and Technology. Report to the President: a transformational effort on patient safety. September 2023 (https://www.whitehouse.gov/wp-content/uploads/2023/09/PCAST_Patient-Safety-Report_Sept2023.pdf).