Stanford University
Human-Centered
Artificial Intelligence

JULY 2021

SPECIAL
HEALTHCARE
SERIES

Improving AI Software for Healthcare Diagnostics

David B. Larson, Daniel L. Rubin, and Curtis P. Langlotz

ONE OF THE MOST PROMISING USES of artificial intelligence (AI) is in radiology, the medical specialization that uses imaging technology to diagnose and treat disease. AI holds great promise for more accurate healthcare diagnostics and even prediction of disease outcomes for patients. AI can improve traditional medical imaging methods like computed tomography (CT), magnetic resonance imaging (MRI), and X-ray by offering computational capabilities that process images with greater speed and accuracy, automatically recognizing complex patterns to assess a patient's health. This sophisticated software needs more robust evaluation methods to reduce risk to the patient, to establish trust, and to ensure wider adoption. A clear example of this can be seen in the difficulty many researchers had in classifying imaging results from early studies of the coronavirus disease that spawned the COVID-19 pandemic.

KEY TAKEAWAYS

- AI-based diagnostics show great promise to improve traditional medical imaging methods, such as CT scans, MRIs, and X-rays. These algorithms offer computational capabilities that process images with greater speed and accuracy than traditional methods and can improve patient outcomes for millions.
- Current proposals for regulatory frameworks do not fully address the necessity to build trust in these systems due to the confusion between the algorithm in question and the task it is designed to perform, inadequate establishment of standard-setting bodies, and insufficient rigor in the evaluation and development process.
- Policymakers should turn to medical societies for the clinical definitions of diagnostic tasks. These groups should extend performance assessments beyond simply testing for accuracy.

In our article in the *Journal of the American College of Radiology,* "Regulatory Frameworks for Development and Evaluation of Artificial Intelligence-Based Diagnostic Imaging Algorithms," we explore three major regulatory frameworks for radiology that have been proposed by the United States Food and Drug Administration (FDA), the European Union, and the International Medical Device Regulators Forum, respectively, and show how they ensure safety, effectiveness, and performance of AI-based applications. However, these regulatory bodies could be doing more to build trust in these systems; we recommend changes that need to be made so diagnostic AI can reach its full potential. The shortcomings we enumerate extend from the tendency to confuse the algorithm with the task it is designed to perform, from the lack of rigor in definitions of medical tasks, and from the lax specification of the task, making it harder to compare similar algorithms directly. We also identify problems with unpredictability in model performance, insufficient testing infrastructure, and inherent conflicts of interest.

A better path forward depends on policymakers and medical societies adopting stronger regulatory guidance to improve testing, enhance safety, and establish performance standards for these algorithms. Gaps in the three regulatory frameworks we examined can be filled by the following four actions:

1. Make sure the algorithm is always distinguished from the definition of the diagnostic task it is automating.
2. Define elements of algorithmic performance beyond accuracy, such as transparency, use of fail-safes, and auditability.
3. Divide the evaluation into discrete steps from the perspective of the potential user or evaluator, including diagnostic task definition, capacity of the algorithm to perform in a controlled environment, evaluation of effectiveness in the real world as compared to performance in the controlled environment, validation of effectiveness in the local setting at each installed site, and durability testing and monitoring to ensure the algorithm performs well over time.
4. Encourage independent assessment by third-party evaluators by implementing a phased testing regime similar to that used by the pharmaceutical industry during drug development and the broader software industry.

Taken together, these steps have the potential to shape diagnostic AI's future for the better and ensure the technology is developed as fairly and as rapidly as possible.

Introduction

Central to the issue of improving diagnostic systems built around artificial intelligence is understanding the proposed regulatory frameworks. Leading the way in discussions about testing and validation of the Software as a Medical Device (SaMD) market—i.e., any software program intended to diagnose, prevent, monitor, treat, or alleviate a disease—is the International Medical Device Regulators Forum (IMDRF). The European Union has codified requirements for clinical evaluation and has specified how manufacturers should prepare and follow up with post market planning. The FDA recently published a working model for software precertification programs. The IMDRF has developed a host of suggestions for SaMD applications, including recommendations to create separate risk categories that track the degree of regulatory scrutiny a given program deserves, principles for creating quality management systems, principles for evaluation of clinical effectiveness, and recommendations for how clinical evaluation reports are to be drafted.

Despite the ongoing activity in these regulatory proposals, our examination identifies a handful of shortcomings that require policymakers' urgent attention: First, the diagnostic tasks that these SaMD systems perform continue to be confused with the algorithm performing the task in question. Though the two are closely linked, activities such as defining what constitutes disease such as pneumonia on a chest radiograph ultimately need to be distinguished from the program that applies the definition of the disease in analyzing an X-ray. Second, because many measurement systems like characterizing a patient's blood flow were designed by a device manufacturer without input from standards-setting bodies, diagnostic task definitions frequently lack the rigor needed to enable widespread adoption. Third, this lack of clarity hampers direct comparison of similar algorithms. Because most diagnostic tasks have been applied informally by humans, formal task definitions do not exist; SaMD system developers are frequently left to interpret informal definitions of diagnostic tasks themselves. As a result, manufacturers have less incentive to strive for optimal performance.

In addition to the concerns detailed above, AI-based algorithms are prone to behaving in unpredictable ways when used in settings different from the one in which they were tested. When new data is of a different quality from its validation testing, safety and efficacy are threatened. Compounding that problem, the resources to assess algorithm performance are not uniformly available, leading to variable algorithm performance when reliability and safety assurances are lacking. Finally, inherent conflicts of interest in testing and validation can undermine trust in AI algorithms when manufacturers portray their products in the most positive light possible even if they lack rigorous evaluation.

Policy Discussion

Overcoming these regulatory gaps will require governments and professional medical societies to adopt a suite of new measures. First, each diagnostic task should be defined, maintained, and updated by an entity free of conflicts of interest. We believe medical societies may be an optimal venue for this task. We propose that task definitions contain the following four elements: 1) a review of relevant background information and medical objectives; 2) a description that includes clinical assessment criteria, measurement definitions and descriptions, and the full universe of potential classification categories; 3) detailed image labeling instructions; and 4) illustrated examples and relevant counterexamples. The medical societies should specify companion references. In some cases developers may need to propose and publish their own task definition. But true standardization will require active management of the ecosystem of task definitions from medical professional societies.

> *Overcoming these regulatory gaps will require governments and professional medical societies to adopt a suite of new measures.*

We also recommend defining how to measure the performance of algorithms. Diagnostic algorithms that determine disease severity can be unpredictable. SaMD systems that perform that task have an alarming tendency not to be able to recognize problems when they arise. Thus, we suggest implementing continuous monitoring of algorithms for all relevant tasks before clinical deployment using performance measures that extend beyond accuracy.

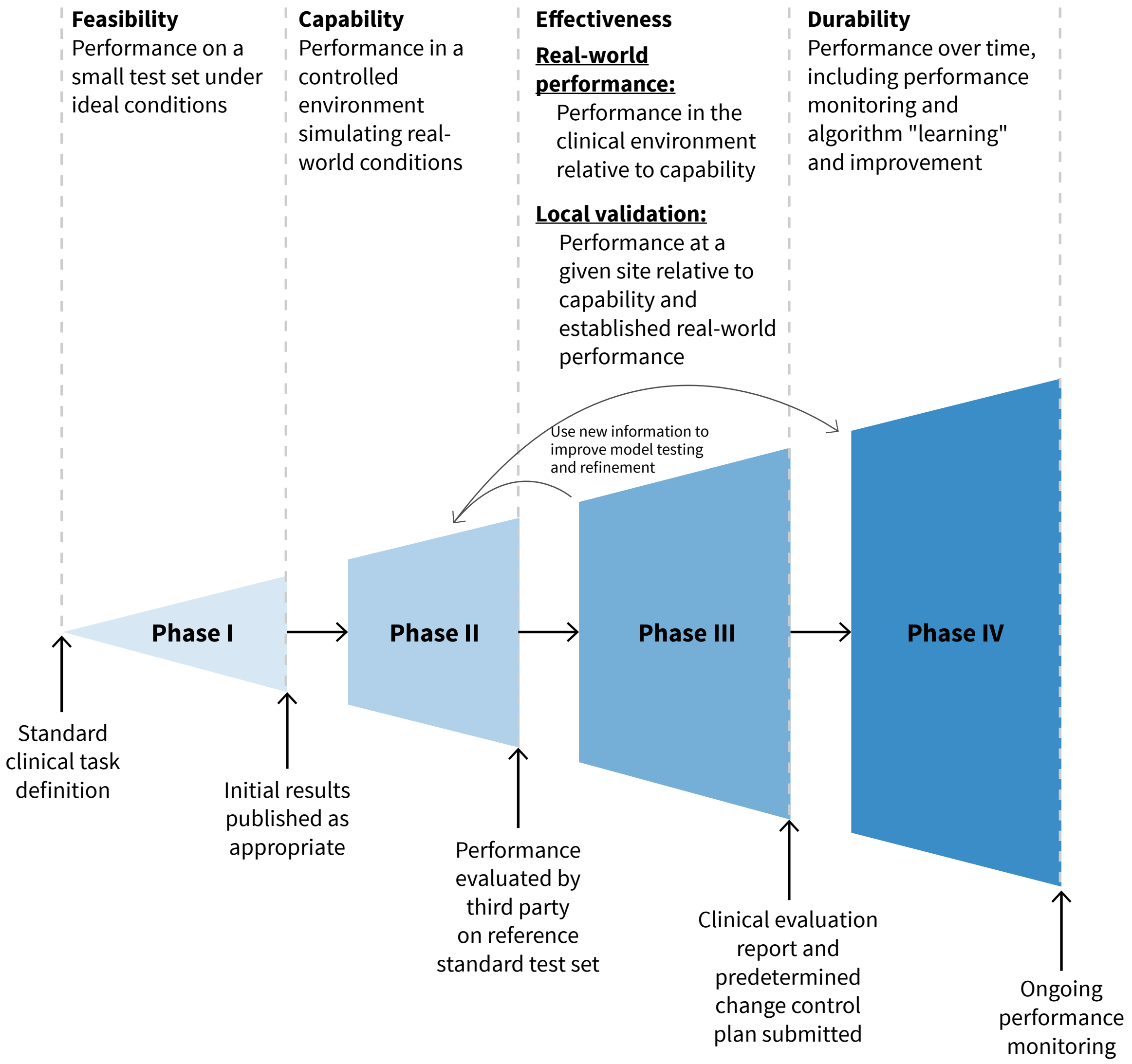
The evaluation process can be divided into five discrete mutually reinforcing steps: 1) Identify the specific diagnostic task performed by the algorithm; 2) Assess algorithm capabilities for its defined task in a controlled environment, comparing it with alternatives; 3) Compare the algorithm's lab-tested capability with real-world performance; 4) Develop specific measures of real-world effectiveness in localized settings rather than in only a few closely monitored sites; and 5) Gauge algorithm performance over time, both to maintain accuracy and improve execution.

> *We suggest implementing continuous monitoring of algorithms for all relevant tasks before clinical deployment using performance measures that extend beyond accuracy.*

Finally, we encourage third-party evaluators to perform exhaustive assessments of diagnostic algorithms collaboratively with clinical research organizations, research laboratories, and other entities that develop and maintain reference data sets. This comprehensive development and evaluation process should be incorporated into manufacturers' development cycles similar to how drug development studies are reviewed by the FDA. Like the approach used for pharmaceutical and software development, designers of AI diagnostic algorithms should subdivide their development process into feasibility, capability, effectiveness, and durability phases.

> *AI-based diagnostic algorithms will not improve patient outcomes unless the public can trust them just as they already trust other medical devices.*

We are beginning to see how AI can enhance quality of life and promote human health. Ensuring that diagnostic algorithms perform effectively both in controlled environments and in real-world settings could improve health outcomes for millions, not just in the United States but around the world. Now is the time to shape these systems' future with more thoughtful and inclusive regulatory guidance. Given the dramatic acceleration in diagnostic activity that SaMD applications will enable as they scale, the pressure on regulators likely will increase. We anticipate the growth of these new AI systems will drive the emergence of a substantial body of research rivaling the literature about algorithm development itself. But AI-based diagnostic algorithms will not improve patient outcomes unless the public can trust them just as they already trust other medical devices.



Feasibility
Performance on a small test set under ideal conditions

Capability
Performance in a controlled environment simulating real-world conditions

Effectiveness

Real-world performance:
Performance in the clinical environment relative to capability

Local validation:
Performance at a given site relative to capability and established real-world performance

Durability
Performance over time, including performance monitoring and algorithm "learning" and improvement

Use new information to improve model testing and refinement
Phase I
Phase II
Phase III
Phase IV
Standard clinical task definition
Initial results published as appropriate
Performance evaluated by third party on reference standard test set
Clinical evaluation report and predetermined change control plan submitted
Ongoing performance monitoring

David B. Larson is a professor of radiology at Stanford Medical Center, where he also serves as the vice chair for education and clinical operations in the Department of Radiology and associate chief quality officer for improvement at Stanford Health Care.

Daniel L. Rubin is a professor of biomedical data science and of radiology, director of biomedical informatics for the Stanford Cancer Institute, and faculty affiliate of Stanford's Institute for Human-Centered Artificial Intelligence.

Curtis P. Langlotz is a professor of radiology and biomedical informatics research, chair of the Health Policy Committee of Stanford's Institute for Human-Centered Artificial Intelligence (HAI), and director of Stanford's Center for Artificial Intelligence in Medicine and Imaging.

The original article, "**Regulatory Frameworks for Development and Evaluation of Artificial Intelligence–Based Diagnostic Imaging Algorithms**," can be accessed at https://www.jacr.org/article/S1546-1440(20)31020-6/fulltext

Other co-authors of the cited journal article include Hugh Harvey of the Institute for Cognitive Neuroscience, University College, in London; Neville Irani of the Department of Radiology, University of Kansas Medical Center; and Justin R. Tse, Department of Radiological Sciences, David Geffen School of Medicine, UCLA.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**