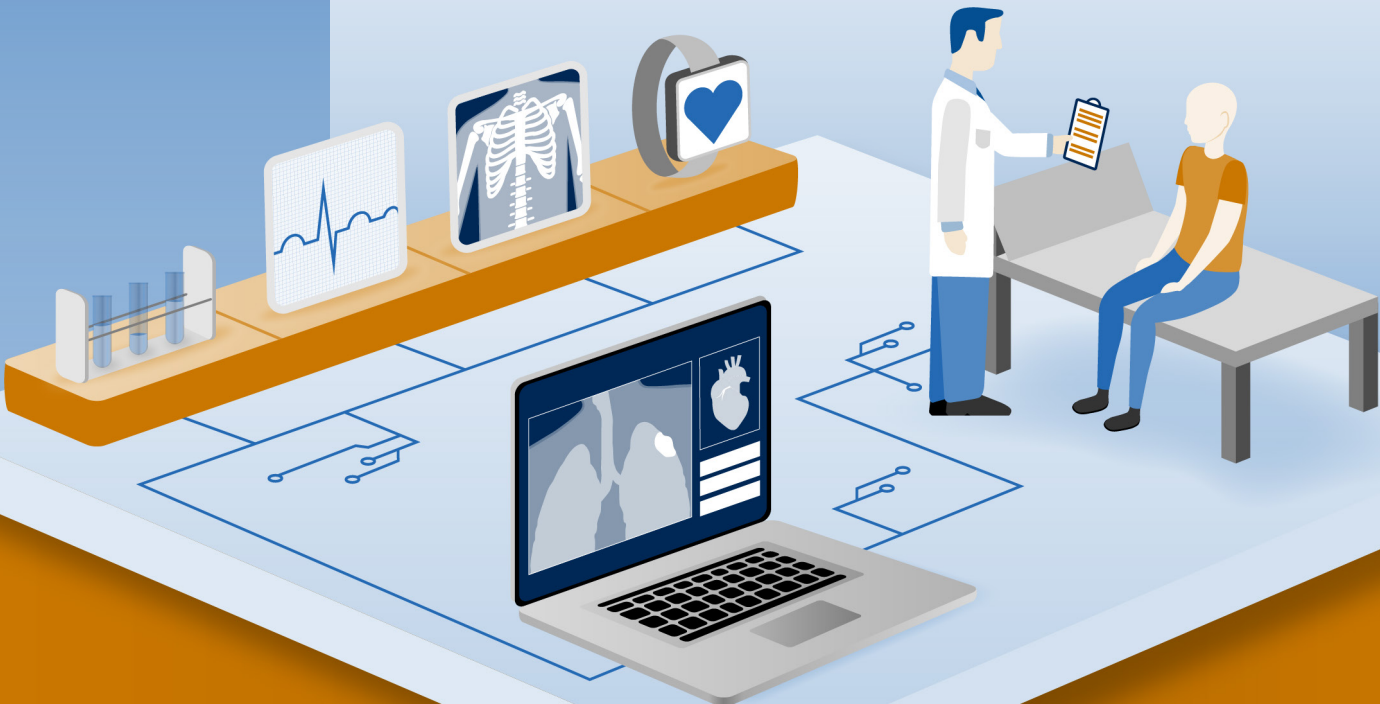


TECHNOLOGY ASSESSMENT

Artificial Intelligence in Health Care

Benefits and Challenges of Machine Learning
Technologies for Medical Diagnostics

With content from the National Academy of Medicine



The cover image displays a stylized representation of data inputs from a variety of sources—including blood testing, electrocardiogram, X-ray image, and wearable device data—to a computer representing artificial intelligence (AI) algorithms. Those algorithms then offer recommendations or other assistance to providers that could augment their ability to diagnose patients.

This report is being jointly published by the Government Accountability Office (GAO) and the National Academy of Medicine (NAM). Part One presents GAO's Technology Assessment *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics*. Part Two presents the NAM publication *Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis* discussing the factors influencing the adoption of non-autonomous point-of-care AI technology that can assist in the diagnosing of a disease. Although GAO and NAM staff consulted with and assisted each other throughout this work, reviews were conducted by NAM and GAO separately and independently, and authorship of the text of Part One and Part Two of the report lies solely with GAO and NAM, respectively.

With the exception of Part Two of this joint publication, this is a work of the U.S. government and is not subject to copyright protection in the United States. The National Academy of Medicine is the author of Part Two and waives its copyright rights for that material. However, because the joint publication may contain copyrighted images or other material, permission from the copyright holder may be necessary if you wish to reproduce this material separately.

Foreword

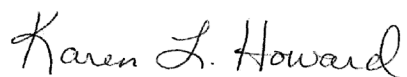
The U.S. health care system is at an important crossroads as it faces major demographic shifts, burgeoning costs, and transformative technologies. By 2030, annual health care spending in the United States is expected to reach \$6.8 trillion. The government share of this spending is projected to be 48 percent by 2030, driven by increases in Medicare enrollment, as more than 10,000 Americans become eligible for Medicare each day. These realities help illustrate the critical need to better address the effectiveness and efficiency of our nation’s health care delivery systems.

Artificial intelligence and machine learning (AI/ML) is a set of technologies that includes automated systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, and decision-making. AI/ML has promising applications in health care, including medical diagnostics. For example, it may result in earlier detection of diseases; more consistent analysis of medical data; and increased access to care, particularly for underserved populations. However, applying AI/ML technologies within the health care system also raises technological, economic, and regulatory questions.

The Government Accountability Office (GAO) and the National Academy of Medicine (NAM), individually and in collaboration, have taken up the charge to explore AI/ML in health care, assess its implications, and identify key options available for optimizing its use. In recognition of mutual interests and obligations, and to reinforce and complement each other’s work, NAM and GAO have cooperated on the development of publications on these topics over the past three years. The two most recent publications in the series are this report, GAO’s technology assessment titled *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics*, and NAM’s Special Publication titled *Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis*.

This cooperative effort included an expert meeting in which we convened a diverse, interdisciplinary, and cross-sectoral group to gather a range of perspectives on the topic. We are grateful to the exceptionally talented staff of NAM and GAO as well as the experts, all of whom worked with enthusiasm, great skill, flexibility, clarity, and drive to make this joint publication possible.

Sincerely,



Karen L. Howard, PhD
Director,
Science, Technology Assessment, and Analytics
U.S. Government Accountability Office



J. Michael McGinnis, MD, MA, MPP
Leonard D. Schaeffer Executive Officer, and
Executive Director, NAM Leadership
Consortium



Executive Summary

This report is being jointly published by the Government Accountability Office (GAO) and the National Academy of Medicine (NAM). Part One of this joint publication is the full presentation of GAO's Technology Assessment: *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics*. Part Two is the full presentation of NAM's Special Publication: *Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis*. Although GAO and NAM staff consulted with and assisted each other throughout this work, reviews were conducted by GAO and NAM separately and independently, and authorship of the text of Part One and Part Two of this Executive Summary and the following report lies solely with GAO and NAM, respectively.

OVERVIEW OF PART ONE – GAO Technology Assessment: *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics*

The GAO report *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics* is the third in a series of technology assessments that GAO conducted at the request of Congress on the use of AI technologies in health care.¹ This report discusses four topics: (1) currently available machine learning (ML) medical diagnostic technologies for five selected diseases, (2) emerging ML medical diagnostic technologies, (3) challenges affecting the development and adoption of ML technologies for medical diagnosis, and (4) policy options to help address these challenges.

Several ML technologies are available to help medical professionals diagnose the five selected diseases we examined: certain cancers, diabetic retinopathy, Alzheimer's disease, heart disease, and COVID-19. These technologies assist medical professionals by augmenting the diagnostic process, resulting in benefits that include earlier detection of diseases; more consistent analysis of medical data; and increased access to care, particularly for underserved populations. We identified a variety of ML diagnostic technologies for the diseases we examined, with most technologies relying on data from imaging. According to our expert meeting participants and interviewees, many technologies are designed to use radiology image data because such images are typically standardized and digitized. Other sources of medical data, such as tissue samples for pathology, are generally less available for training ML technologies due to additional steps including collecting and digitizing data.

Although these technologies have potential benefits and are available to assist in diagnosing the diseases we examined, they are generally not widely adopted. Companies we interviewed

¹Part One of this Joint Publication presents the GAO Technology Assessment: *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics*. Although NAM staff and leadership provided assistance and advice in the identification of issues and experts consulted during the development process (identified in app. II), the contents and resulting policy options of this technology assessment are solely those of GAO and the responsibility of GAO

reported varying levels of adoption. For example, one company with an ECG monitoring technology told us that its technology was being used in most major U.S. medical centers, while another company using a technology to detect COVID-19 infection said its technology was only being used in a handful of universities and research institutions.

Academic, government, and private sector researchers are working to expand the capabilities of ML-based medical diagnostic technologies for the five diseases we examined. We also identified three emerging approaches—autonomous, adaptive, and consumer-oriented ML diagnostics—that could enhance medical professionals’ capabilities and improve patient treatment. However, these approaches also have certain limitations. For example, adaptive ML diagnostic technologies, which update their algorithms by incorporating new patient data, may provide more accurate diagnoses or improve features for users, but changes in the algorithm data may also lead to adverse outcomes such as inconsistent or poorer algorithmic performance.

Despite the promise of these technologies, we identified challenges affecting the development and adoption of ML in medical diagnostics. These challenges, which include demonstrating real-world performance, meeting medical needs, and addressing regulatory gaps, affect technology developers, medical providers, and patients. For example, medical providers may be reluctant to adopt an ML technology until its real-world performance has been adequately demonstrated in relevant and diverse clinical settings, according to experts, stakeholders, and literature. However, developers face difficulties accessing high-quality data to validate their technologies and may not be willing to incur the significant costs needed to rigorously evaluate them. Medical providers are also less likely to adopt ML technologies that do not address a clear clinical need, such as improved accuracy or increased efficiency, and many ML diagnostic technologies do not progress from development to adoption for this reason. Lastly, gaps in the regulatory framework may also pose a challenge to the development and adoption of ML technologies, including emerging types such as adaptive algorithms. Some of these challenges are similar to those identified previously by GAO in its first and second publications in this series.²

In this report, GAO describes three options that policymakers—which GAO defines broadly to include Congress, federal agencies, state and local governments, academic and research institutions, and industry, among others—could use in addressing the challenges for medical diagnostics technologies:

- **Evaluation.** Policymakers could create incentives, guidance, or policies to encourage or require the evaluation of ML diagnostic technologies across a range of deployment conditions and demographics representative of the intended use. This could help address the challenge of demonstrating real world performance.

²GAO, *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development*, GAO-20-215SP (Washington, D.C.: Dec. 20, 2019). GAO, *Artificial Intelligence in Health Care: Benefits and Challenges of Technologies to Augment Patient Care*, GAO-21-75P (Washington, D.C.: Nov. 30, 2020).

- **Data Access.** Policymakers could develop or expand access to high-quality medical data to develop and test ML medical diagnostic technologies. Examples include standards for collecting and sharing data, creating data commons, or using incentives to encourage data sharing. This could help address the challenge of demonstrating real world performance.
- **Collaboration.** Policymakers could promote collaboration among developers, providers, and regulators in the development and adoption of ML diagnostic technologies. For example, policymakers could convene multidisciplinary experts together in the design and development of these technologies through workshops and conferences. This could help address the challenges of meeting medical needs and addressing regulatory gaps.

OVERVIEW OF PART TWO – NAM: *Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis*

Artificial intelligence (AI) carries significant potential in aiding clinical diagnostic decision-making. AI-assisted diagnostic decision support tools (AI-DDS), given their processing power and continuous learning capabilities, can synthesize large volumes of data and perform advanced pattern analysis tasks to make diagnostic processes more effective, efficient, and accurate. While AI-DDS systems grow increasingly sophisticated and robust, their practical value and broader success are contingent upon their adoption by health care providers. To this end, the authors present a framework for evaluating and promoting provider adoption of new AI-DDS tools, centered on four integrated domains: 1) Reason to Use, 2) Means to Use, 3) Methods to Use, and 4) Desire to Use.

Domain 1, Reason to Use: An initial consideration of the adoption of a novel AI-DDS tool is based on its alignment with the patient care missions of health systems and providers. The tool must cater to an important clinical need or gap and ultimately contribute to improved patient outcomes. Having demonstrated clinical utility, integrating, deploying, and maintaining the tool in clinical workflows requires significant financial investment. Therefore, a second critical factor in the adoption of a new AI-DDS system is its overall affordability and value proposition to the health system, provider, and patients, including appropriate insurance coverage for the use of a given tool.

Domain 2, Means to Use: Once adopted by a health system, a new AI-DDS systems requires robust infrastructure to support the efficient and sustainable implementation of the tool. The first set of infrastructural elements is the computing hardware and software needed to 1) collect and organize relevant health data used by the algorithm, 2) construct and validate an AI algorithm at the point of care, and 3) conduct regular maintenance, including troubleshooting of technical problems. The second set is the human and operational resources needed to effectively maintain AI-DDS systems. Key roles include, but are not limited to, frontline IT staff, data architects, and AI-machine learning specialists to understand the context of use and tailor the solution to be fit for purpose. The infrastructure also requires information security and data

privacy officers, legal and industrial contract officers for business and data use agreements, and IT educators to train and retrain providers and staff.

Domain 3, Methods to Use: Clinical operations differ substantially among health care systems, medical specialties, patient populations, and geographic areas. Therefore, operationalizing and scaling new AI-DDS technologies, including AI-DDS within and across health systems, can be expensive and complex. Effective integration of new AI-DDS tools into existing clinical workflows is essential. Equally critical is thoughtful development and deployment of new tools to optimize workflow efficiency and enable providers to prioritize cognitive and emotional energy for patient interactions. Additionally, AI-DDS must be deliberately designed to minimize detracting from the diagnostic process, including limiting interface distractions and data obfuscation. Finally, health systems must ensure the technical proficiency of providers in relation to new AI-DDS tools with in-depth onboarding training and continuous medical education.

Domain 4, Desire to Use: It is important to attend to psychological factors surrounding the use of AI-DDS, such as addressing how these tools can facilitate professional fulfillment among providers, including mitigating burnout while enabling provider autonomy. The other indispensable element within the desire to use core domain is trust, including the legal and ethical considerations of AI-DDS systems. Two significant sources of distrust relevant to the adoption of AI-DDS tools by clinicians explored in this paper are 1) bias (real or perceived) and 2) liability. Clinicians must be able to trust that these products can deliver quality care outcomes for their patients without creating harm or error and align with both patients' and clinicians' ethics and values. A key factor affecting clinicians' willingness to adopt AI-DDS tools is whether the tools will receive a rigorous, data-driven review of safety and effectiveness before moving into clinical use.

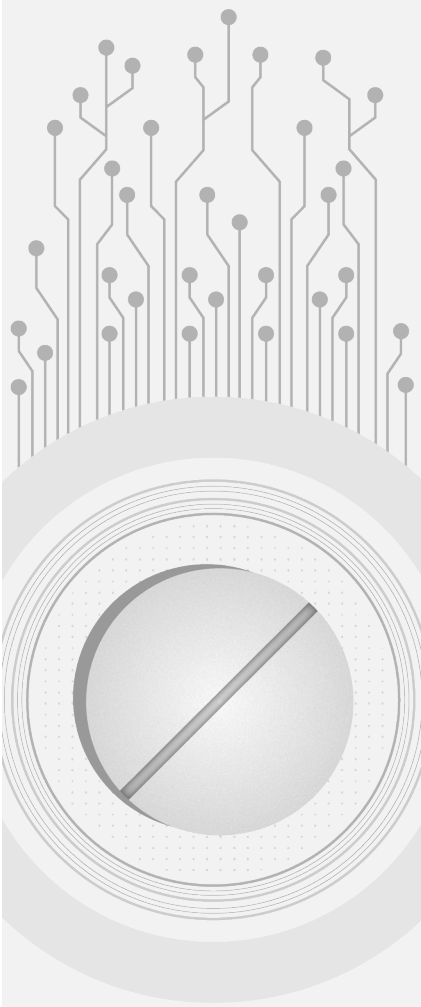
Crosscutting these considerations is the need to be cognizant of the equity implications that accompany the adoption of AI-DDS tools. While there is excitement and demonstrated benefits to bringing AI-DDS tools into clinical practice, poor data quality and prevalent biases in health care can jeopardize progress towards achieving health equity and fuel ongoing uncertainties and hesitations about adopting these tools. In addition, preventing widening disparities in the implementation of AI-DDS tools will require addressing the digital gap by developing and implementing infrastructure that will support the equitable use of AI.

AI-DDS systems are becoming increasingly prevalent, sophisticated, and reliable. Across medical specialties, these tools demonstrate potential to make the clinical diagnostic process more efficient and accurate, ultimately improving patient outcomes. Focused efforts to create equitable and robust AI-DDS algorithms, streamline integration of new AI-DDS tools into clinical workflows, and train health care providers to effectively use such tools – coupled with strong regulatory oversight and financial incentives – will optimize the likelihood that innovative, clinically impactful AI-DDS systems are adopted and used responsibly by health care providers to the ultimate benefit of their patients.

Table of Contents

Foreword.....	i
Executive Summary.....	ii
Part One—Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics	viii
Introduction	1
1 Background	3
1.1 Medical diagnosis and diagnostic tests	3
1.2 Selected diseases	5
1.3 Machine learning (ML).....	7
1.4 Roles and responsibilities in the development of ML medical diagnostic technologies...7	
2 Available ML Diagnostic Technologies.....	10
2.1 Potential benefits of ML diagnostic technologies	10
2.2 ML diagnostic technologies by disease	11
2.3 Adoption	14
3 Emerging ML Diagnostic Technologies	16
3.1 Emerging improvements to ML diagnostic technologies, by disease	16
3.2 Emerging approaches to ML diagnostic technologies across diseases	17
4 Challenges Affecting ML Technologies for Medical Diagnostics	23
4.1 Demonstrating real-world performance	23
4.2 Meeting medical needs	25
4.3 Addressing regulatory gaps	26
5 Policy Options to Enhance Benefits or Address Challenges of ML Diagnostic Technologies	28
5.1 Policy Option: Evaluation.....	28
5.2 Policy Option: Data Access	29
5.3 Policy Option: Collaboration.....	30
6 Agency and Expert Comments.....	32

Part Two—(NAM) Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis	35
Introduction	35
1 A Primer on AI-Diagnostic Decision Support Tools	37
2 Facilitating Provider Adoption of AI-Diagnostic Decision Support Tools	41
3 Ensuring and Promoting Health Equity in the Deployment of AI-Assisted Diagnostic Tools.....	64
4 Path Forward – Policy Implications and Action Priorities	66
References	69
Appendix I: Objectives, Scope, and Methodology	83
Appendix II: Expert Participation	86
Appendix III: GAO Contacts and Staff Acknowledgments	88



PART ONE

Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics

U. S. Government Accountability Office (GAO)

Part One of this Joint Publication presents the GAO Technology Assessment: *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics*. Although NAM staff and leadership provided assistance and advice in the identification of issues and expertise consulted during the development process (identified in Appendix II), responsibility for the text, findings, and options lies solely with GAO.

Why GAO did this study

Diagnostic errors affect more than 12 million Americans each year, with aggregate costs likely in excess of \$100 billion, according to a report by the Society to Improve Diagnosis in Medicine. ML, a subfield of artificial intelligence, has emerged as a powerful tool for solving complex problems in diverse domains, including medical diagnostics. However, challenges to the development and use of machine learning technologies in medical diagnostics raise technological, economic, and regulatory questions.

GAO was asked to conduct a technology assessment on the current and emerging uses of machine learning in medical diagnostics, as well as the challenges and policy implications of these technologies. This report discusses (1) currently available ML medical diagnostic technologies for five selected diseases, (2) emerging ML medical diagnostic technologies, (3) challenges affecting the development and adoption of ML technologies for medical diagnosis, and (4) policy options to help address these challenges.

GAO assessed available and emerging ML technologies; interviewed stakeholders from government, industry, and academia; convened a meeting of experts in collaboration with the National Academy of Medicine; and reviewed reports and scientific literature. GAO is identifying policy options in this report.

View [GAO-22-104629](#). For more information, contact Karen L. Howard at (202) 512-6888 or howardk@gao.gov.

ARTIFICIAL INTELLIGENCE IN HEALTH CARE

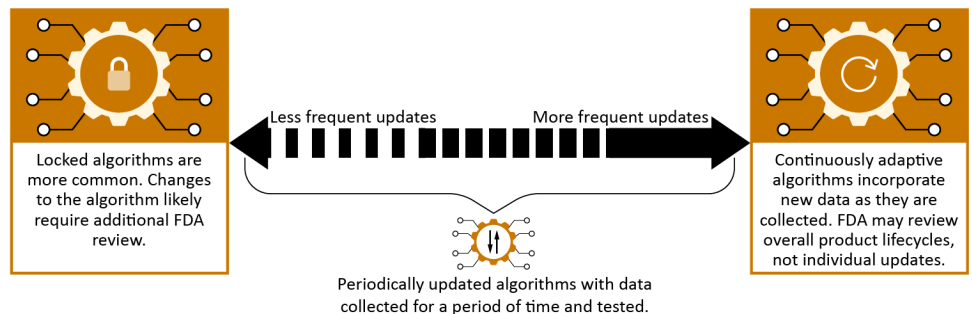
Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics

What GAO found

Several machine learning (ML) technologies are available in the U.S. to assist with the diagnostic process. The resulting benefits include earlier detection of diseases; more consistent analysis of medical data; and increased access to care, particularly for underserved populations. GAO identified a variety of ML-based technologies for five selected diseases — certain cancers, diabetic retinopathy, Alzheimer’s disease, heart disease, and COVID-19 —with most technologies relying on data from imaging such as x-rays or magnetic resonance imaging (MRI). However, these ML technologies have generally not been widely adopted.

Academic, government, and private sector researchers are working to expand the capabilities of ML-based medical diagnostic technologies. In addition, GAO identified three broader emerging approaches—autonomous, adaptive, and consumer-oriented ML-diagnostics—that can be applied to diagnose a variety of diseases. These advances could enhance medical professionals’ capabilities and improve patient treatments but also have certain limitations. For example, adaptive technologies may improve accuracy by incorporating additional data to update themselves, but automatic incorporation of low-quality data may lead to inconsistent or poorer algorithmic performance.

Spectrum of adaptive algorithms



Source: GAO analysis of Food and Drug Administration (FDA) information. | GAO-22-104629

We identified several challenges affecting the development and adoption of ML in medical diagnostics:

- Demonstrating real-world performance across diverse clinical settings and in rigorous studies.
- Meeting clinical needs, such as developing technologies that integrate into clinical workflows.
- Addressing regulatory gaps, such as providing clear guidance for the development of adaptive algorithms.

These challenges affect various stakeholders including technology developers, medical providers, and patients, and may slow the development and adoption of these technologies.

GAO developed three policy options that could help address these challenges or enhance the benefits of ML diagnostic technologies. These policy options identify possible actions by policymakers, which include Congress, federal agencies, state and local governments, academic and research institutions, and industry. See below for a summary of the policy options and relevant opportunities and considerations.

Policy Options to Help Address Challenges or Enhance Benefits of ML Diagnostic Technologies

	Opportunities	Considerations
<p>Evaluation (report page 28)</p> <p>Policy makers could create incentives, guidance, or policies to encourage or require the evaluation of ML diagnostic technologies across a range of deployment conditions and demographics representative of the intended use.</p> <p><i>This policy option could help address the challenge of demonstrating real world performance.</i></p>	<ul style="list-style-type: none"> Stakeholders could better understand the performance of these technologies across diverse conditions and help to identify biases, limitations, and opportunities for improvement. Could inform providers' adoption decisions, potentially leading to increased adoption by enhancing trust. Information from evaluations can help inform the decisions of policymakers, such as decisions about regulatory requirements. 	<ul style="list-style-type: none"> May be time-intensive, which could delay the movement of these technologies into the marketplace, potentially affecting patients and professionals who could benefit from these technologies. More rigorous evaluation will likely lead to extra costs, such as direct costs for funding the studies. Developers may not be incentivized to conduct these evaluations if it could show their products in a negative light, so policymakers could consider whether evaluations should be conducted or reviewed by independent parties, according to industry officials.
<p>Data Access (report page 29)</p> <p>Policy makers could develop or expand access to high-quality medical data to develop and test ML medical diagnostic technologies. Examples include standards for collecting and sharing data, creating data commons, or using incentives to encourage data sharing.</p> <p><i>This policy option could help address the challenge of demonstrating real world performance.</i></p>	<ul style="list-style-type: none"> Developing or expanding access to high-quality datasets could help facilitate training and testing ML technologies across diverse and representative conditions. This could improve the technologies' performance and generalizability, help developers understand their performance and areas for improvement, and help to build trust and adoption in these technologies. Expanding access could enable developers to save time in the development process, which could shorten the time it takes for these technologies to be available for adoption. 	<ul style="list-style-type: none"> Entities that own data may be reluctant to share them for a number of reasons. For example, these entities may consider their data valuable or proprietary. Some entities may also be concerned about the privacy of their patients and the intended use and security of their data. Data sharing mechanisms may be of limited use to researchers and developers depending on the quality and interoperability of these data, and curating and storing data could be expensive and may require public and private resources.
<p>Collaboration (report page 30)</p> <p>Policy makers could promote collaboration among developers, providers, and regulators in the development and adoption of ML diagnostic technologies. For example, policymakers could convene multidisciplinary experts together in the design and development of these technologies through workshops and conferences.</p> <p><i>This policy option could help address the challenges of meeting medical needs and addressing regulatory gaps.</i></p>	<ul style="list-style-type: none"> Collaboration between ML developers and providers could help ensure that the technologies address clinical needs. For example, collaboration between developers and medical professionals could help developers create ML technologies that integrate into medical professionals' workflows, and minimize time, effort, and disruption. Collaboration among developers and medical providers could help in the creation and access of ML ready data, according to NIH officials. 	<ul style="list-style-type: none"> As previously reported, providers may not have time to both collaborate with developers and treat patients; however, organizations can provide protected time for employees to engage in innovation activities such as collaboration. If developers only collaborate with providers in specific settings, their technologies may not be usable across a range of conditions and settings, such as across different patient types or technology systems.

Source: GAO. | GAO-22-104629

Abbreviations

AI	artificial intelligence
CDC	Centers for Disease Control and Prevention
CT	computed tomography
DOE	Department of Energy
ECG, EKG	electrocardiogram
FDA	Food and Drug Administration
FFDCA	Federal Food, Drug, and Cosmetic Act
FTC	Federal Trade Commission
HHS	Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act of 1996
ML	machine learning
MRI	magnetic resonance imaging
NIH	National Institutes of Health
NAM	National Academy of Medicine
SaMD	software as a medical device
VA	Department of Veterans Affairs
VHA	Veterans Health Administration



September 29, 2022

Congressional Requesters

Effective treatment depends on accurate and timely diagnosis which explains a patient’s health problem and informs treatment. Diagnostic errors are the most common, catastrophic, and costly of medical errors, with annual aggregate costs likely in excess of \$100 billion, according to a report by the Society to Improve Diagnosis in Medicine.³ Citing a 2014 study, the report states diagnostic errors affect more 12 million Americans each year, with perhaps one-third of those suffering serious harms.⁴ Further, a National Academy of Medicine report on improving diagnosis states that diagnostic errors contribute to approximately 10 percent of patient deaths and 6 to 17 percent of adverse events in hospitals.⁵

Artificial intelligence (AI) has emerged as a powerful tool for solving complex problems in diverse domains.⁶ Machine learning (ML), a subfield of AI, could revolutionize diagnosis by augmenting clinical diagnostics practice resulting in earlier and better diagnoses, lives saved, and avoided costs of time and money. In recent years, for example, ML technology was reported to be equivalent to medical professionals in interpreting medical data from fields like radiology and dermatology.⁷ ML technology can assist medical professionals in completing repetitive tasks without getting tired, and flagging potential medical issues at the point of care.

However, challenges to the development and use of ML medical diagnostic technologies (diagnostic) raise technological, economic, and regulatory questions. For example, as we have

³Society to Improve Diagnosis in Medicine, “The Roadmap for Research to Improve Diagnosis, Part 1: Converting National Academy of Medicine Recommendations into Policy Action” (February 7, 2018), accessed January 28, 2022, https://www.improvediagnosis.org/wp-content/uploads/2018/10/policy_roadmap_for_diagnosti.pdf.

⁴H. Singh, A.N.D. Meyer and E.J. Thomas, “The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations,” *BMJ Quality & Safety*, 23 (2014): 727-731.

⁵National Academies of Sciences, Engineering, and Medicine. *Improving Diagnosis in Health Care*. (Washington, DC: The National Academies Press, 2015). <https://doi.org/10.17226/21794>.

⁶Section 5002 of the National Defense Authorization Act for Fiscal Year 2021, defines AI as: a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. AI systems use machine and human-based inputs to—(A) perceive real and virtual environments; (B) abstract such perceptions into models through analysis in an automated manner; and (C) use model inference to formulate options for information or action. William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021 (NDAA FY21), Pub. L. No. 116-283, § 5002, 134 Stat. 3388 (2021). The National Artificial Intelligence Initiative Act of 2020, was enacted as Division E of the NDAA FY21. For additional characteristics of AI, see GAO, *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, GAO-21-519SP (Washington, D.C.: June 30, 2021).

⁷Esteva, A., Chou, K., Yeung, S. *et al.* Deep learning-enabled medical computer vision. *npj Digital. Medicine*, 4, 5 (2021). <https://doi.org/10.1038/s41746-020-00376-2>.

previously reported, AI tools developed using historical data could unintentionally perpetuate biases, reduce safety and effectiveness for different groups of patients, and produce disparities in treatment.⁸

In view of the potential for ML in diagnostics, you asked us to conduct a technology assessment in this area. This report discusses (1) currently available ML diagnostics technologies, (2) emerging ML diagnostics technologies, (3) challenges affecting the development and adoption of ML technologies for medical diagnosis, and (4) policy options to address these challenges. See appendix I for additional information on our scope and methodology.

We conducted our work from November 2020 through September 2022 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement to obtain sufficient and appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

⁸GAO, *Artificial Intelligence in Health Care: Benefits and Challenges of Technologies to Augment Patient Care*, GAO-21-7SP (Washington, D.C.: Nov. 30, 2020).

Part One—(GAO) Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics

1 Background

1.1 Medical diagnosis and diagnostic tests

Medical diagnosis is a key step in patient care, in which medical professionals use patient history, symptoms, and test results to characterize and understand a patient's health problems and inform treatment plans. An accurate and timely diagnosis can significantly improve a patient's chance for positive health outcomes.





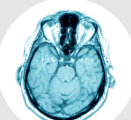
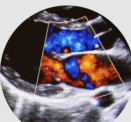

According to a medical journal, medical professionals can use diagnostic testing for a number of purposes, such as obtaining a diagnosis, monitoring the effectiveness of therapeutic interventions, or conducting disease surveillance.⁹ Diagnostic testing can also be used to screen patients to help identify a condition before signs and symptoms become apparent.¹⁰ For example, certain imaging tests can help identify coronary artery disease by indicating the presence of coronary artery blockage, even in the absence of symptoms like chest pain.

Medical professionals use a variety of diagnostic tests to inform diagnosis; different tests can generate different types of information depending on the physiological system being examined. Figure 1 describes examples of diagnostic tests that medical professionals use to help diagnose selected diseases.

⁹K.A. Fleming, S. Horton, M.L. Wilson, R. Atun, K. DeStigter, J. Flanigan, S. Sayed et al. "The Lancet Commission on diagnostics: Transforming access to diagnostics." *The Lancet* 398, no. 10315 (2021): 1997-2050.

¹⁰Screening tests are functionally similar to diagnostic tests, and can use the same types of tests and technologies.

Figure 1: Description of diagnostic tests for selected diseases

	Test	Description
	<ul style="list-style-type: none"> ▶ Blood test 	<ul style="list-style-type: none"> ▶ A small sample of blood is taken from the body and analyzed to characterize the different parts of blood, including chemicals and proteins.
	<ul style="list-style-type: none"> ▶ Computed tomography (CT) 	<ul style="list-style-type: none"> ▶ Many X-ray images (see entry for X-ray below) are recorded as the detector moves around the patient's body. A computer reconstructs all the individual images into cross-sectional images or "slices" of the patient's internal organs and tissues.
	<ul style="list-style-type: none"> ▶ Electrocardiogram (ECG, EKG) 	<ul style="list-style-type: none"> ▶ Electrodes placed on the patient's body record the electrical activity of the heart. The results are analyzed to measure any damage to the heart, determine whether the heart is beating normally, or measure the size and position of the heart's chambers.
	<ul style="list-style-type: none"> ▶ Fundus (e.g. retinal) photography 	<ul style="list-style-type: none"> ▶ A specialized camera captures a photograph of the interior surface of the eye to document conditions like diabetic retinopathy and retinal detachment.
	<ul style="list-style-type: none"> ▶ Magnetic resonance imaging (MRI) 	<ul style="list-style-type: none"> ▶ A non-invasive imaging technology using magnetic fields and radio waves to produce three-dimensional, detailed anatomical images. MRI is preferred when frequent imaging is required for diagnosis or therapy because it does not use x-rays or other radiation.
	<ul style="list-style-type: none"> ▶ Ultrasound 	<ul style="list-style-type: none"> ▶ A non-invasive imaging technology using sound waves with frequencies above the threshold of human hearing to produce images of internal organs. Additionally, diagnostic ultrasound results can help visualize changes or differences in function within an organ.
	<ul style="list-style-type: none"> ▶ X-ray 	<ul style="list-style-type: none"> ▶ Patients are exposed to a type of electromagnetic radiation which results in pictures of the inside of body in different shades of black and white. X-rays can be used to check for broken bones, and in other ways, such as in mammograms.

Source: GAO analysis of interviews, literature, and documentation (text); GAO and Tryfonov/iztverichka/ververidis/stock.adobe.com (icons). | GAO-22-104629

1.2 Selected diseases

Six in 10 Americans live with at least 1 chronic condition, such as cancer, diabetes, Alzheimer's disease or heart disease according to a journal article.^{11,12} The article further identifies chronic diseases represent seven of the 10 causes of death in the U.S., are the leading causes of disability in the U.S., and are the leading drivers of the nation's annual health care spending. We explored chronic diseases that may benefit from ML






medical diagnostic technologies, as well as COVID-19 because of the continuing pandemic. From this examination, we focused this technology assessment on the following five diseases: the top three causes of death in 2020 - heart disease, cancer, and COVID-19; a leading cause of disability - Alzheimer's disease; and the leading cause of adult blindness - diabetic retinopathy.¹³

¹¹Chronic diseases are defined broadly as conditions that last 1 year or more and require ongoing medical attention or limit activities of daily living or both according to CDC.

¹²K.A. Hacker, P.A. Briss, L. Richardson, J. Wright, and R. Petersen, "COVID-19 and Chronic Disease: The Impact Now and in the Future," *Preventing Chronic Disease*, 18 (2021).

¹³Although the results of our assessment cannot be generalized to other diseases, these diseases reflect a range of disease types and the potential for distinct challenges to the use of machine learning for diagnosis.

Figure 2: Selected diseases and their impacts in the U.S.

	Disease	Description	Effect in the U.S.
	<ul style="list-style-type: none"> ▶ Cancer 	<ul style="list-style-type: none"> ▶ A disease of more than 100 types in which cells begin to grow out of control and spread into surrounding tissues. Each type is usually named for the part of the body where it started. We focus on five specific cancers representing the highest age-adjusted mortality in the U.S., according to 2018 data: <ul style="list-style-type: none"> ▶ Lung and bronchus ▶ Breast (female) ▶ Prostate ▶ Colon and rectum ▶ Pancreas 	<ul style="list-style-type: none"> ▶ The second leading cause of death, with more than 1.8 million people estimated to be diagnosed with new cases of cancer in 2020. In 2019, the national economic burden associated with cancer care (including patient out-of-pocket costs and patient time costs) was over \$21 billion, according to the Annual Report to the Nation on the Status of Cancer Part 2.
	<ul style="list-style-type: none"> ▶ Diabetic retinopathy 	<ul style="list-style-type: none"> ▶ An eye condition in which high blood sugar, caused by diabetes, damages blood vessels in the retina, causing vision loss and blindness. 	<ul style="list-style-type: none"> ▶ The leading cause of new cases of blindness in adults. Early diagnosis and timely treatment of diabetic retinopathy can reduce the risk of vision impairment or loss. However, CDC estimates diabetes-related blindness costs about \$500 million annually, and as many as 50 percent of patients are not getting their eyes examined or are diagnosed too late for treatment to be effective.
	<ul style="list-style-type: none"> ▶ Alzheimer’s disease 	<ul style="list-style-type: none"> ▶ A degenerative brain disease that slowly destroys memory skills, and eventually affects the ability to carry out basic bodily functions like walking and swallowing. It is the most common cause of dementia in older adults accounting for 60-80 percent of dementia cases according to the Alzheimer’s Association. Scientists do not yet fully understand what causes Alzheimer’s disease in most cases. 	<ul style="list-style-type: none"> ▶ The seventh leading cause of death and a leading cause of disability in 2020. As of 2021, NIH estimated that more than 6 million Americans had Alzheimer’s disease. The Alzheimer’s Association estimated 2020 costs for Alzheimer’s or other dementias at \$305 billion, not including the value of informal caregiving.
	<ul style="list-style-type: none"> ▶ Heart disease 	<ul style="list-style-type: none"> ▶ The most common type of heart disease—coronary heart disease—develops when arteries cannot deliver enough oxygen-rich blood to the heart because of the buildup of plaque, which impedes blood flow. 	<ul style="list-style-type: none"> ▶ The leading cause of death in the U.S. As of 2021, more than 1 of every 9 American adults had been diagnosed with heart diseases according to the National Institutes of Health (NIH). Heart disease cost the U.S. about \$363 billion each year from 2016 to 2017, according to the Centers for Disease Control and Prevention (CDC).
	<ul style="list-style-type: none"> ▶ COVID-19 	<ul style="list-style-type: none"> ▶ COVID-19 is a disease caused by a coronavirus called SARS-CoV-2. A wide range of symptoms have been reported: fever or chills, cough, difficulty breathing, fatigue, and new loss of taste or smell. Although most people with COVID-19 get better within weeks of illness, some people experience post-COVID conditions, according to the CDC. “Long COVID,” also known as post-COVID conditions, can be considered a disability under the Americans with Disabilities Act of 1990, according to the Department of Health and Human Services. 	<ul style="list-style-type: none"> ▶ We previously reported that at the beginning of January 2022, the U.S. had about 56 million reported cases of COVID-19 and over 830,000 reported deaths, according to CDC. Additionally, the federal government had spent \$3.5 trillion of \$4.0 trillion obligated for COVID-19 relief, as of November 30, 2021. This spending included non-healthcare costs such as unemployment insurance and business loans.⁹

Source: GAO review of literature and agency documentation (text); GAO (icons). | GAO-22-104629

Note: Cause of death rankings in this table are from provisional 2020 CDC data.

⁹GAO, COVID-19: Significant Improvements Are Needed for Overseeing Relief Funds and Leading Responses to Public Health Emergencies, [GAO-22-105291](#) (Washington, D.C.: January 27, 2022).

1.3 Machine learning (ML)

ML is the leading AI approach in recent diagnostics development.¹⁴ ML technologies are trained (see text box) by processing data to identify patterns that may be hidden or complex. ML relies on large amounts of data for this training process. The increased availability of such data has enabled many recent ML advances, such as in image recognition.¹⁵

Selected methods to train ML algorithms

Supervised ML. An algorithm is provided with labeled data to identify logical patterns in the data and use those patterns to predict a specified answer to a problem. For example, an algorithm trained on many labeled images of malignant (cancerous) or benign lesions could then classify whether a new unlabeled image with a lesion is cancerous.

Unsupervised ML. An algorithm is provided with unlabeled data to allow the algorithm to identify structure in the data, for example by clustering similar data, without a preconceived idea of what clusters to expect. In this technique, an algorithm could cluster images into groups based on similar features, such as a group of malignant lesion images and a group of benign lesion images, without the images in the training set being labeled as cancerous or not.

Source: GAO-21-75P. | GAO-22-104629

1.4 Roles and responsibilities in the development of ML medical diagnostic technologies

Three groups of stakeholders are involved in the development and deployment of ML medical diagnostic technologies: research and development entities, end users, and regulators.^{16,17}

Research and development. Research and development for ML diagnostics continues across multiple stakeholder groups. For example, one study by an academic team developed an ML technology to assist pathologists to differentiate between subtypes of liver cancer. Additionally, there are commercial efforts to bring ML diagnostics to market. For example, one company markets an AI-based technology to assist pathologists in detecting prostate cancer.¹⁸

At the federal level, NIH develops ML technologies as well as collaborates with others to study such technologies.¹⁹

According to NIH's website, ML technologies are being developed across all 27 of its institutes and centers. For example, NIH's National Institute on Aging's Artificial Intelligence for Alzheimer's Disease Initiative

¹⁴For our analysis, we focused on ML methods relying on statistical learning using observed or simulated data. One ML algorithm is an artificial neural network; inspired by the brain, it contains an input layer that receives data, hidden layers that process data, and an output layer. Deep learning uses deep neural networks, which contain a large number of hidden layers.

¹⁵For additional information on other advances in AI, see *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications*, GAO-18-142SP (Washington, D.C.: March 2018).

¹⁶For the remainder of this report, we will refer to ML medical diagnostic technologies as ML diagnostics.

¹⁷Though not always directly involved in the development and deployment of ML diagnostics, patients, who are affected by medical diagnostic decision, are also stakeholders.

¹⁸See chapters 2 and 3 for additional commercial examples.

¹⁹NIH is an agency within the Department of Health and Human Services (HHS).

aims to leverage ML technologies to support diagnostics.²⁰

Additionally, NIH's National Cancer Institute collaborates with the Department of Energy (DOE) and several DOE national laboratories in the Joint Design of Advanced Computing Solutions for Cancer program. The stated goals include understanding the impact of new diagnostics through the application of advanced computational capabilities to population-based cancer data.

End users. The end users of diagnostic technologies, including those using ML, are generally medical professionals, including nurses, doctors, and others. Medical professionals apply clinical reasoning skills as they collect and integrate information from a patient's history, interview, physical exam, diagnostic testing, and consultations with or referrals to other medical professionals.

At the federal level, government agencies are at various stages of adopting ML-based diagnostic tests. For example, the Department of Veterans Affairs (VA), through its Veterans Health Administration (VHA), operates the largest integrated health care system in the U.S. and provides diagnostic services in support of 9 million enrolled veterans. VHA diagnostic services include clinical services of pathology and laboratory medicine, radiology, and nuclear medicine.

²⁰NIH Grant PAR-19-269 "Cognitive Systems Analysis of Alzheimer's Disease Genetic and Phenotypic Data (U01 Clinical Trial Not Allowed)", <https://reporter.nih.gov/project-details/10028746>

²¹See 21 U.S.C. § 360c. High-risk devices generally require FDA premarket review and approval to determine whether the device meets the statutory standard of reasonable assurance of safety and effectiveness for its intended use. See 21 U.S.C. § 360e(c). Moderate-risk and some lower-risk devices may require premarket clearance, whereby sponsors demonstrate

VA officials provided an example of one VA facility that recently started using AI to detect hemorrhages. According to these officials, the process for adopting diagnostic technologies, including those using ML, is unique to each VHA facility and depends on local mechanisms and funding. In addition, the Department of Defense's Defense Innovation Unit reported it was working with the Defense Health Agency in training ML to help diagnose cancer.

Regulators. The Food and Drug Administration (FDA) has a role in regulating medical devices under the Federal Food, Drug, and Cosmetic Act (FFDCA) and implementing regulations. Specifically, FDA generally approves or clears devices before they can be marketed in the United States, in accordance with the level of risk the device poses to patients or users.²¹ In 2019, FDA issued a proposed regulatory framework for machine learning-based software as a medical device (SaMD), but it has not yet promulgated any regulations.²²

The Federal Trade Commission (FTC) also has a role in protecting consumers from false and deceptive advertising, such as by evaluating whether health claims made in an advertisement are substantiated and

that the new device is substantially equivalent to a device already on the market. See 21 U.S.C. § 360(k). For the purposes of our report, we refer to devices that have been approved or cleared by FDA as "authorized devices."

²²FDA, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)," Washington, D.C.: Apr. 2, 2019.

truthful.²³ For example, in 2015, FTC challenged and reached settlements with two marketers of mobile apps deceptively claiming to use algorithms that could detect symptoms of melanoma, a form of skin cancer.²⁴ FTC also coordinates with FDA to

protect consumers from deceptive advertising and labeling; for example, FTC officials told us these agencies may issue joint warning letters to companies making deceptive claims.

²³15 U.S.C. § 45(a)(1) and (a)(2), and Federal Trade Commission, FTC Policy Statement on Deception, appended to *Cliffdale Associates, Inc.*, 103 F.T.C. 110, 174 (1984), at 1 (1983).

²⁴Federal Trade Commission (FTC), “FTC Cracks Down on Marketers of “Melanoma Detection” Apps” (Washington, D.C.: February 23, 2015), accessed September 24, 2021, <https://www.ftc.gov/news-events/press-releases/2015/02/ftc-cracks-down-marketers-melanoma-detection-apps>

2 Available ML Diagnostic Technologies

Several ML diagnostic technologies are available in the U.S. These technologies assist medical professionals by augmenting the diagnostic process, resulting in benefits that include earlier detection of diseases; more consistent analysis of medical data; and increased access to care, particularly for underserved populations. We identified a variety of ML diagnostic technologies for the diseases we examined, with most technologies relying on data from imaging. However, these technologies have generally not been widely adopted.

2.1 Potential benefits of ML diagnostic technologies

Several ML technologies are available to help medical professionals diagnose the selected diseases we examined. These technologies typically do not provide a diagnosis; rather, they typically augment the decision-making process of medical professionals. While some of these technologies can suggest a specific diagnosis, they are not intended or used to determine a final diagnosis. Other technologies may highlight information, such as abnormalities in an MRI image, for a medical professional to evaluate more closely.

We identified three key potential benefits of ML diagnostic technologies:

- **Early detection.** Some ML diagnostic technologies can detect certain diseases earlier in their progression than conventional methods. These technologies can accomplish this by identifying features before a medical

professional would be able to or by enabling the medical professional to screen more patients. Earlier detection of diseases may improve treatment plans and patient outcomes. According to NIH officials, ML technologies can identify features that medical professionals may not detect because they have higher sensitivity and specificity, which allows the technologies to better recognize patterns in data. The officials also stated that ML technologies would allow medical professionals to quickly screen patients and reduce referral wait times for high-risk patients because clinics would not need to see as many patients. A private company official provided an illustrative example of such a benefit. The official told us that in one locale where the company deployed its technology for detecting diabetic retinopathy, non-specialists were able to conduct screenings. This meant that patients who exhibited early signs of diabetic retinopathy could see a specialist quickly, sometimes on the same day, instead of potentially waiting weeks or months to see a specialist. Early detection and treatment of this condition can prevent vision loss and blindness in patients with diabetes.

- **Consistency.** ML technologies can also provide medical professionals with a more consistent analysis of patient data and can help them better diagnose a variety of diseases. For example, ML technologies that analyze medical images can provide consistent results, whereas human specialists, such as radiologists, may misinterpret images due to fatigue, among other possible factors. Similarly, ML

technologies that analyze mammograms can reduce the high level of variation among human interpretations. This helps ensure that more patients receive high-quality screening recommendations, according to an expert meeting participant specializing in breast cancer detection, diagnosis, and treatment. Finally, ML technologies can also help medical professionals track disease progression consistently over time. For example, according to a company official, one available ML technology improves upon conventional approaches for diagnosing Alzheimer’s disease by taking consistent measurements of a patient’s brain images over time. This helps medical professionals track how the patient’s brain is degenerating and compares the degeneration to that of healthy individuals.

- **Access.** Interviewees said that ML technologies could enable more patients to access care, particularly in underserved areas. NIH officials stated that ML diagnostic technologies could allow medical professionals to reach larger segments of the population in at-home care or smaller clinical settings, particularly in areas of the country with limited resources. These technologies can automate certain tasks, which in turn reduces the workloads of some medical professionals and empowers non-specialists to perform specialist tasks, such as cardiac imaging and analysis. For example, in addition to reducing wait times, the diabetic retinopathy screening technology noted above allows more

patients to receive timely care by medical professionals and specialists.

2.2 ML diagnostic technologies by disease

ML technologies can analyze a variety of medical data, but most technologies rely on medical images, according to agency officials.²⁵ According to our expert meeting participants and interviewees, many technologies are designed to use radiology image data because such images are typically standardized and digitized. The majority of such technologies for imaging have been for cancer, but applications for cardiovascular and neurological imaging are becoming more numerous, NIH officials told us. Other types of medical data are less available for training ML technologies due to additional steps including collecting and digitizing data. For example, one cancer researcher noted that data from tissue samples are more difficult to acquire because they require pathologists to first prepare microscope slides and then scan and digitize the images. As a result, some ML technologies to analyze pathology specimens are available but not as mature as ML-based medical imaging technologies.

We identified available ML diagnostic technologies for the diseases we examined, as shown in table 1 and detailed below.

²⁵While some developing ML technologies examine text-based electronic health records for diagnostic purposes, such technologies are not yet widely available for clinical use. In September 2021, FDA published an initial list of AI/ML-enabled

medical devices marketed in the U.S. as a resource to the public about these devices and FDA’s work in this area. Not all devices listed are specifically medical diagnostic technologies. (See the full list [here](#).)

Table 1: Types of data used by available ML diagnostic technologies for five selected diseases

Disease	Data used by available ML technologies
Cancer	Imaging (e.g., magnetic resonance imaging (MRI), computed tomography (CT), X-ray)
Diabetic retinopathy	Imaging (e.g., retinal photos)
Alzheimer’s disease	Imaging (e.g., MRI)
Heart disease	Electrocardiogram (ECG), heart sounds, imaging (e.g., ultrasound)
COVID-19	Biomarker analysis (e.g., immunoassay)

Source: GAO analysis of literature and agency documentation. | GAO-22-104629

Note: These are examples of types of data used, not an exhaustive list.

- Cancer.** Available ML technologies for cancer diagnosis use data from images—collected using MRI, CT, pathology slide microscopy, and X-rays—to help specialists detect, measure, and analyze tumors. One company’s ML technology analyzes breast MRI data and provide radiologists with information such as the densities and sizes of lesions. A company official told us radiologists can use this information to follow up on suspicious features or determine whether a lesion is cancerous. ML technologies can also be used to track the progression of certain cancers over time, which can help medical professionals better assess treatments, according to interviewees

and expert meeting participants. However, the ability to validate ML technologies for diagnosing cancer varies by the type of cancer. For example, an official at a VA medical clinic told us that it is easier to validate image-based ML technologies for diagnosing lung and breast tumors than prostate cancer because lung and breast tumors are more well-defined.

- Diabetic retinopathy.** Available ML technologies can detect signs of diabetic retinopathy by interpreting retinal images captured by a specialty camera. The technologies also recommend a diagnosis to medical professionals. As noted above, these technologies allow medical professionals to screen patients efficiently and consistently to detect the disease earlier than conventional methods, which may better inform treatment and improve patient outcomes. One company’s website states that the technology can return a result in less than a minute. Further, a research paper published by individuals from this company noted that this technology can be scaled up more effectively than manual screening to help meet the needs of a growing population with diabetes.²⁶
- Alzheimer’s disease.** Available ML technologies augment a clinician’s process for diagnosing Alzheimer’s disease by analyzing brain images. These analyses, based on MRI, are intended to help clinicians distinguish changes to brain structure resulting from normal

²⁶Bhaskaranand, Malavika et al, “The Value of Automated Diabetic Retinopathy Screening with the EyeArt System: A Study of More Than 100,000 Consecutive Encounters from People with Diabetes,” *Diabetes Technology and Therapeutics*, vol. 21, no. 11 (2019): 635.

aging and those resulting from Alzheimer's disease. For example, one company developed an ML technology that automatically labels and measures brain structures from a set of MRI scans, but it does not suggest a diagnosis. Other interviewees stated that it can be difficult to validate technologies to detect and diagnose Alzheimer's disease, in part because of the disease's ambiguous clinical definition and diagnostic criteria. An industry coalition official explained that some technologies focus on alerting clinicians about potential features to monitor or analyze closely, such as a plaque in the brain, but these features may not always be a sign of the disease. Similarly, an official from a VA medical clinic stated that, in many cases, clinicians disagree on the features, biomarkers, and definitions for diagnosing Alzheimer's disease.

- **Heart disease.** Available technologies include devices, sold directly to consumers, which track an individual's electrocardiogram (ECG) to detect conditions such as atrial fibrillation.²⁷ For example, individuals can collect and track their ECGs using wearables or other smartphone-enabled devices. We identified three technology companies that have developed wearables to monitor ECGs. In addition, one smartphone-enabled technology records an ECG, analyzes it using an ML algorithm,

and detects several heart conditions, such as atrial fibrillation, bradycardia (slow heart rate), and tachycardia (fast heart rate). These technologies are not intended for consumers to self-diagnose specific medical conditions but rather to help medical professionals better diagnose patients by providing ECG information between visits. In addition to technologies that monitor ECGs, FDA has authorized devices that examine radiological images, score the amount of calcification in blood vessels, segment the amount of plaque buildup within blood vessels, and provide an early alert to radiologists that a patient may have a pulmonary embolism, according to FDA officials.

- **COVID-19.** Technology developers are marketing ML technologies to help improve COVID-19 detection methods. For example, one company created an ML technology that is a non-diagnostic screening device to screen asymptomatic people who may have active COVID-19 infections by assessing the pulse characteristics within a patient's arm.²⁸ According to company officials, their technology is advantageous because (1) it is faster than a standard molecular test which may require samples to be shipped to a laboratory for processing and (2) their technology can detect active infection in the early stages of infection when the viral load may not be high

²⁷ According to CDC, atrial fibrillation, often called AFib or AF, is the most common type of treated heart arrhythmia. When a person has AFib, the normal beating in the upper chambers of the heart is irregular and blood doesn't flow as well as it should to the lower chambers.

²⁸ As of September 2022, this device was available under an emergency use authorization from FDA. FDA may temporarily authorize the emergency use of an unapproved medical product, provided certain statutory criteria are met. See 21 U.S.C. § 360bbb-3. For example, it must be reasonable to believe that the product may be effective and that the known and potential benefits of the product outweigh the known and potential risks.

enough to be reliably detected by other tests. Officials stated that use of this technology could help patients quickly determine the need to isolate themselves, helping to reduce the spread of the disease. Another company's technology analyzes biomarkers from laboratory blood samples to identify patients who may have been infected with SARS-CoV-2, the virus that causes COVID-19.²⁹ This technology measures antibodies against the virus in a patient's blood, and company officials stated that the technology can deliver results within minutes using only a drop of blood. A study, conducted by individuals from the company, suggests that the accuracy of this technology was comparable to, or better than, three other tests.³⁰ Additionally, company officials stated that this technology could differentiate between those who recovered from infection with the COVID-19 virus and those who were vaccinated. Such information could help enhance our understanding of protection from infection by variants of the COVID-19 virus.

2.3 Adoption

Although ML diagnostic technologies have potential benefits and are available to assist in diagnosing the diseases we examined, deep learning technologies are generally not widely adopted in medical clinics, according to our expert meeting participants and interviewees. For example, an industry coalition official stated that medical professionals have used some ML-based tools for over 20 years, but deep learning technologies are newer and therefore less commonly available. Additionally, a survey from the American College of Radiology found a modest 30 percent adoption of AI and ML among radiologists.³¹ This survey also found that large practices were more likely to use the technologies than smaller ones.³²

The companies we interviewed reported various levels of technology adoption. For example, one company with an ECG monitoring technology told us that its technology was being used in most major U.S. medical centers, while another company using a technology to detect COVID-19 infection said its technology was only being used in a handful of universities and research institutions. In addition, some developers create ML technologies to use at specific institutions and do not market them commercially, which limits the extent of their dissemination, according to an expert

²⁹As of September 2022, this device was available under an emergency use authorization from FDA.

³⁰Ikegami, Sachie et al. "Target specific serologic analysis of COVID-19 convalescent plasma," *PLOS One*, vol. 16, no. 4 (2021).

³¹Allen, Bibb, et al., "2020 ACR Data Science Institute Artificial Intelligence Survey," *Journal of the American College of Radiology*, vol. 18, no. 8, (2021): 1153 -1159. (Note: This survey may have included AI technologies that may not be diagnostic devices, and not all AI technologies are ML. Also, the sample used for the survey only included members of the American College of Radiology and is not necessarily representative of all members.)

³²Allen, Bibb, et al. "2020 ACR Survey," 1153.

meeting participant. Chapter 4 discusses reasons medical professionals may be reluctant to adopt ML technologies.

Wider adoption could help improve access to these technologies by medical professionals and patients across various health care settings, geographic locations, and demographics. It could also allow broader realization of the potential benefits of these technologies, including earlier detection of disease and improved consistency of diagnoses.

3 Emerging ML Diagnostic Technologies

Academic, government, and private sector researchers are working to expand the capabilities of AI and ML-based medical diagnostic technologies for the five diseases we examined. We also identified three emerging approaches—autonomous, adaptive, and consumer-oriented ML diagnostics—that can be applied to diagnose a variety of diseases. These advances could enhance medical professionals’ capabilities and improve patient treatment but also have certain limitations.

3.1 Emerging improvements to ML diagnostic technologies, by disease

Academic, government, and private sector organizations continue to research improvements to AI and ML technologies that would enhance or expand upon available capabilities for diagnosing selected diseases. We found examples across the five diseases we examined, including the following:

- **Cancer.** NIH is funding projects to improve available approaches to detect lung, prostate, and colon cancer, using medical imaging and other data sources such as biomarkers. In particular, NIH’s National Cancer Institute Cancer Imaging Program funds research—primarily conducted outside NIH—to reduce diagnostic uncertainty and improve early detection of aggressive cancers. Similarly, another group within the National Cancer

Institute told us that they are developing ML algorithms to help radiologists and pathologists identify prostate cancers, score them for aggressiveness, and predict the cancer’s severity.³³

- **Diabetic retinopathy and other diseases.** Researchers and companies are working to adapt existing ML diabetic retinopathy technologies to help diagnose other eye diseases. For example, an official from a company that developed a diabetic retinopathy technology told us that the company is working to apply its algorithm to detect other eye diseases such as macular degeneration and glaucoma. Researchers are also exploring the use of retinal photos to detect diseases elsewhere in the body, including coronary artery, liver, and gallbladder diseases, according to an expert meeting participant.
- **Alzheimer’s disease.** NIH officials told us that they expect future ML technologies to be able to predict an individual’s risk of developing Alzheimer’s disease and identify the disease subtypes. In addition, researchers have published studies using AI that demonstrate analyses of voice recordings to detect cognitive impairments, including Alzheimer’s and other types of dementia.
- **Heart disease.** An ML diagnostic technology in development can interpret ECGs to determine a patient’s ejection fraction, according to a medical expert at

³³NIH officials also told us that the National Cancer Institute’s Cancer Research Data Commons Imaging Data Commons is intended to support and speed development of new imaging-based ML diagnostics.

our meeting whose organization is developing the algorithm.³⁴ The expert further noted that this algorithm can detect conditions from ECGs that medical professionals cannot easily detect.

- **COVID-19.** Researchers are developing ML technologies to analyze chest X-rays for COVID-19 detection, according to an industry coalition official. Some studies report that X-rays may help medical professionals detect the disease when there is a shortage of conventional testing kits. Additionally, a medical researcher from our expert meeting described using ML technologies based on medical imaging to evaluate responses to various treatments for COVID-19 patients.




3.2 Emerging approaches to ML diagnostic technologies across diseases

In addition to the examples above, which expand on available capabilities for our selected diseases, we identified three emerging, cross-cutting ML-based approaches that can be applied to diagnose a variety of diseases: autonomous, adaptive, and consumer-oriented technologies (see figure 3).³⁵ Interviewees and expert meeting participants expect continued development of technologies that use these approaches.

³⁴Ejection fraction is a measurement of the percentage of blood leaving a patient's heart each time it contracts. Medical professionals can use the measurement to help determine if a patient may have certain types of heart failure.

³⁵These approaches are not mutually exclusive and emerging technologies may incorporate one or multiple approaches.

Figure 3: Potential benefits and limitations of emerging approaches to ML diagnostic technologies

	Description	Potential Benefit	Potential Limitation
 <p>Autonomous technologies</p>	<ul style="list-style-type: none"> Technologies that independently interpret images or other patient data to render a diagnosis. 	<ul style="list-style-type: none"> Fast, consistent information at the point of care. Improved clinician capacity and patient access. Earlier and more accurate detection. 	<ul style="list-style-type: none"> Developers may not be able to create and medical professionals may not adopt algorithms that diagnosis certain diseases autonomously.
 <p>Adaptive algorithms</p>	<ul style="list-style-type: none"> Technologies that update their algorithms by incorporating new patient data. 	<ul style="list-style-type: none"> May provide more accurate diagnoses or information by incorporating additional population or individual data. Food and Drug Administration may be able to streamline its regulatory review of adaptive algorithms by reviewing potential changes to an algorithm during the initial review phase, rather than reviewing individual updates to algorithms. This could allow for rapid improvement of algorithms. Could expand or improve features for users. 	<ul style="list-style-type: none"> Changes in the algorithm data may lead to adverse outcomes such as inconsistent or poorer algorithmic performance.
 <p>Consumer-oriented technologies</p>	<ul style="list-style-type: none"> Technologies such as wearables and at-home devices that are marketed to consumers and may assist medical professionals in monitoring a patient's medical conditions. 	<ul style="list-style-type: none"> Can give medical professionals more information about patients to improve diagnosis and treatment. May increase access to care for consumers, particularly in underserved areas, such as rural settings, that lack specialists. 	<ul style="list-style-type: none"> Need further research to understand whether some devices improve patient outcomes. Effectiveness may depend on patient's ability to understand or willingness to accept the health information presented.

Source: GAO analysis of interviews, literature, and documentation. | GAO-22-104629

Note: These approaches are not mutually exclusive and emerging technologies may incorporate one or multiple approaches.

3.2.1 Autonomous ML diagnostic technologies

Autonomous ML diagnostic technologies would interpret images or other patient data to render a diagnosis, in contrast to the approach discussed in Chapter 2, in which medical professionals interpret results from ML technologies alongside other information to diagnose patients. Such technologies could have several benefits. First, they may reduce costs and

provide faster, more consistent information to patients and medical professionals at the point of care and in real time, according to a company official. Second, they may improve clinician capacity and patient access by removing the need for some patients to see specialists. For example, a company official noted that ML technologies may in the future be able to rule out breast cancer, and companies are already applying autonomous ML technologies to detect diabetic

retinopathy.³⁶ Third, such technologies may result in earlier and more accurate detection than traditional diagnostics and thereby improve treatment and patient outcomes.

Some federal agency and company officials we spoke with noted that they expect researchers to develop an increasing number of autonomous ML diagnostic technologies. However, other interviewees cautioned that such technologies may not be widely developed or adopted because diagnostics may always need to work in tandem with human clinicians.

In particular, developers may not be able to create (and medical professionals may not adopt) autonomous algorithms that diagnose certain diseases, especially those with more complex clinical definitions. FDA officials stated that the complexity of the diagnosis process may limit autonomous diagnostics because many diagnostic tests are not binary (i.e., positive or negative) and require multiple steps or pieces of information, and medical professionals need to fill the gap between the tool and the outcome. For example, two interviewees told us it could be difficult for developers to create an autonomous ML technology to diagnose Alzheimer’s disease because there is little consensus among medical professionals about the diagnostic criteria. In addition, one VA official noted that medical professionals prefer ML technologies that inform rather than provide a diagnosis, particularly for

diagnoses that may affect life insurance rates or disability payments.

3.2.2 Adaptive algorithms

Adaptive ML diagnostic technologies update their algorithms by incorporating new patient data. In contrast to adaptive algorithms, available ML diagnostic technologies are typically “locked,” meaning that manufacturers cannot update algorithms without FDA review. In January 2021, FDA released an action plan on how it may review adaptive algorithms moving forward.³⁷ The process outlined in the plan requests that companies describe their plans for updating an algorithm as part of their submissions for premarket approval, so FDA can review how these algorithms may be modified after entering the market.

Developers can choose the characteristics of the population whose data are used to update the algorithm, as well as the frequency of updates. For example, the technologies may incorporate individual patient data to improve performance of an individual device, or they may aggregate patient data to improve all devices that use a given algorithm or that are used for a given subpopulation. In addition, the frequency of updates can be either continuous or periodic (see figure 4). Continuous updates change the algorithm as new data arrive. With periodic updates, data are collected for some period of time, followed by a discrete update. This makes it easier for developers to confirm that

³⁶ However, available technologies for detecting diabetic retinopathy still involve oversight by medical professionals.

³⁷ Food and Drug Administration, Artificial Intelligence/ Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan (January 12, 2021). See action plan [here](#).

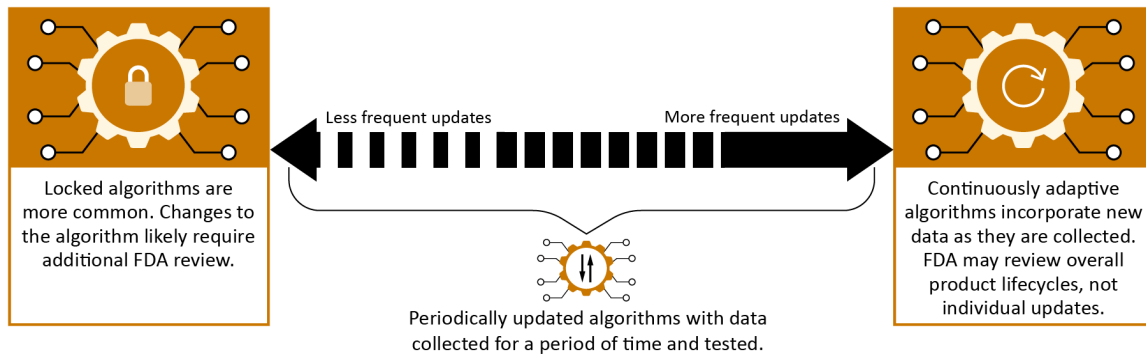
performance has indeed improved after the update, according to two company officials.

Adaptive ML diagnostic technologies may provide more accurate diagnoses or information by incorporating additional population or individual data. According to a participant at our expert meeting, if the technology learns from additional patient data, it may be able to develop a better diagnosis or information that assists in a diagnosis. Also, FDA may be able to limit its regulatory review of adaptive algorithms by reviewing the total product lifecycle, rather than individual updates to algorithms within technologies.³⁸ This could allow for rapid improvement of algorithms while maintaining safety and effectiveness. Additionally, in contrast to the “locked” approach, adaptive ML diagnostic technologies could expand features for certain users. For example, an expert

meeting participant noted that because ML diagnostic technologies may vary in performance from location to location, adaptive technologies could facilitate “tuning” of the algorithm to improve local performance.

However, changes in the algorithm data may lead to adverse outcomes. For example, automatic incorporation of low-quality data may lead to inconsistent or poorer algorithmic performance. According to one industry official, it is important for developers to ensure that changes to a technology’s algorithm in the field do not negatively impact patient outcomes. An official at a different company said periodic updating is the most feasible way to update technologies because it allows for testing and validation before implementing changes.

Figure 4: Spectrum of adaptive algorithms



Source: GAO analysis of Food and Drug Administration (FDA) information. | GAO-22-104629

³⁸For high-risk devices that require FDA premarket review and approval, device sponsors are generally required to submit an application supplement to FDA before making a change affecting the safety or effectiveness of the approved device. 21 C.F.R. § 814.39(a) (2021). Similarly, sponsors of

moderate and low-risk devices that have been cleared by FDA must submit a premarket notification to the agency before making a change to the device that could significantly affect its safety or effectiveness. 21 C.F.R. § 807.81(a)(3) (2021).

3.2.3 Consumer-oriented ML technologies

Available ML diagnostic technologies are typically used in clinical settings such as hospitals and doctor's offices, but certain technologies may be used in homes and other settings by consumers. This approach could help clinicians better monitor patients. For example, some ML-enabled monitors and wearable devices already collect patient data, such as ECGs, at home or throughout the day. Developers continue to advance sensors and wearables for diagnosis and monitoring of different conditions, such as coronary heart diseases and sleep disorders, according to a researcher at our expert meeting. Further, a number of consumer electronics companies have developed and marketed wearables with health monitoring features. According to one company official, consumers and medical professionals are increasingly confident in using the results from these technologies to understand patient conditions. Also, NIH officials projected that more wearable technologies will enter the market in the future.

By deploying ML technologies outside of clinical settings, medical professionals can collect more information about a patient between visits, which may lead to a more accurate diagnosis and treatment plan.³⁹ According to one company official, such data provide patients and physicians with new opportunities for personalized medicine, which involves tailoring information and outputs for specific patients. Deploying these technologies outside of clinical settings may also increase access to care for consumers, particularly in underserved areas that lack specialists, such as rural areas. For example, technologies for diagnosing diabetic retinopathy can be used in optometry chains or pharmacies, which may be easier and more convenient for consumers to access.

However, two expert meeting participants cautioned that additional research is needed for some wearables to understand whether they improve patient outcomes. For example, one expert noted that some companies overstate the abilities of their wearable technologies. Additionally, according to another expert meeting participant, a patient's ability to use and understand health information from wearables may contribute to the wearable's effectiveness, and some general wellness applications have not improved health

³⁹The use of these data may be limited by existing privacy regulations. For example, the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and its implementing regulations, the Privacy and Security Rules, protect individually identifiable health information that is used within the patient and provider relationship. However, the HIPAA protections do not apply to technologies that use or disclose health data outside of this relationship and for which the technology developer is not creating, receiving, transmitting, or maintaining protected health information on behalf of a covered entity or another business associate.

outcomes in the past. Further, while some of these wearables are reviewed by FDA before marketing, others that are not considered medical devices do not fall

under FDA’s jurisdiction.⁴⁰ Without FDA review, such devices may not be independently evaluated for safety and effectiveness before being marketed.

⁴⁰FDA generally only regulates wearables that meet the definition of medical devices, including medical devices that are intended for use by consumers who are not medical

professionals. See 21 U.S.C. §§ 321(h) (defining the term “device”) and 360j(o) (describing software functions that are excluded from the definition of device).

4 Challenges Affecting ML Technologies for Medical Diagnostics

Drawing on information from experts, stakeholders, and the scientific literature, we identified several challenges affecting the development and adoption of ML in medical diagnostics. These challenges affect technology developers, medical providers, and patients and may slow the adoption of these technologies. We highlight the following three challenges below: demonstrating real-world performance, meeting medical needs, and addressing regulatory gaps.

4.1 Demonstrating real-world performance

Medical providers may be reluctant to adopt ML technologies until its real-world performance has been adequately demonstrated in relevant and diverse clinical settings, according to experts, stakeholders, and literature.⁴¹ Before deciding to adopt a technology, medical providers want to know that it is appropriate for their patients and will improve outcomes. According to a review article of AI technologies, it is important to conduct rigorous studies, publish the results in peer-reviewed journals, and establish clinical validation in real-world environments before roll-out and implementation of a

technology in patient care.⁴² In order to establish clinical validity, ML technologies can be trained using high-quality data that are representative of the intended patient population, then tested and validated on diverse external datasets representing a range of clinical settings, conditions, and patient populations. This can help identify biases and limitations of the technology and ensure that results are generalizable.

However, many available technologies have not been adequately tested or validated across generalizable data sets and settings and, as a result, may not transfer from development to adoption in clinical environments. A review of 516 studies that evaluated the performance of image-based AI algorithms found only 6 percent of the studies performed external validation against data sets from institutions or time periods that differed from the training data.⁴³

Further, among ML technologies that have been validated externally, performance can vary substantially. Participants in our expert meeting noted that a key challenge for these technologies is that the performance may vary across different settings. For example, as we have previously reported, a technology

⁴¹In this report, the term medical providers may include healthcare systems, such as clinics, hospitals, or medical centers, as well as medical professionals, such as physicians.

⁴²Topol, Eric J. "High-performance medicine: the convergence of human and artificial intelligence". *Nature Medicine* vol. 25 (2019): 44-56.

⁴³Kim, Dong Wook, et al. "Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers." *Korean Journal of Radiology*, vol. 20(3), (2019): 405-410. The literature search in this review limited the search period to year 2018, with the search updated until August 17, 2018. The report notes that the study design features addressed in this study, including external validation, are crucial for validating the real-world clinical performance of AI but would be excessive for proof-of-concept technical feasibility studies, which constituted nearly all of the studies published in the study period.

may work well in a large high-resource health system but may not work as well in smaller, low-resource systems.⁴⁴ Furthermore, a journal article reported that the performance of diagnostic imaging algorithms varies substantially from site to site in the real world, and went on to highlight the need for validation of algorithm performance at each clinical site before installation.⁴⁵

A lack of prospective studies of these technologies may also be hindering adoption. Prospective studies are those where the outcome has not occurred when the study starts and participants are followed over time to track eventual outcomes. Most ML technologies rely on retrospective data for validation studies, but studies based on retrospective data may not show clinical validity or impact and may not accurately reflect real-world conditions. According to a 2019 review of available AI-based diagnostic technologies, there has been little prospective validation of the algorithms reviewed, and stakeholders will not know how well AI can predict key outcomes in the health care setting until there is robust validation in prospective studies with rigorous statistical methodology and analysis.⁴⁶ A National Academy of Sciences workshop proceedings

report on improving cancer diagnosis stressed that for clinical validation, algorithms should be evaluated in well-designed prospective studies.⁴⁷ Further, two expert meeting participants identified a lack of sufficient prospective evaluation in relevant clinical settings as a key challenge for these technologies.

However, developers face several challenges evaluating and validating ML diagnostic technologies. First, developers have difficulty accessing high-quality representative data to train and validate their technologies. According to an industry developer, access to sufficient amounts of nonbiased, ethnically diverse, real-world training data is their primary challenge, in part because partnering with hospitals and academic centers to obtain data sets takes time, including time to build trust with these institutions. Institutions are often reluctant to share data, especially protected health information, due to privacy concerns.⁴⁸ DOE officials stated that those who have data are reluctant to share them unless the developers can determine how the data will be used. Additionally, officials stated that data use agreements can take a long time to arrange, partly because they typically require approval by an institutional review board to help ensure adequate patient

⁴⁴ GAO-21-7SP

⁴⁵ Larson, David B. et al. "Regulatory Frameworks for Development and Evaluation of Artificial Intelligence–Based Diagnostic Imaging Algorithms: Summary and Recommendations." *Journal of the American College of Radiology*, (October 2022).

⁴⁶ Topol, "High-performance medicine," 49.

⁴⁷ National Academies of Sciences, Engineering, and Medicine. *Improving Cancer Diagnosis and Care: Clinical Application of Computational Methods in Precision Oncology: Proceedings of a Workshop*. The National Academies Press (Washington, D.C.: 2019).

⁴⁸ The HIPAA Privacy Rule generally prohibits the use or disclosure of protected health information except in the circumstances set out in the regulations. Protected health information is individually identifiable health information and includes information collected from an individual, including demographic information, that 1) is created or received by a health care provider, health plan, or health care clearinghouse, and 2) relates to the past, present or future physical or mental health condition of the individual, or the payment for the provision of health care, and 3) identifies the individual or with respect to which there is a reasonable basis to believe the information can be used to identify the individual.

privacy. In addition to concerns over privacy, institutions may be reluctant to share valuable proprietary data if doing so could hurt their competitive advantage.

Performing and funding evaluations is also time and cost intensive and may not be in the best interest of developers, according to literature and a participant in our expert meeting. Rigorous evaluations are expensive and could delay the adoption of some technologies. In addition, a journal article stated that manufacturers of these technologies have a strong financial interest in showing their products in a positive light, and further noted that there is an inherent conflict of interest if they are expected to fund, conduct, and publish results of objective and rigorous evaluations that may highlight deficiencies in their products.⁴⁹ Similarly, one expert meeting participant also expressed concern about the incentive structure for post-market validation, stating that it may not be in a developer's interest to bear the costs of such an evaluation if it could show that the technology does not work.

4.2 Meeting medical needs

Medical providers are less likely to adopt ML technologies that do not address a clear clinical need, and many ML diagnostic technologies do not progress from development to adoption for this reason. Expert meeting participants told us that developers may not understand the clinical

needs of medical providers and professionals. These technologies bring the most value in settings where clinical uncertainty is high or knowledge is quickly changing, where a need exists to reduce specialist referrals or other types of diagnostic tests, or where the technology can demonstrate significant clinical productivity gains, according to a white paper by the Duke Margolis Center for Health Policy.⁵⁰

Developers may struggle to adequately define and communicate the uses and benefits of their technologies. For example, according to an NIH workshop report, the most impactful challenge to the adoption of these technologies is that clinically effective uses for AI have been poorly defined.⁵¹ Similarly, an expert meeting participant stated that providers may not understand the value of these technologies and will not change their practices to adopt them until developers can show the value of their products to help with accuracy, increase efficiencies, or reduce likelihood of error.

Additionally, providers and professionals are more likely to adopt technologies that integrate into existing health care systems and clinical workflows, according to studies and interviewees. For example, VA officials told us that integration with existing technology and security tools and systems is an important consideration when evaluating whether to adopt a technology. These technologies may also not scale commercially

⁴⁹Larson, et al, "Regulatory Frameworks for Development and Evaluation of Artificial Intelligence–Based Diagnostic Imaging Algorithms: Summary and Recommendations", 5.

⁵⁰Duke Margolis Center for Health Policy, *Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care*, (Durham, N.C.: 2019).

⁵¹Bibb, Allen Jr. et al, "A Road Map for Translational Research on Artificial Intelligence in Medical Imaging: From the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop", *Journal of the American College of Radiology*, vol. 16, issue 9, part A (Sept 2019): 1179-1189.

if they do not integrate into clinical workflows. Clinicians are more likely to adopt technologies that integrate into their workflows without adding time or effort to their workload; for example, they may prefer technologies that do not require the clinician to log into separate systems or repeat tasks or thought processes, as this adds extra time and effort for the clinician. Technologies that are integrated into the optimal point in the clinical workflow can reduce burden and maximize effectiveness. The Duke Margolis Center for Health Policy white paper found that the appropriate fit in the workflow will vary based on the technology and application; for example, some technologies may work best if they proactively alert the clinician while others may work best if only activated at the request of the clinician.⁵²

Lastly, providers and professionals are also more likely to adopt technologies that they can understand, according to studies and experts. In particular, they may want to understand how a technology works, its performance, clinical evidence, and potential limitations or biases. In addition, users are particularly likely to adopt an ML technology if they can verify its findings or recommendations, according to a VA medical center official.

However, certain information —such as how the technology works — may be confidential, unknown, or unexplainable. Developers may

consider certain information proprietary and important to their competitive edge. Further, as we previously reported, the decision-making of AI algorithms can be difficult or impossible to explain or understand, even for their developers.⁵³ Though some users may desire full explainability and transparency of ML technologies, achieving this goal may be unrealistic.⁵⁴ According to one study, it may not be possible to explain how the technology arrived at an individual result or decision; rather, the authors recommend thorough and rigorous validation across diverse and distinct populations to show that patient and health care outcomes are improved and that marginalized groups are not disadvantaged.⁵⁵ This approach may be more consistent with the adoption of non-ML technologies; according to participants from our expert meeting, medical professionals routinely use technologies without understanding how the technologies work if their performance has been adequately demonstrated.

4.3 Addressing regulatory gaps

Gaps in the regulatory framework may also pose a challenge to the development and adoption of ML technologies. Regulatory requirements and standards for demonstrating real world performance and clinical validity are insufficient for wide clinical adoption, according to experts, stakeholders, and the research literature. As previously mentioned, medical providers may

⁵²Duke Margolis Center for Health Policy, *Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care*, 24.

⁵³GAO-21-7SP

⁵⁴Explainability refers to methods and techniques in the application of artificial intelligence such that the results of the solution can be understood by humans.

⁵⁵Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care”, *Lancet Digital Health*, vol. 3 (Nov 2021): 745-750.

be reluctant to adopt ML diagnostic technologies without adequate evidence of performance and efficacy across diverse clinical environments. An industry group expressed concern that developers may not understand the importance of clinical validation and that FDA guidance and requirements for clinical validation may not address the needs of clinicians and patients. As discussed earlier in this report, FDA reviews medical devices for safety and effectiveness. However, reviews do not always include comprehensive information on real world performance, clinical outcomes or other information that users may deem relevant to their adoption decisions.⁵⁶ FDA recognizes the need for more evidence of the real world performance of these technologies, according to an action plan FDA released in January 2021.⁵⁷ This plan identifies the need for improved methods to evaluate bias, generalizability, and robustness, as well as the need for clearer guidance on real world performance monitoring.

The existing regulations may also limit the development of emerging types of ML

diagnostic technologies. For example, industry officials stated that regulators would need to change the regulatory environment, standards, and expectations in order to support the development of autonomous technologies. Regulatory gaps may also impact the development of adaptive algorithms. Changes or modifications to a device may require additional review and authorization by FDA, which may limit their ability to improve by learning from real-world use and experience. FDA is working to update regulatory guidance; in 2019, FDA issued a discussion paper on a proposed regulatory framework that includes a “Predetermined Change Control Plan” that could allow devices to learn and iteratively improve after they are in use.⁵⁸ FDA collected stakeholder input on this plan and set a goal to publish draft guidance in 2021; however, as of March 2022, the draft guidance had not been published and, as we have previously reported, draft guidance is issued for comment purposes only and is not for implementation.⁵⁹

⁵⁶According to FDA officials, reviews are limited to utilizing the least burdensome amount of information required to meet the particular regulatory standard, set forth in the Federal Food, Drug, and Cosmetic Act (e.g., substantial equivalence), and FDA is not able to request information that would not aid its authorization decision. This information may not be the same as what would lead to user adoption of a product.

⁵⁷Food and Drug Administration, Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan (January 12, 2021).

⁵⁸FDA, “Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD),” Washington, D.C.: Apr. 2, 2019.

⁵⁹Guidance includes recommendations; stakeholders may use an alternative approach if it satisfies the requirements of the applicable statutes and regulations.

5 Policy Options to Enhance Benefits or Address Challenges of ML Diagnostic Technologies

We developed three policy options that policymakers—Congress, federal agencies, state and local governments, academic research institutions, and industry, among others—could take to enhance the benefits of ML technologies or to mitigate the challenges discussed in the previous chapter. These options include encouraging evaluation of these technologies, improving high-quality data access, and promoting collaboration across stakeholders. We present potential opportunities and considerations for each option.

While we present options to address the major challenges we identified, the list of options is not intended to be exhaustive. We intend our policy options to provide policymakers with a broader base of information for decision-making. We also did not rank the options in any way. Additionally, depending on the options selected, additional steps might need to be taken on potential design and legal issues. We did not conduct work to assess how effective the options may be, and express no view regarding the extent to which legal changes would be needed to implement them.

5.1 Policy Option: Evaluation

Policymakers could create incentives, guidance, or policies to encourage or require the evaluation of ML diagnostic technologies across a range of deployment conditions and demographics representative of the intended use.

This policy option could help address the challenges of demonstrating real world performance.

Description:

- Policymakers could encourage evaluations by providing funding or other incentives for more rigorous evaluations, according to participants in our expert meeting. Policymakers could also create guidance, standards or best practices for evaluation of these technologies. Policymakers could also require post-adoption evaluation under certain conditions. For example, a participant in our expert meeting said regulators might consider formal processes for ongoing review of the accuracy of the technology after adoption, especially when the algorithm is adapting to different institutions or patient demographics.

Opportunities:

- More comprehensive evaluation could help developers, providers, and policymakers better understand the performance of ML technologies across a diverse spectrum of patients, providers,

and other factors. Evaluating technologies through rigorous studies, such as through external validation or peer-reviewed prospective studies, can help stakeholders determine a technology's clinical validity, ability to predict healthcare outcomes, potential biases and limitations, and opportunities for improvement.

- Evaluation could inform providers' adoption decisions. A better understanding of these technologies can potentially lead to increased adoption by enhancing trust, according to FDA officials.
- Information from evaluations can help inform the decisions of policymakers, such as decisions about regulatory requirements.

Considerations:

- Rigorous evaluations can be time-intensive and require collaboration between stakeholders that may already have limited time, such as medical professionals. This could also delay the development and adoption processes, according to VA officials. This could negatively affect the lives of patients and professionals who could benefit from earlier availability.
- More rigorous evaluation will likely lead to extra costs, such as direct costs for funding the studies. As previously mentioned, developers may not be incentivized to conduct these evaluations if it could show their products in a negative light, so policymakers could consider whether evaluations should be conducted or reviewed by independent parties, according to industry officials.

5.2 Policy Option: Data Access

Policymakers could develop or expand access to high-quality medical data to develop and test ML medical diagnostic technologies.

This policy option could help the challenge of demonstrating real world performance.

Description:

- Policymakers can explore opportunities to make data sharing easier, faster, or cheaper. For example, policymakers could reach agreement about data standards or share best practices for collecting and sharing data. Policymakers could also increase data access by, when appropriate, creating and participating in mechanisms for data sharing, such as data commons— cloud-based platforms where users can store, share, access, and interact with data and other digital objects. Policymakers could also use incentives, such as grants or access to databases, to encourage data sharing.

Opportunities:

- Developing or expanding access to high-quality datasets could help facilitate training and testing ML technologies across diverse and representative conditions, which could improve their performance and generalizability. This in turn could help developers and other stakeholders understand the performance of these technologies under varied conditions, identify biases or limitations, and identify opportunities for improvement. According to FDA officials, if clinicians can better understand model

outcomes, it could build trust and adoption in these technologies.

- Expanding access could enable developers to save time in the development process, which could shorten the time it takes for these technologies to be available for adoption.

Considerations:

- As discussed in chapter 4, entities that own data may be reluctant to share them for a number of reasons. For example, these entities may consider their data valuable or proprietary. Some entities may also be concerned about the privacy of their patients and the intended use and security of their data.
- As previously reported, data sharing mechanisms may be of limited use to researchers and developers depending on the quality and interoperability of these data, and curating and storing data could be expensive and may require public and private resources.

5.3 Policy Option: Collaboration

Policymakers could promote collaboration between developers, providers, and regulators in the development and adoption of ML diagnostic technologies.

This policy option could help address the challenges of meeting medical needs and addressing regulatory gaps.

Description:

- Policymakers could promote multidisciplinary collaboration between medical professionals and developers to foster innovation and create technical solutions. Policymakers could convene multidisciplinary experts together in the design and development of these technologies. For example, according to an NIH working group report, policymakers could convene cross-disciplinary collaborators, such as through workshops, conferences, and other opportunities for convening experts from different fields.⁶⁰ Another example of collaboration we previously reported are hackathons, where computer engineers, other technology experts, and providers collaborate to solve technical problems. Regulators could continue to provide public notice and seek public comment from stakeholders to tailor the regulatory framework or create guidance, standards, or best practices for the use and development of ML technologies.

Opportunities:

- Collaboration between ML developers and providers could help ensure that the technologies address clinical needs. For example, collaboration between developers and medical professionals could also help developers create ML technologies that integrate into medical professionals' workflows, and minimize time, effort, and disruption.

⁶⁰Bibb et al, "A Road Map for Translational Research on Artificial Intelligence in Medical Imaging".

- Collaboration among developers and medical providers could help in the creation and access of ML ready data, according to NIH officials.

Considerations:

- As previously reported, providers may not have time to both collaborate with developers and treat patients; however, organizations can provide protected time for employees to engage in innovation activities such as collaboration.⁶¹

We also reported that if developers only collaborate with providers in specific settings, their technologies may not be usable across a range of conditions and settings, such as across different patient types or technology systems.

⁶¹GAO-21-7SP

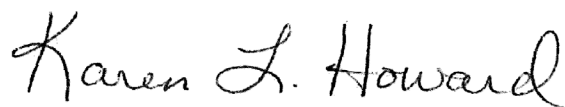
6 Agency and Expert Comments

We provided a draft of this report to the Department of Health and Human Services (Food and Drug Administration and the National Institutes of Health), the Department of Veterans Affairs, the Department of Energy, and the Federal Trade Commission with a request for technical comments, and incorporated agency comments into this report as appropriate.

We also provided a draft of this report to 16 participants from our expert meeting and incorporated comments received as appropriate, consistent with previous technology assessment methodologies.

We are sending copies of this report to the appropriate congressional committees, relevant federal agencies, and other interested parties. In addition, the report is available at no charge on the GAO website at <http://www.gao.gov>.

If you or your staff members have any questions about this report, please contact me at (202) 512-6888 or howardk@gao.gov. Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. GAO staff who made key contributions to this report are listed in appendix III.



Karen L. Howard, PhD
Director
Science, Technology Assessment, and Analytics

List of Requesters

The Honorable Richard Burr

Ranking Member
Committee on Health, Education, Labor, and Pensions
United States Senate

The Honorable Cathy McMorris Rodgers

Republican Leader
Committee on Energy and Commerce
House of Representatives

The Honorable Brett Guthrie

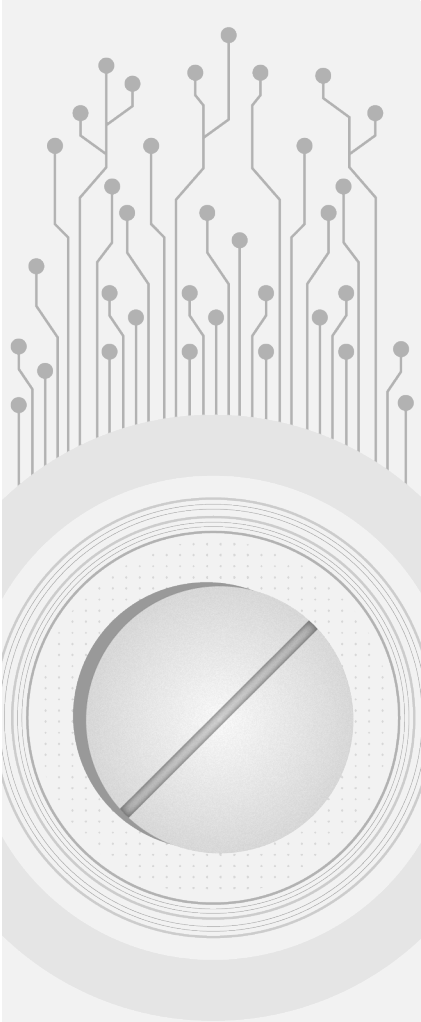
Republican Leader
Subcommittee on Health
Committee on Energy and Commerce
House of Representatives

The Honorable H. Morgan Griffith

Republican Leader
Subcommittee on Oversight and Investigations
Committee on Energy and Commerce
House of Representatives

The Honorable Michael C. Burgess

House of Representatives



PART TWO

Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis

National Academy of Medicine

Part Two presents the NAM publication *Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis* discussing the factors influencing the adoption of non-autonomous point-of-care AI technology that can assist in the diagnosing of a disease. Although GAO and NAM staff consulted with and assisted each other throughout this work, reviews were conducted by GAO and NAM separately and independently, and authorship of the text of Part One and Part Two of the report lies solely with GAO and NAM, respectively.

Part Two—(NAM) Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis

Julia Adler-Milstein, PhD, University of California-San Francisco; **Nakul Aggarwal, BS**, University of Wisconsin-Madison; **Mahnoor Ahmed, MEng**, National Academy of Medicine; **Jessica Castner, PhD, RN-BC**, Castner Incorporated; **Barbara J. Evans, PhD, JD**, University of Florida; **Andrew A. Gonzalez, MD, JD, MPH**, Regenstrief Institute; **Cornelius A. James, MD**, University of Michigan; **Steven Lin, MD**, Stanford University; **Kenneth D. Mandl, MD, MPH**, Boston Children’s Hospital; **Michael E. Matheny, MD, MS, MPH**, Vanderbilt University Medical Center and Veterans Affairs; **Mark P. Sendak, MD, MPP**, Duke University; **Carmel Shachar, JD, MPH**, Harvard University; and **Asia Williams, MPH**, National Academy of Medicine

September 29, 2022

Introduction

Clinical diagnosis is essentially a data curation and analysis activity through which clinicians seek to gather and synthesize enough pieces of information about a patient to determine their condition. The art and science of clinical diagnosis dates to ancient times, with the earliest diagnostic practices relying primarily on clinical observations of a patient’s state, coupled with methods of palpation and auscultation (Berger, 1999; Mandl and Bourgeois, 2017). Following a period of stagnation in clinical diagnostic practices, the 17th through 19th centuries marked a period of discovery that transformed modern clinical diagnostics, with the advent of the microscope, laboratory analytic techniques, and more precise physical examination and imaging tools (e.g., the stethoscope, ophthalmoscope, X-ray, and electrocardiogram) (Walker, 1990). These foundational achievements, among many others, laid the groundwork for modern clinical diagnostics. However, the volume and breadth of data for which clinicians are responsible has exponentially grown,

generating challenges for human cognitive capacity to assimilate.

Computerized diagnostic decision support (DDS) tools emerged to alleviate the burden of data overload, enhance clinicians’ decision making capabilities, and standardize care delivery processes. DDS tools are a subcategory of clinical decision support (CDS) tools, with the distinction that DDS tools focus on diagnostic functions, whereas CDS tools more broadly can offer diagnostic, treatment, and/or prognostic recommendations. Debuting in the 1970s and 1980s, expert-based DDS tools such as MYCIN, Iliad, and Quick Medical Reference operated by encoding then-current knowledge about diseases through a series of codified rules, which rendered a diagnostic recommendation (Miller and Geissbuhler, 2007). While these early DDS tools initially achieved pockets of success, the promise of many of these tools diminished as several shortcomings became evident. Most prominently, the capacity of data collection and the complexity of

knowledge representation prevented accurate representation of the pathophysiological relationships between a disease and treatments. Programmed with a limited set of information and decision rules, several expert-based DDS tools could not generalize to all settings and cases. Some suffered from performance issues as well, often struggling to generate a result or yielding an errant diagnosis. Moreover, users were frustrated. Since these tools existed outside of the main clinical information systems, clinicians had to reenter a long list of information to use them, which created significant friction in their workflows. Similarly, updating the knowledge base of a DDS system often required cumbersome manual entry. Finally, there was a lack of incentives to drive adoption. Thus, provider acceptance remained low, and expert-based DDS tools faded from use (Miller, 1994).

The revitalization of the artificial intelligence (AI) field—the ability of computer algorithms to perform tasks that typically require human intelligence—offers an opportunity to augment human diagnostic capabilities and address the limitations of expert-based DDS tools (Yu, Beam, and Kohane, 2018). Current AI techniques possess not only remarkable processing power, speed, and ability to link and organize large volumes of multimodal data, but also the ability to learn and adjust based on novel inputs, building upon previous knowledge to generate new insights. For this reason, AI approaches, specifically machine learning (ML), are especially well suited to the problems of clinical diagnosis, shortening the time for disease detection, diagnostic accuracy, and reducing medical errors. By doing so, AI

diagnostic decision support (AI-DDS) tools could reduce the cognitive burden on providers, mitigate burnout, and further enhance care quality.

While contemporary AI-DDS tools are more sophisticated than their expert-based predecessors, concerns about their development, interoperability, workflow integration, maintenance, sustainability, and workforce requirements remain, hampering the adoption of AI-DDS tools.

Additionally, the “black box” nature of some AI systems poses liability and reimbursement challenges that can affect provider trust and adoption. This paper examines the key factors related to the successful adoption of AI-DDS tools, organized into four domains: **reason to use, means to use, method to use, and desire to use**. Additionally, the paper discusses the crosscutting issues of bias and equity as they relate to provider trust and adoption of these tools. Addressing biases and inequities perpetuated by AI tools is paramount to preventing the widening of disparities experienced by certain populations and to engendering confidence and trust among clinicians who are responsible for providing care to these populations. To conclude, the authors discuss the policy implications around the adoption of AI-DDS systems and propose action priorities for providers, health systems leaders, legislators, and policy makers to consider as they engage in collaborative efforts to advance the longevity and success of these tools in supporting safe, effective, efficient, and equitable diagnosis.

1 A Primer on AI-Diagnostic Decision Support Tools

AI-DDS tools come in various forms, use myriad AI techniques (see *Table 1*), and can be applied to a growing number of conditions and clinical disciplines. In this paper, the authors focus on adoption factors as they relate to *assistive* AI-DDS tools. Unlike autonomous AI tools, which operate independently from a human, assistive AI

tools involve a human to some degree in the analysis and decision-making process (see *Figure 1*) (Bitterman, Aerts, and Mak, 2020). The authors in this paper focus on AI-DDS tools designed to support health care professionals in decision-making processes, rather than consumer-facing tools in which a layperson interacts with an AI-DDS system.

Table 1: A Non-Exhaustive Glossary of Key Terms Related to Artificial Intelligence






Artificial Intelligence (AI)	A collection of computer algorithms displaying aspects of human-like intelligence for solving specific tasks.
Machine Learning (ML)	A subset of AI that harnesses a family of statistical modeling approaches to automatically learn trends from the input data and improve the prediction of a target state.
Deep Learning (DL)	A subset of ML consisting of multiple computational layers between the input and output that form a “neural network” used for complex feature learning.
Convolutional Neural Networks (CNN)	A subset of DL techniques that is particularly efficient in AI-based pattern recognition. It is the foundation of many image processing AI algorithms, for instance in radiology.
Supervised Learning	A type of AI/ML algorithm that is trained to “learn” associations from labeled data (i.e., input and desired output data).
Unsupervised Learning	A type of AI/ML algorithm that is trained on unlabeled data and intended to “independently” find underlying structures of patterns in input data.
Random Forests Method	A type of ML/AI algorithm involving several decision trees, whose output is the statistical mode (in classification) or mean (in regression) of each of the decision trees.
Natural Language Processing (NLP)	A type of AI that refers to algorithms that employ computational linguistics to understand and organize human speech.
Computer Vision (CV)	Scientific field that deals with how computers process, evaluate, and interpret digital images or videos.
AI Diagnostic Decision Support (AI-DDS)	A computer-based tool, driven by AI algorithms, that uses clinical knowledge and patient-specific health information to inform, aid, and augment health care providers’ diagnostic decision making processes.

Source: Adapted from Abdulkareem, M. and S. E. Petersen. 2021. The Promise of AI in Detection, Diagnosis, and Epidemiology for Combating COVID-19: Beyond the Hype. *Frontiers in Artificial Intelligence*. <https://doi.org/10.3389/frai.2021.652669> and Aggarwal, N., M. Ahmed, S. Basu, J. J. Curtin, B. J. Evans, M. E. Matheny, S. Nundy, M. P. Sendak, C. Shachar, R. U. Shah, and S. Thadaney-Israni. 2020. Advancing Artificial Intelligence in Health Settings Outside the Hospital and Clinic. *NAM Perspectives*. Discussion Paper, National Academy of Medicine, Washington, DC. <https://doi.org/10.31478/202011f>.

Current AI-DDS tools reflect artificial narrow intelligence (ANI), i.e., the application of high-level processing capabilities on a single, predetermined task, as opposed to artificial general intelligence (AGI), which refers to human-level reasoning and problem-solving skills across a broad range of domains. AI-aided diagnostic tools are designed to address specific clinical issues

related to a prescribed range of clinical data. They do not (and are not intended to) comprise omniscient, science-fiction-like algorithmic interfaces that can span all disease contexts. Ultimately, the purpose of AI-DDS tools is to augment provider expertise and patient care rather than dictate it.

Figure 1: Levels of automation of medical artificial intelligence systems

	Assistive AI algorithms		Autonomous AI algorithms		
	Level 1  Data presentation	Level 2  Clinical decision-support	Level 3  Conditional automation	Level 4  High automation	Level 5  Full automation
Event monitoring	AI	AI	AI	AI	AI
Response execution	Clinician	Clinician and AI	AI	AI	AI
Fallback	Not applicable	Clinician	AI, with a backup clinician available at AI request	AI	AI
Domain, system, and population specificity	Low	Low	Low	Low	High
Liability	Clinician	Clinician	Case dependent	AI developer	AI developer
Example	AI analyses mammogram and highlights high-risk regions	AI analyses mammogram and provides risk score that is interpreted by clinician	AI analyses mammogram and makes recommendation for biopsy, with a clinician always available as backup	AI analyses mammogram and makes biopsy recommendation, without a clinician available as backup	Same as level 4, but intended for use in all populations and systems

Source: Bitterman, D. S., H. J. W. L. Aerts, and R. H. Mak. 2020. Approaching autonomy in medical artificial intelligence. *The Lancet Digital Health* 2(9):447-449. [https://doi.org/10.1016/S2589-7500\(20\)30187-4](https://doi.org/10.1016/S2589-7500(20)30187-4). Reprinted with permission under Creative Commons Attribution (CC BY 4.0).

Generally, assistive AI-DDS tools currently use a combination of computer vision and ML techniques such as deep learning, working to identify complex non-linear relationships between features of image, video, audio, *in vitro*, and/or other data types, and anatomical correlates or disease labels. The authors highlight a few representative examples below.

Most prominently, assistive AI-DDS tools can be found in the field of diagnostic imaging, given the highly digital and increasingly computational nature of the field. In fact, radiology boasts more Food and Drug Administration (FDA)-authorized (that is, cleared or approved) AI tools than any other medical specialty (Benjamins et al., 2020). A well-studied algorithm within the cardiac imaging space is HeartFlow FFR_{CT}. Trained on large amounts of computed tomography (CT) scans, this algorithm employs deep learning to create a precise 3D visualization of a patient's heart and major vessels to assist in the detection of arterial blockage (Heartflow, 2014). Deep learning methods can also be applied to gauge minute variations in cardiac features such as ventricle size and cardiac wall thickness to make distinctions between hypertrophic cardiomyopathy and cardiac amyloidosis—two conditions which have similar clinical manifestations and can often be misdiagnosed (Duffy et al., 2022). Within oncology, ML techniques in the form of computer-aided detection systems have been used since the 1990s to support early detection of breast cancer (Fenton et al., 2007; Nakahara et al., 1998). Since then, the FDA has approved several AI-based cancer detection tools to help detect anomalies in breast, lung, and skin images, among others (Shen et al., 2021; Ray and Gupta, 2020; Ardila et al., 2019). Many of these models have been shown to improve

diagnostic accuracy and prediction of cancer development well before onset (Yala et al., 2019).

Beyond imaging, AI applications include the early recognition of sepsis, one of the leading causes of death worldwide. Electronic health record (EHR)-integrated decision tools such as Hospital Corporation of America (HCA) Healthcare's Sepsis Prediction and Optimization Therapy (SPOT) and the Sepsis Early Risk Assessment (SERA) algorithm developed in Singapore draw on a vast repository of structured and unstructured clinical data to identify signs and symptoms of sepsis up to 12–48 hours sooner than traditional methods. In this regard, natural language processing (NLP) of unstructured clinical notes is particularly promising. NLP helps to discern information from a patient's social history, admission notes, and pharmacy notes to supplement findings from blood results, creating a richer picture of a person's risk for sepsis (SPOT, 2018; Goh et al., 2021). However, there are significant concerns about the clinical utility and generalizability of these tools across different geographic settings (Wong et al., 2021).

In the fields of mental health and neuropsychiatry, AI-DDS tools hold potential for combining multimodal data to uncover pathological patterns of psychosocial behavior that may facilitate early diagnosis and intervention. For instance, the FDA recently authorized marketing of an AI-based diagnostic aid for autism spectrum disorder (ASD) developed by Cognoa, Inc. As a departure from deep learning and CNNs, the Cognoa algorithm is based in random forest decision trees. It integrates information from three sources to provide a binary prediction of ASD diagnosis:

1. a brief parent questionnaire regarding child behavior completed via mobile app,
2. key behaviors identified in videos of child behaviors, and
3. a brief clinician questionnaire.

The tool has demonstrated safety and efficacy for ASD diagnosis in children ages 18 months to five years, performing at least as well as conventional autism screening tools (Abbas et al., 2020). There have also been promising demonstrations of AI for diagnosing depression, anxiety, and post-traumatic stress disorder (Lin et al., 2022; Khan et al., 2021; Marmar et al., 2019).

AI-DDS systems are also becoming increasingly common in the field of pathology, particularly *in vitro* AI-DDS tools. Akin to the radiological examples, AI techniques can analyze blood and tissue samples for the presence of diagnostic biomarkers and characterize cell or tissue morphology. For example, a model developed by PreciseDx uses CNNs to calculate the density of Lewy-type synucleinopathy, a biomarker of early Parkinson's disease, in the peripheral nerve tissue of saliva glands (Signaevsky et al., 2022).

2 Facilitating Provider Adoption of AI-Diagnostic Decision Support Tools

Despite the significant potential AI-DDS tools hold in augmenting medical diagnosis, these tools may fail to achieve wide clinical uptake if there is insufficient clinical acceptance. A particularly telling example is that of many early expert-based DDS examples (the forerunners to modern AI-DDS systems, as discussed in the Introduction), which disappointed provider expectations because of a host of usability and performance issues as discussed in the Introduction.

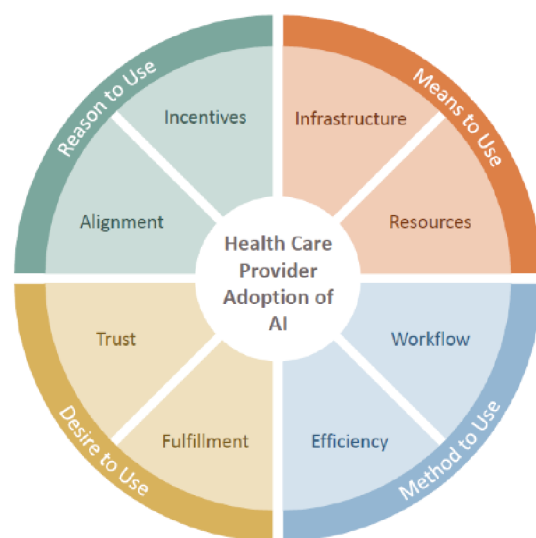
However, the deficiencies of these early AI-DDS tools are instructive for facilitating the adoption of contemporary AI-DDS tools. Additionally, lessons learned from implementing current non-AI-based DDS tools, or systems that generate recommendations by matching patient information to a digital clinical knowledge base, can offer insight. The authors of this paper present a model for understanding the key drivers of clinical adoption of AI-DDS tools by health systems and providers alike, drawing from these historical examples and the current discourse around AI, as well as notable frameworks of human behavior (Ajzen, 1985; Ajzen, 1991). This model focuses on eight major determinants across four interrelated core domains, and the issues covered within each domain are as follows (see *Figure 2*):

- Domain 1: **Reason to use** explores the alignment of incentives, market forces, and reimbursement policies that drive health care investment in AI-DDS.
- Domain 2: **Means to use** reviews the data and human infrastructure

components as well as the requisite technical resources for deploying and maintaining these tools in a clinical environment.

- Domain 3: **Method to use** discusses the workflow considerations and training requirements to support clinicians in using these tools.
- Domain 4: **Desire to use** considers the psychological aspects of provider comfort with AI, such as the extent to which the tools alleviate clinician burnout, provide professional fulfillment, and engender overall trust. This section also examines medicolegal challenges, one of the biggest hurdles to fostering provider trust in and the adoption of AI-DDS.

Figure 2: Core Domains to Support Provider Adoption of AI-DDS Tools



Source: Created by authors

Domain 1: Reason to Use

At the outset, the adoption and scalability of a given AI-DDS tool are driven by two simple but critical factors that dictate the fate of nearly any novel technology being introduced into a health setting. The first factor is the ability of a tool to address a pressing clinical need and improve patient care and outcomes (*alignment* with providers' and health systems' missions). Considering that these tools require sufficient financial investment for deployment and maintenance, the second factor is the tool's affordability both to the patient and health system, including the *incentives* for the provider, patient, and health system to justify the costs of acquiring the tool and investments needed to implement it. The issues related to *Alignment* and *Incentives and Reimbursements* are, in practice, deeply intertwined and codependent. However, for the purposes of the discussion that follows, the authors have separated the two for clarity, emphasizing the logistical and technical steps relevant to *Incentives and Reimbursement*.

Alignment with Health Care Missions

AI-DDS tools must facilitate the goals and core objectives of the health care institution and care providers they serve, although the specific impetus and pathway for AI-DDS tool adoption can vary by organization. For instance, risk prediction and early diagnosis AI-DDS tools being developed and implemented by the Veterans Health Administration (VHA)—the largest integrated health care system in the United States—were initiated by governmental mandates and congressional acts requiring VHA to improve specific patient outcomes in this population (i.e., the Comprehensive

Addiction and Recovery Act) (114th Congress, 2016b). Such initiatives, mandated on a national level, benefit immensely because the VHA is a nationalized health care service, capable of deploying resources in an organized fashion and on a large scale. Another pathway by which these tools can be introduced into clinical settings is through private AI developers collaborating with academic health centers or other independent health systems. These collaborations can result in the creation of novel AI-DDS tools or the customization of “off-the-shelf” commercial tools. A recent example of this type of partnership is Anumana, Inc., a newly founded health technology initiative between Nference (a biomedical start-up company) and Mayo Clinic focused on leveraging AI for early diagnosis of heart conditions based on ECG data (Anumana, 2022). In this context, the AI-DDS development process may be geared toward a given health system's specific needs or strategic missions. However, this does not necessarily preclude its broader utility in other health systems.

A useful framework for evaluating the necessity and utility of AI-DDS tools relates to the Quintuple Aim of health care—better outcomes, better patient experiences, lower costs, better provider experiences, and more equitable care (Matheny et al., 2019). Given the link between patient outcomes and provider experience, it is also important to establish and validate the accuracy of new AI-DDS tools at the start of the adoption process and throughout its use. However, there are often discrepancies between AI-DDS developers' scope and the realities of clinical practice, resulting in tools that can be either inefficient or only tangentially useful. To reassure providers that their tools are optimized for clinical

effectiveness, health system leaders must be committed to regular evaluations of AI-DDS models and performance, as well as efficient communication with developers and companies to update algorithms based on changes like diagnosis prevalence and risk-factor profiles. As algorithms are deployed, and their output is presented to providers in EHR systems, special attention must be paid to the information design and end-user experience to optimize providers' ability to extract key information and act on it efficiently (Tadavarthi et al., 2020). Another critical step in proving robust clinical utility of an AI-DDS tool will be to demonstrate low burden of unintended harms and consequences with use of a given tool (i.e., high sensitivity *and* high specificity) (Unsworth et al., 2022). The degree to which provider reasoning impacts the AI-DDS will also play a role in this regard. Finally, in implementing care plans based in part on AI-DDS output, all care team members must be coordinated in their response and long-term follow-up roles (see **Domain 2: Means to Use** for discussion about requisite resources and roles to accomplish these tasks).

Incentives and Reimbursement

Many health care systems operate on razor-thin financial margins (Kaufman Hall & Associates, 2022). Moving forward, robust insurance reimbursement programs for the purchase and use of AI-DDS tools will be critical to promoting greater adoption by providers and health systems (Chen et al., 2021). However, incentive structures and payer reimbursement protocols for AI-DDS tools are in their nascent stages. Furthermore, insurance dynamics, including for AI-DDS systems, are particularly complex in the U.S., due in part to the heterogeneity of potential payers that

range from governmental entities to private insurers to self-insured employers.

In the current fee-for-service environment, a general trend is for the Centers for Medicare and Medicaid Services (CMS), the federal agency that is the nation's largest health care payer, to be the first to establish payment structures for new technologies and for private payers to then emulate the standards set by CMS (Clemens and Gottlieb, 2017). In determining whether to reimburse the use of a novel AI-DDS tool (and to what extent), a primary consideration for payers, regardless of type, is to assess whether the technology in question pertains to a condition or illness that falls under the coverage benefits of the organization. For instance, an AI-DDS system may be deemed as a complementary or alternative health tool, which may fall outside the scope of many insurance plans and, therefore, be ineligible for reimbursement. If the AI-DDS tool is indeed related to a covered benefit by an insurer [for examples of AI-DDS tools currently reimbursed by US Medicare, see (Parikh and Helmchen, 2022)], developers must provide payers with an adequate evidentiary basis for the utility and safety of the new tool. For this assessment, payers often require data similar to what the FDA would require for premarket approval of a device—for example, clinical trial data showing effectiveness (clinical validity and utility) or other solid evidence that clinical use of the tool improves health care outcomes (Parikh and Helmchen, 2022). Developers bringing new DDS systems to market through FDA's other market authorization pathways, such as 510(k) clearance or de novo classification, may lack such data and need to generate additional evidence of safety and effectiveness to satisfy payers' data requirements (Deverka

and Dreyfus, 2014). Ongoing post-marketing surveillance to verify the clinical safety and effectiveness of new AI-DDS tools thus is important not only to support the FDA's continuing safety oversight but also as a source of data to support payers' evaluation processes.

Experts in health care technology assessment highlight two components of AI-DDS evaluation that are of particular interest to payers: potential algorithm bias and product value. Payers must be convinced that a given AI-DDS will perform accurately and improve outcomes in the specific populations they serve. As described later in this paper, algorithm bias can arise with the use of non-representative clinical data in AI-DDS algorithm development and testing and may lead to suboptimal performance in disparate patient populations based on geographic or socioeconomic factors, as well as in historically marginalized populations (e.g., the elderly and disabled, homeless/displaced populations, and LGBTQ communities). To avoid such biases, monitoring and local validation need to be incorporated into reimbursement frameworks. With regard to product value, payers may weigh the potential clinical benefits of an AI-DDS tool relative to standard diagnostic approaches against the logistical and workflow disruptions that introducing and integrating a new tool into health systems may cause (Tadavarthi et al., 2020; Parikh and Helmchen, 2022). Furthermore, payers can also seek assurance of long-term technical support from algorithm developers.

Although there are not direct reimbursement channels for many types of AI-DDS tools, within the scope of CMS payment systems, there are currently two

primary mechanisms through which AI-DDS services can be reimbursed. The first is that CMS reimburses physician office payments through the Medicare Physician Fee Schedule (MPFS). Within MPFS, payment details are specified via the Current Procedure Terminology (CPT), maintained by the American Medical Association (AMA). CPT codes denote different procedures and services provided in the clinic. New AI-CDS/DDS systems that receive approval for reimbursement by CMS may be assigned a CPT code, as was done in 2020 for IDx-DR, an autonomous AI tool for the diagnosis of diabetic retinopathy (Digital Diagnostics, 2022). The second CMS mechanism is through the Inpatient Prospective Payment System (IPPS) for hospital outpatient services. Within IPPS, the Diagnosis Related Groups (DRG) coding system describes bundles of procedures and services provided to clusters of medically similar patients. Novel AI-DDS tools can be reimbursed in the context of a DRG via a mechanism known as the New Technology Add-on Payment (NTAP). NTAP, created to encourage the adoption of promising new health technologies, provides supplemental payment to a hospital for using a given new technology in the context of a broader care plan that may be covered in the original DRG (Chen et al., 2021).

As AI-DDS systems become more prevalent, sophisticated, and integrated into broader diagnostic workflows, distinguishing their specific role in the diagnostic process and ascribing specific reimbursement values to an algorithm may become difficult. AI-DDS tools may fare better and enjoy greater adoption under value-based payment frameworks, where efficiency and overall quality of care are incentivized rather than individual procedures (Chen et al., 2021).

Domain 2: Means to Use

Paramount to establishing the value proposition is ensuring that clinical environments are properly equipped to support and sustain the implementation of AI-DDS tools. This consists of two interrelated elements: (a) the data and computing *infrastructure* required to collect and clean health care data, develop and validate an AI algorithm at the point of care, and perform routine maintenance and troubleshooting of technical problems in a high-throughput environment; and (b) the human and operational *resources* needed to conduct these technical functions so clinicians can seamlessly interface with these tools.

Infrastructure

Building the necessary infrastructure to deploy AI-DDS relies on developing the hardware and software capabilities to support a range of functions beginning with data processing and curation. Concurrent with developing and implementing a working AI-DDS pipeline, several health IT infrastructure and data flow steps are required to support the implementation and sustainment of an AI-DDS tool. The first point of entry into the pipeline is data ingestion. This step requires linking a data producer, such as an MRI machine, into a data collection and processing workflow to maintain and represent the data in a way that can be leveraged by an AI-DDS algorithm. Many AI-DDS systems currently in use are “locked,” which means that the algorithms are static. However, in the case of a continuous learning/adaptive AI system, in which the system continuously ingests new data to update the algorithm in “real-time,” this could be performed on a fixed schedule (e.g., every day, month, etc.)

or a trigger. The next consideration is determining where and how the raw data is stored (e.g., enterprise data warehouse [EDW] versus a data lake). In practice, these considerations are constrained by, first, the specific clinical problem being addressed and, second, the extent to which the available resources can accommodate the complexity of the pipeline. An EDW, which contains structured, filtered data for specific uses, may be preferred for operational analysis, whereas a data lake house, which is a large repository of raw data for purposes yet to be specified, may be selected by institutions seeking to perform deep research analysis. While model development is a distinct step in building an AI pipeline, it is nonetheless interdependent on deployment considerations. For example, an institution seeking to build analytic tools that are robust to future changes in imaging (e.g., adding a new MRI machine) may opt for a more flexible architecture of a data lake house instead of a traditional EDW. This, in turn, creates dependency cascades since data storage choice changes the order and extent to which data cleaning and other pre-processing pipelines are implemented. Thus, AI-DDS development and implementation choices are both business operations and data science decisions since their steps are codependent.

Some clinical problems may require more frequent data updates or “data meals” to ensure that adaptive AI systems can appropriately address rapidly evolving issues with a nascent foundation of data. For instance, a COVID-19 diagnostic model at the beginning of the pandemic might have been built around admission vital signs and complete blood count (CBC) results. However, as knowledge about the natural

history of the illness progressed, the model may have evolved to include additional data types such as erythrocyte sedimentation rates (ESR), chest X-ray (CXR) images, and metabolic panel data. In many hospital systems, adding the ESR values is not particularly challenging from a data ingestion standpoint because this data originates from the same system that provides the CBC values. However, the addition of CXR images is challenging because it requires working with another department—radiology, in this instance—and interfacing with another information system (picture archiving and communication system [PACS]). Finally, extending predictions from a single outcome at a discrete point in time (i.e., cross-sectional analysis) to multiple predictions or ones relying on time series data can impact upstream choices for data ingestion pipelines.

It is also important to consider that health care AI needs to be deployed in clinical workflows. In these settings, the demand for near real-time data can result in added hardware complexity, expense, and risk. Notably, for most AI-DDS systems, raw data is insufficient; high-quality data that has been curated and annotated is required for robust algorithm training. At a minimum, redundant storage and processing cores capable of model training and validation are essential. While the granular technical requirements are specific to the algorithm employed, the amount and type of data (e.g., images vs. audio vs. text) institutions seek to implement AI-DDS tools may necessitate the ability to access storage on the terabyte and potentially petabyte scale. However, not all data are required to be

available for real-time access. Furthermore, while discussion of data privacy and security is beyond the scope of this section, there are numerous Health Insurance Portability and Accountability Act (HIPAA)-compliant cloud solutions that could address the issues of availability of real-time data access and storage. These issues should be carefully considered in an institution's data plan when seeking to develop and deploy AI-DDS tools.

Another major consideration beyond storage is processing power, particularly for model development and model updating. The types and number of specific chipsets that would be most beneficial should be determined by expert consultation once there is some understanding of the clinical use case and the amount and type of medical data involved. Due to the computational requirements, deep learning-based models might require use of graphical processing units (GPUs), which, in contrast to central processing units (CPUs), offer the ability to do parallel processing with multiple cores, which is particularly useful in deep learning models. While such models could be run on conventional CPUs, efficiency may be reduced by several orders of magnitude depending on model complexity, resulting in models that take weeks to train rather than hours.

Finally, with respect to deployment, it is essential that there is a local solution permitting any mission-critical AI-DDS tools to continue to function at times when internet connectivity is disrupted. Previously, these “downtime” events were often limited to a few hours or days. However, in the age of hospitals becoming an increasing target for ransomware attacks, some planning should be made for

what to do if a downtime event lasts weeks or months.

With respect to software needs, the ability of models to run on mobile devices is becoming increasingly important. As such, the ability to either securely log on to a hospital's server or perform the computations for an AI-DDS on a mobile device is becoming the industry standard, rather than a bespoke one-off requirement for providers enthusiastic about technology. The extent to which health systems should invest in such technology depends on the amount and type of data, the complexity and efficiency of AI/ML models, and the clinical scenario the AI-DDS is addressing. To illustrate, consider an AI-DDS that predicts the need for hospital admission based on data collected from traveling wound care nurse checking capillary blood glucose and uploading a picture of a patient's worsening extremity wound. All of this can now be done on a mobile device. A model could be implemented such that a traveling wound care nurse takes a picture and runs the model at the point of care using an application on a mobile device.

Another key consideration for deployment of AI-DDS tools is system interoperability. This issue can be conceptualized from many different "pain points". One occurs at the data ingestion stage, as discussed previously. This may be due to incompatible EHR systems (e.g., the hospital's inpatient system uses Cerner, but the outpatient clinics use Epic), which cannot "speak" to one another. Alternatively, a health system may have hospitals that use the same EHR, but the EHRs do not share a common data storage repository. Although everyone uses the same PACS system, pulling imaging data from hospitals A, B, and C requires

accessing one server, while pulling data from hospitals X, Y, and Z across the state requires accessing a different server, an issue of interoperability related to information exchange. A second ingestion scenario would require harmonization of different sensors into the same repository. For example, the hospital may use multiple types of point-of-care glucose monitors. The workflow workaround is often that the bedside technician looks at the monitor reading and then types it into the EHR. However, if this data needed to be transitioned into an automatically collected format, there may need to be different integrations for each type of glucose monitor. A second "pain point" occurs in the data cleaning stage, known as the data curation stage. Consider the ramifications of a hospital changing from reporting hemoglobins to hematocrits or traditional troponins to high sensitivity troponins. While this makes little difference at the bedside, it has the potential to significantly complicate AI/ML modeling if the change is not recognized and a standardized process for addressing the inconsistency is not developed. Although a hospital's primary focus should be on selecting tools that enhance value for patients, some attention should be devoted to considering how these tools may impact AI-DDS pipelines. As the reliance on cyber-physical systems grows, health systems should plan to mitigate how physical equipment upgrades change AI/ML data ingestion and use pipelines. Usually, such changes have a trivial effect on overall model performance; however, they can significantly impact the time and effort required to pre-process data. The most efficient way would be to have members of the AI-DDS team with expertise in cyber-physical systems and extract, transform, and load (ETL) data pipelines.

In addition, ensuring providers can readily access AI-DDS tools is critical to adoption. Successfully deploying an AI-DSS tool requires optimizing the multitude of human and software factors involved in the patient care workflow. However, as a preliminary consideration, the essential task is building infrastructure that avoids clinician devising workarounds. There is ample evidence that clinicians will avoid using or develop workarounds for poorly tailored solution or requirements that are perceived as being foisted on them and otherwise constitute yet another inefficiency in an already inefficient system. Regarding software, developers must be prepared to ensure that the tool can be used and viewed on both desktop and mobile devices and potentially by provider-facing and patient-facing versions of the EHR software. Transitioning between these various contexts should be seamless and, more importantly, provide the same information.

Resources

Apart from the data and computational infrastructure necessary to develop, implement, and maintain a health care AI-DDS solution, there are also significant human capital requirements. Practices and health systems often lack the required human resources to run a minimum data infrastructure that can support AI-powered applications. Key requirements include, but are not limited to, frontline IT staff, data architects, and AI-machine learning specialists to understand the context of use and tailor the solution to be fit for purpose. The infrastructure also requires information security and data privacy officers, legal and industrial contract officers for business and data use agreements, and IT educators to train and retrain providers and staff.

To ensure sustainable and safe integration of AI-DDS tools into clinical care, it is crucial that the tools meet the clinical needs of the institution while also maintaining alignment with best practice guidelines, which change over time (Sutton et al., 2020). This requires a governance process in the health care system, with time investments from executive leadership and sponsorship as well as committee and oversight mechanisms to provide regular review (Kawamanto et al., 2018). Direct clinical champions must also have dedicated time to interface between front-line clinicians and the leadership, informatics, and data science teams. These models and tools need to be assessed for accuracy in the local environment and modified and updated if they do not perform as expected. Lastly, they must be surveilled over time and checked regularly to ensure performance maintenance.

One of the major challenges in effectively deploying AI in health care is managing implementation and maintenance costs. Nationally, non-profit hospital systems report an average profit margin of around 6.5%. (North Carolina State Health Plan and Johns Hopkins Bloomberg School of Public Health, 2021). These relatively slim margins encourage health care systems to be conservative in investing in unproved or novel technologies. Robust analysis of cost savings and cost estimates in the deployment of AI in health care is still lagging, with only a small number of articles found in recent systematic reviews, most of which focus on specific cost elements (Wolff et al., 2020). In general, industry estimates the overall cost of development and implementation of such tools can range from \$15,000 to \$1 million, depending on the complexity of the system and integration with workflow (Sanyal, 2021).

Another challenge is the tension between hiring a health care technology firm to develop or adapt the algorithms and tools into a health care environment versus hiring and supporting internal staff, which could cost between \$600 and \$1,550 a day (Luzniak, 2021). Even when much of the core data science expertise is hired into a system, data scientists spend about 45% of their time on data cleaning (GlobeNewswire, 2020). Because familiarity and ongoing business intelligence and clinical operations needs require managing data, many systems choose to hire internally for a portion of their infrastructure needs, which require a continued injection of capital.

Domain 3: Method to Use

Operationalizing and scaling innovations within the health care delivery system is costly and challenging. This is partly due to the heterogeneity of clinical *workflows* across and within organizations, medical specialties, patient populations, and geographic areas. Thus, AI-DDS tools must contend with this heterogeneity by plugging into key process steps that are universally shared. However, a weakness that limits options for reshaping physician *workflows* is the still nascent implementation science for deploying interventions that change provider behavior as well as the non-modularity and non-modifiability of extant, sometimes antiquated point-of-care software, including EHRs (Mandl and Kohane, 2012).

Coupled with workflow challenges is the issue of developing and deploying these tools in a manner that improves *efficiency of practice* and frees up cognitive and emotional space for providers to interact with their patients. The risk of unsuccessful

systems interfering with or detracting from the diagnostic process, through user interface distractions or data obfuscation, exists and must be guarded against. In addition, extensive user training, both onboarding and ongoing and equally nimble educational infrastructure, is necessary to ensure technical proficiency.

Workflow

AI-DDS tools must be effectively integrated into clinical workflows to impact patient care. Unfortunately, many integrations of AI solutions into clinical care fail to improve outcomes because context-specific factors limit efficacy when tools are diffused across sites. Although numerous details are crucial to integrating AI/ML tools into practice, three key insights have emerged from experiences integrating AI/ML tools into practice at various locations and drawn from literature reviews of the AI clinical care translation process (Kellogg et al., 2022; Sendak et al., 2020a; Yang et al., 2020; He et al., 2019; Wiens et al., 2019; Kawamoto, 2005).

First, health systems looking to use AI-DDS tools must recognize the factors that shape adoption and be willing to restructure roles and responsibilities to allow these tools to function optimally. The current state of health information technology centers workflows around the EHR, and AI tools often automate tasks that historically required manual data entry or review. Similarly, AI tools often codify clinical expertise and can prompt concern from clinicians who value autonomy (Sandhu et al., 2020). To navigate these complexities, health systems may need to develop new workflows that change clinical roles and responsibilities, including new ways for interdisciplinary teams to respond to AI

alerts. For example, an increasing number of AI tools require staff in a remote, centralized setting to support bedside clinical teams (Escobar et al., 2020; Sendak et al., 2020b). Many hospitals already benefit from more manual remote, interdisciplinary support through services such as cardiac telemetry, eICU, and overnight teleradiology. Similarly, AI can decentralize the location of specialized services. For example, instead of diabetic retinopathy screening requiring a visit to a retina specialist, Digital Diagnostics now hosts automated AI machines at grocery stores (Digital Diagnostics, 2019).

Second, health systems must closely examine the unique impacts of AI integration on different stakeholders along the care continuum and balance stakeholder interests. This is a key facet in establishing the value proposition for the introduction of a new AI-DDS tool. Experience in AI integration reveals that “predictive AI tools often deliver the lion’s share of benefits to the organization, not to the end user” (Kellogg et al., 2022). Predictive AI tools often identify events before they happen, meaning the optimal setting for AI use is upstream of the setting typically affected by the event. For example, patients with sepsis die in the hospital and often in intensive care units, but timely intervention to prevent complications must occur within the emergency department (ED). Similarly, patients with end-stage renal disease often present to the ED to initiate dialysis, but preventive interventions must occur in primary care. Project leaders looking to integrate AI into workflows must map out value streams, and if value is captured by downstream stakeholders in a different setting, project leaders must identify other opportunities to create value for end users.

One approach is to identify “how a tool can help the intended end users fix problems they face in their day-to-day work” (Kellogg et al., 2022). For example, when a team of cardiologists and vascular surgeons aimed to reduce unnecessary hospital admissions for patients with low-risk pulmonary embolisms (PEs), ED clinicians initially pushed back. Scheduling outpatient follow-up for a low-risk PE had historically been challenging, so the specialists offered to coordinate care for patients identified by the AI/ML tool and block off outpatient appointments to ensure timely follow-up, allowing both the tool and the clinicians to operate as efficiently as possible (Vinson et al., 2022).

Third, workflows should be continuously monitored and adapted to respond to optimize the labor effort required to effectively use AI tools. For example, when a chronic kidney disease algorithm was implemented on a Duke Health Medicare population of over 50,000 patients, many patients identified by the algorithm as high risk for dialysis were already on dialysis or seeing a nephrologist outside of Duke (Sendak et al., 2017). Early intervention was no longer as relevant for these patients, so the team agreed to establish a new pre-rounding process by which a nurse filtered out patients already impacted by the outcome of interest. However, after months of manually reviewing alerts for patients identified by an AI tool as high risk of inpatient mortality, the lead nurse felt confident that the algorithm identified appropriate patients (Braier et al., 2020). The team agreed to remove the manual review step and directly automate emails to hospitalist attendings to consider goals of care conversations. Lastly, there must also be feedback loops with end users to ensure that the AI tool continues to be

appropriately used. For example, hospitalists using the inpatient mortality tool inquired about using the tool to triage patients to intensive care units. Similarly, nurses responding to sepsis alerts began asynchronously messaging clinicians in the ED through the EHR rather than calling and talking directly with provider. These changes in communication approach and intended use may seem subtle but can undermine validity of the tool and potentially harm patients. To avoid drift in workflow or use of AI tools, project leaders should clearly document algorithms and regularly train staff on appropriate use (Sendak et al., 2020c).

Efficiency of Practice

The impact of AI-DDS tools and systems on the cognitive and clerical burdens of health care providers remains unclear. Successful tools would ideally reduce both burdens by delivering just-in-time diagnostic assistance in the most unobtrusive manner to providers while minimizing clerical tasks that might be generated by their use (e.g., extra clicks, menu navigation, more documentation). Experience with traditional CDS systems has shown that these tools are significantly more likely to be used if they are integrated into EHRs instead of existing as stand-alone systems. However, integration alone is insufficient. How that integration is executed—from the design of the user interfaces to the way alerts and notifications are displayed (e.g., triggers, cadence) or handled (e.g., non-interruptive versus interruptive alert)—is critical to practice efficiency and, ultimately, provider acceptance and adoption.

One major impediment is the high degree of difficulty integrating new software with vendor EHR products. Most integrations are

“one-offs,” and, therefore, the technology fails to diffuse broadly. The 21st Century Cures Act (“Cures Act”) specifies a new form of health IT interoperability underpinning the redesign of provider-facing applications as modular components that can be launched within the context of the EHR, and which may be instrumental in delivering AI capabilities to the point of care (114th Congress, 2016a). The Cures Act and the federal rule that implements interoperability provisions require that EHRs have an application programming interface (API) granting access to patient records “with no special effort” (Wu et al., 2021; HHS, 2020). “APIs are how modern computer systems talk to each other in standardized, predictable ways. The Substitutable Medical Applications, Reusable Technologies (SMART) on Fast Healthcare Interoperability Resource (FHIR) API, required under the rule, enables researchers, clinicians, and patients to connect applications to the health system across EHR platforms” (Wu et al., 2021). Top EHR vendors have all incorporated common API standards (“SMART on FHIR”) into their products, creating a substantial opportunity for innovation in software and data-assisted health care delivery. Illustrative of the transformative potential of the integration of AI-DDS with EHRs is Apple’s decision to use the SMART API to connect its Health App to EHRs at over 800 health systems, giving 200 million Americans the option to acquire standardized and computable copies of their medical record data on their phones. The implementation science underpinning translation of machine learning to practice is nascent, however. Cultivating support for standards is driving an emerging ecosystem of substitutable apps, which can be added to or deleted from EHRs (like apps on a smartphone

can). Such apps yield opportunity to deliver the output of diagnostic algorithms within the provider workflow during an EHR session within a patient context (Barket and Johnson, 2021; Kensaku et al., 2021; Khalifa et al., 2021).

EHR alert fatigue is a widespread and well-studied phenomenon among providers that has been linked to avoidable medical errors and burnout (Ommaya et al., 2018). How the introduction of AI-DDS systems into next-generation EHRs might affect alert fatigue and the provider experience is unclear. Successful deployment of these AI-DDS tools likely requires use of both human factors engineering and informatics principles, as the problem arises from the technology and how busy humans interact with it. Diagnostic outputs provided by the DDS should be specific, and clinically inconsequential information should be reduced or eliminated. Outputs should be tiered according to severity with any alternative diagnoses presented in a way that signals providers to clinically important data. Alerts must be designed with human factors principles in mind (e.g., format, content, legibility, placement, colors). Only the most important, high-level, or severe alerts should be made interruptive.

While thoughtful human-centered design can facilitate adoption to an extent, some degree of health care provider training will be required to ensure the necessary competencies to use AI-based DDS tools. The rapid pace of technological change requires such educational infrastructure to be equally nimble. Training opportunities must be integrated across undergraduate medical education, graduate medical education, and continuing medical education. To the extent that some AI-DDS tools are designed to support collaborative

team workflows, interprofessional and multidisciplinary training is also necessary. While competencies surrounding the use of AI-DDS systems are still evolving and yet to be established, the authors of this paper have identified the following core areas as essential:

1. Foundational knowledge (“What is this tool?”);
2. Critical appraisal (“Should I use this tool?”);
3. Clinical decision making (“When should I use this tool?”);
4. Technical use (“How should I use this tool?”);
5. Addressing unintended consequences (“What are the side effects of this tool and how should I manage them?”)

For *foundational knowledge*, health care providers need to understand the fundamentals of AI, how AI-DDS are created and evaluated, their critical regulatory and medicolegal issues, and the current and emerging roles of AI in health care. For *critical appraisal*, providers need to be able to evaluate the evidence behind AI-DDS systems and assess their benefits, harms, limitations, and appropriate uses via validated evaluation frameworks for health care AI. For *clinical decision making*, providers need to identify the appropriate indications for and incorporate the outputs of AI-DDS into decision making such that effectiveness, value, and fairness are enhanced. For *technical use*, providers need to perform the tasks critical to operating AI-based DDS in a way that supports efficiency, builds mastery, and preserves or augments patient-provider relationships. To address *unintended consequences*, providers need to anticipate and recognize the potential

adverse effects of AI-DDS systems and take appropriate actions to mitigate or address them. Determining how to integrate this education into an already crowded training space, whether extra certification or credentialing is required for providers to use AI-DDS, and how institutions can adapt to rapidly changing training needs on the frontlines remain open questions.

Domain 4: Desire to Use

Ultimately, the success of AI-DDS tools in optimizing health system performance is dependent on the desire of clinicians to incorporate these tools into routine practice. Indeed, the factors discussed in the previous three core domain sections are crucial variables in the “desire to use” calculus. Additionally, it is important to attend to psychological factors, such as addressing how these tools can facilitate *professional fulfillment* among providers, including mitigating burnout. The other indispensable element within the desire to use core domain is *trust*. Clinicians must be able to trust that these tools can deliver quality care outcomes for their patients without creating harm or error and align with both patients’ and clinicians’ ethics and values.

Professional Fulfillment

Continued alignment of AI technology with the element of the Quintuple Aim to improve the work-life balance of health care professionals remains an indispensable aspect of the potential success and adoption of AI tools. Health care providers report high levels of professional burnout, partially attributable to EHRs and related technologies (Melnick et al., 2020). Generally, for every one hour spent with patients, providers spend another two

hours in front of their computers (Colligan et al., 2016). The exponential rise in digital work since the COVID-19 pandemic began has exacerbated burnout and amplified some providers’ deeply rooted reluctance to adopt new technologies (Lee et al., 2022). Successful AI-DDS tools will need to overcome this hesitancy and tap into positive sources of fulfillment for providers, including facilitating professional pride, autonomy, and security; reassessing or expanding their scope of practice; and augmenting their sense of proficiency and mastery.

A major contributing source of professional fulfillment is the strength of the patient-provider relationship. As discussed, AI-DDS tools hold the potential to greatly improve diagnostic accuracy and reduce medical errors. If seamlessly integrated, they could also unburden providers of rote tasks, enabling them to allocate more attention to engaging and establishing meaningful bonds with patients. However, by deferring certain higher-order data analysis and synthesis tasks—functions traditionally within the scope of providers—to an AI-based system, providers may experience a sense of detachment from their work. There also is concern that AI systems could erode the patient-provider relationship if patients begin to preferentially value the diagnostic recommendation of an AI system. While the personal qualities of interacting with a human might be preferred, some believe that AI’s ability to emulate human conversation (via chatbots or conversational agents) could eventually supplant providers (Goldhahn et al., 2018). However, it should be noted that this concern only applies to autonomous systems, and the assistive systems this paper focuses on by definition involve, by

definition, a health care professional in the workflow.

As observed in previous cycles of AI diffusion, potential threats to job security have negatively impacted provider receptivity to AI. Anxiety has been particularly acute in certain specialties, such as radiology, where in 2016, speculation arose that radiologists would be irrelevant in five years (Hinton, 2016). However, instead of replacing providers, AI in radiology has assumed an assistive role, supporting providers in the sorting, highlighting, and prioritizing key findings that might otherwise be missed (Parakh, 2019). Therefore, to foster the adoption of AI-DDS, it is important to uphold the paradigm of augmented intelligence—in which these tools enhance human cognition, and the human is ultimately the arbiter of the action recommended. A key element of this is to empower providers to co-exist in an increasingly digital world through skill-building and instilling trust and transparency in AI systems. It is also important to reconsider expectations about provider roles and responsibilities. With the potential of increased practice efficiency, AI-DDS tools may expand provider bandwidth and purview. In this regard, providers could see patients in greater numbers, through multiple media, and in geographically distant areas.

Despite increasingly sophisticated AI algorithms, it is imperative to value the human qualities that can correct or counteract the shortcomings of AI systems. For instance, biased algorithms struggle with diagnosing melanoma in darker-skinned patients (Krueger, 2022). Having a provider carefully review and assess results produced or interpreted by an AI tool is essential to avoiding a missed or erroneous

diagnosis in this case. Above all, provider involvement is critical in shared decision making. Even in circumstances when an AI-DDS tool is highly accurate, providers are indispensable in helping patients select the right course of treatment based on their health goals and preferences.

Trust

Trust within human-AI-diagnostic partnerships requires a human willingness to be vulnerable to an AI system. Trust overall is a complex concept and trust in technology is equally complex (Lankton et al., 2015). A human user may distrust an AI-DDS tool whose recommendations go against their intuitive conclusions, especially if that person has professional training and significant experience. A user may also distrust AI-DDS recommendations if the user finds something faulty with the development process of the tool, such as inadequate testing or a lack of process transparency. Another potential impediment can include concern that the tool's development and use is motivated by profits over people or a lack of professional values alignment (Rodin and Madsbjerg, 2021). Clarity in individual clinician and health care organizational governance and standards setting for various AI tools remains unclear, which also may inhibit trust. Drivers of trust, on the other hand, can include positive past experiences with a particular manufacturer or service provider, seamless interoperability of a new application with an existing suite of tools from a familiar and currently trusted company or product, or company reputation among the professional health care community (Adiekum et al., 2018; AI HLEG, 2019; Benjamin, 2021).

In this section of the paper, the authors focus on two significant sources of distrust with AI-DDS products as especially relevant to the adoption of AI-DDS by clinicians:

1. bias (real or perceived) and
2. liability.

Providers may be concerned that AI-DDS tools underperform in care for certain patients, especially marginalized populations, as AI trained on biased data can produce algorithms that reproduce these biases. However, it is critical to recognize that bias has multiple sources. It could arise, for example, if the data used to train the AI did not adequately represent all population subgroups that eventually will rely on the AI-DDS tool. It is crucial to ensure that training data are as inclusive and diverse as the intended patient populations, and that deficiencies in the training data are frankly disclosed. Using all-male training data for a tool intended for use only in males to detect a male health condition would not result in bias, but using all-male data would cause bias in tools intended for more general use. Other bias types could exist, for example, if AI tools are trained using real-world data incorporating systemic deficiencies in past health care. For example, if doctors in the past systematically underdiagnosed kidney disease in Black patients, the AI can “learn” that bias and then underdiagnose kidney disease in future Black patients. Thus, it is crucial to design and monitor AI tools with a lens toward preventing, detecting, and correcting bias and disclosing limitations of the resulting AI-DDS tools.

Complicating this issue is the fact that it can be very difficult to understand the inner workings of many AI-DDS algorithms. The

terms “transparency” and “explainability” can have various technical meanings in different contexts, but this paper conceives them broadly to denote that the user of an AI tool, such as a health care professional, would be able to understand the underlying basis for its recommendations and how it arrived at them. It can be challenging, and at times impossible, to understand how an AI arrives at its output and to determine whether the tool in question problematically replicates social biases in its predictions. Furthermore, developers rarely reveal the underlying data sets used to train AI-DDS algorithms, making it difficult for providers to ascertain if a particular product is trained to reflect their patient populations. There may also be tension between the AI-DDS purchasing decisions made by hospital leadership and the providers affiliated with the institutions, with the perception that hospital leadership is “imposing” use of specific AI-DDS algorithms on the providers.

To foster trust among clinician users, a regulatory framework that prospectively aims to prevent injuries (see discussion in *Tools to Promote Trust*), coupled with mechanisms to assign accountability and compensate patients if problematic outcomes occur, must exist. Because AI-DDS tools sit at the intersection of technology and clinical practice, there are two potential avenues for compensating patient injuries through the American tort system. The first is medical malpractice, which implies that the ultimate responsibility for problematic clinical decisions rests with the provider. The second is product liability, which implies that the responsibility for problematic clinical decisions rests instead with the developer and manufacturer of the AI-DDS tool.

Currently, the dividing line appears to be whether an independent professional, such as an end-user provider, could review the recommendations from an AI tool and understand how it arrived at them. As commentators note:

The Cures Act parses the product/practice regulatory distinction as follows: Congress sees it as a medical practice issue (instead of a product regulatory issue) to make sure health care professionals safely apply CDS [clinical decision support] software recommendations that are amenable to independent professional review. In that situation, safe and effective use of CDS software is best left to clinicians and to their state practice regulators, institutional policies, and the medical profession. When CDS software is not intended to be independently reviewable by the health care provider at the point of care, there is no way for these bodies to police appropriate clinical use of the software. In that situation, the Cures Act tasks the FDA with overseeing its safety and effectiveness. Doing so has the side effect of exposing CDS software developers to a risk of product liability suits (Evans and Pasquale, 2022).

This distinction is a workable and sensible one, reflecting the limitations of the average provider's abilities to evaluate new AI-DDS tools. It would be helpful to educate providers and hospital administrators on the dividing line between explainable CDS tools, which allow health care providers to understand and challenge the basis for algorithmic decision making and "black box" algorithms, for which the basis of algorithmic decisions making is obscure, on the other hand. This distinction carries

implications for liability insofar as courts may hesitate to hold providers accountable for "black box" tools that precluded the possibility of provider control. Providers who hesitate to adopt AI-DDS out of fear of medical malpractice liability may find that distinction comforting and trust-building. For patient injuries arising when AI-DDS systems are in use, policymakers and courts may wish to consider shifting the balance of liability from the current norm (which focuses almost entirely on medical malpractice) to one that also includes product liability in situations where the AI tool, rather than the provider, appears primarily at fault. This shift could further encourage trust and desire to use these tools among providers and would incentivize developers to design algorithms and select training data with a view to minimizing poor outcomes.

Product liability generally arises when a product inflicts "injuries that result from poor design, failure to warn about risks, or manufacturing defects" (Maliha et al., 2021). Product liability, to date, has only been applied in limited and inconsistent fashion to software in general and to health care software in particular (Brown and Miller, 2014). For example, in *Singh v. Edwards Lifesciences Corp*, the court permitted a jury to award damages against a developer because its software resulted in a catheter malfunctioning (CaseText, 2009b). On the other hand, in *Mracek v. Bryn Mawr Hospital*, a court rejected via summary judgment the plaintiff's argument that product liability should be imposed when the da Vinci surgical robot malfunctioned in the course of a radical prostatectomy (CaseText, 2009a). Further complicating the product liability landscape, the Supreme Court concluded in *Riegel v. Medtronic* that devices going through the

FDA premarket approval process, as opposed to other market authorization pathways such as 510(k) clearance, can enjoy certain protection against state product liability cases (CaseText, 2008). Thus, available redress for patients can vary depending on the market authorization pathway for the specific AI tool. The conflicting and limited case law in this area suggests that there is room to explore an expanded product liability landscape for AI-DDS software. One clear point from prior case law is that clinicians will bear the brunt of liability for injuries that occur when using AI-DDS tools “off-label” (e.g., using a tool that warns it is only intended for use on one patient population on a different population). This fact may help incentivize AI tool developers to disclose limitations of their training data since doing so can shift liability to providers who venture beyond the tool’s intended use.

It is also important to note that opening the door to product liability suits does not foreclose the potential for medical malpractice suits against providers who use AI-DDS tools. A provider who relies on AI-DDS tools in good faith could still face medical malpractice liability if their actions fall below the generally accepted standard of care for use of such tools or if the AI-DDS tool is used “off label”, i.e. using an AI-DDS tool developed for one type of MRI interpretation on another type of MRI image (Prince et al., 2019). Overall, courts are reluctant to excuse physician liability, allowing malpractice claims to proceed against physicians even in cases where:

1. there was a mistake in the medical literature or an intake form;
2. a pharmaceutical company failed to warn of a therapy’s adverse effect; or

3. there were errors by system technicians or manufacturers (Maliha et al., 2021).

These cases, taken together, suggest that providers cannot simply point to an AI-DDS error as a shield from medical malpractice liability.

Eventually, widespread adoption of AI-DDS could open the door for medical malpractice liability for providers who do not incorporate these tools into their practice, i.e., “failure to use”. Physicians, specifically, open themselves to medical malpractice liability when they fail to deliver care at the level of a competent physician of their specialty (Price et al., 2019). Currently, the standard of care does not include relying on AI-DDS tools. But as more and more providers incorporate AI-DDS tools into their practice, that standard may shift. Once the use of AI-DDS is considered part of the standard of care, medical malpractice liability will create a strong incentive for all providers to rely on these tools, regardless of their personal views on appropriateness.

Tools to Promote Trust

Two of the most impactful mechanisms to promote trust in AI-DDS among clinicians (and, thus, improving desire to use) would be to further refine the existing regulatory landscape for AI-DDS tools and to promote collaborations between stakeholders. This section of the paper explores avenues to promote trust.

To minimize concerns about liability, nuanced, thoughtful regulation and governance from all levels of the U.S. government—federal, state, and local—can reassure providers that they can trust

available AI-DDS tools and move forward with implementation. A key factor affecting clinicians' willingness to adopt AI-DDS tools is likely whether the tools will receive a rigorous, data-driven review of safety and effectiveness by the FDA before moving into clinical use. A potential concern is that some, but not necessarily all, AI-DDS software is subject to FDA medical device regulation under the Cures Act. It remains difficult for providers to intuit whether a given type of AI-DDS tool is or is not likely to have received oversight under FDA's medical device regulations. Uncertainty about which tools will receive FDA oversight—and which marketing authorization process the FDA may require (e.g., premarket approval, 510(k), or de novo classification)—likely fuels provider discomfort with using AI-DDS tools.

A key source of this uncertainty, at present, is that the Cures Act addresses the scope of the FDA's power to regulate various types of medical software but does not itself define or use the terms DDS or CDS software (114th Congress, 2016a; 21 U.S. Code § 360j, 2017). As used in this paper, AI-DDS tools broadly refer to computer-based tools, driven by AI algorithms, that use clinical knowledge and patient-specific health information to inform health care providers' diagnostic decision-making processes (see *Table 1*), with DDS tools being a subset of CDS tools more generally. This paper thus follows the definition

provided by the Office of the National Coordinator for Health Information Technology (ONC), which stresses that CDS tools “provide ... knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care” (ONC, 2018). The FDA has used this ONC definition when discussing how CDS software is broadly understood (FDA, 2019b). Central to the ONC definition, and this paper, is the notion that DDS and CDS tools combine general medical “knowledge” with patient-specific information to produce recommended diagnoses. With AI-DDS systems, that knowledge can include inferences generated internally by an AI/ML algorithm.

The Cures Act authorizes the FDA to regulate only some of the software that might fit into the broader, more common conception of AI-DDS systems just described. Thus, FDA lacks authority to regulate all of the tools that clinicians might think of as being DDS/CDS tools. The Cures Act expressly excludes five categories of medical software from the definition of a “device” that FDA can regulate (114th Congress, 2016a [21 U.S.C. § 360j(o)(1), 2017]). One of these exclusions places restrictions on FDA's power to regulate CDS and DDS software (114th Congress, 2016a [21 U.S.C. § 360j(o)(1)(E)]). *Box 1* below shows the specific wording of the relevant Cures Act exclusion.

Box 1 | Provisions of the Cures Act that Exclude Some AI-DDS Tools from FDA Oversight

Section 3060 of the Cures Act, codified at Title 21 of the U.S. Code, Section 360j(o)(1)(E), excludes certain medical software from being treated as a “device” that the FDA can regulate.

Basic exclusion from the medical device definition. Subject to the two specific exceptions noted below, software is not an FDA-regulable medical device if it is intended:

“for the purpose of –

- (i) displaying, analyzing, or printing medical information about a patient or other medical information (such as peer-reviewed clinical studies and clinical practice guidelines);
- (ii) supporting or providing recommendations to a health care professional about prevention, diagnosis, or treatment of a disease or condition; and
- (iii) enabling such health care professional to independently review the basis for such recommendations that such software presents so that it is not the intent that such health care professional rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient.” (21 U.S.C. § 360j(0)(1)(E)(i)-(iii)).

Exceptions. Two exceptions allow software that meets the above description to nevertheless be regulated by the FDA as a medical device. These exceptions are:

1. Jurisdictional saving clause. The opening passage of Section 360j(o)(1)(E) contains a “saving” clause preserving the FDA’s authority to regulate certain software that meets the above three conditions. This clause states that the basic exclusion just quoted applies to a software tool “unless the function is intended to acquire, process, or analyze a medical image or a signal from an *in vitro* diagnostic device or a pattern or signal from a signal acquisition system” [emphasis added]. Put more simply, a tool is not excluded from being an FDA-regulable device, if its function is to acquire, process, or analyze images or signals of such types.
2. Procedure for overriding the basic exclusion. The Secretary of HHS can restore the FDA’s power to regulate a CDS or DDS tool that otherwise would fit into the basic exclusion, by making a finding that use of the tool “would be reasonably likely to have serious adverse health consequences” and issuing a final order after notice and public comment (21 U.S.C. § 360j(o)(3)). Through this procedure, the Secretary has the power to determine that the tool is a medical device and therefore subject to FDA oversight.

Source: 114th Congress, 2016a.

Looking at the basic exclusion in *Box 1*, the first two conditions, (i) and (ii), describe CDS and DDS software without using those names. The third condition, shown at (iii), bears on the concept this paper refers to as explainability, again without using that term. When all three conditions are met, this passage of the Cures Act creates a potential exclusion from FDA regulation for CDS/DDS software that meets the criterion for explainability set out in condition (iii) of *Box 1*. This exclusion, however, is subject to the two exceptions shown at the bottom of *Box 1*.

The first exception—the saving clause—confirms the FDA’s power to regulate many

types of software whose function supports diagnostic testing, such as software used in the bioinformatics pipeline for genomic testing. Before the Cures Act, FDA’s medical device authority included oversight covering both *in vitro* diagnostic devices (which support clinical laboratory testing of biospecimens) and *in vivo* devices (such as X-rays and MRI machines that produce images of tissues within a patient’s body). FDA has long regulated software embedded in diagnostic hardware devices, for example, software internal to sequencing analyzers and MRI machines. The saving clause confirms FDA’s power to regulate “stand-alone” diagnostic software that is not necessarily part of a hardware device

but processes signals from in vitro and in vivo testing devices.

This power is crucial in light of the modern trend for many clinical laboratories to use third-party software service providers and vendors for data analysis supporting complex diagnostic tests, such as genomic tests (Curnutte et al., 2014). *In vitro* diagnostic testing by clinical laboratories is subject to the Clinical Laboratory Improvement Amendments of 1988 (CLIA) regulations (100th Congress, 1988). The CLIA framework focuses on the quality of clinical laboratory services but does not provide an external, data-driven regulatory review of the safety and effectiveness of tests used in providing those services, nor does it evaluate the software laboratories use when analyzing and interpreting test results. FDA’s authority to regulate stand-alone diagnostic software positions FDA to oversee clinical laboratory software, even in situations where FDA exercises discretion and declines to regulate an underlying laboratory-developed test (Evans et al., 2020). In a 2019 draft guidance document, circulated for comment purposes only, the FDA noted that “bioinformatics products used to process high volume ‘omics’ data (e.g., genomics, proteomics, metabolomics) process a signal from an in vitro diagnostic (IVD) and are generally not considered to be CDS” tools (FDA, 2019b). The saving clause clarifies that FDA can regulate such software, even in situations where it might technically be considered CDS software falling within the basic exclusion in *Box 1* (114th Congress, 2016a [21 U.S.C. § 360j(o)(1)(E)]).

Much of the AI-DDS software providers use in clinical health care settings would not fall under the saving clause (see *Box 1*), which seems directed at software processing

signals from diagnostic devices as part of the workflow for producing finished diagnostic test reports and medical images. However, there is some ambiguity. An example would be an AI-DDS tool that analyzes several of a patient’s gene variants along with the patient’s reported symptoms, clinical observations, treatment history, and environmental exposures to recommend a diagnosis to a clinician. It is unclear if the fact that the tool processes gene variant data means that it is “processing a signal from an IVD device” and thus FDA-regulated, or if the saving clause only applies when the signal is directly fed to the software as part of the clinical laboratory workflow. Without knowing how the FDA interprets the breadth of the saving clause, it is hard for clinicians to understand what is and is not regulated.

Assuming the saving clause does not apply, AI/DDS tools are generally excluded from FDA regulation if they meet all three of the conditions listed at (i)-(iii) in *Box 1*. The first two conditions are fairly straightforward, but it is still not clear how the FDA plans to assess whether the third condition, bearing on the concept of explainability, has been met. How, precisely, the FDA will decide whether an AI/DDS tool is “intended” to be “for the purpose of” “enabling [a] health care professional to independently review the basis for [its] recommendations” (see *Box 1*) is unknown. The FDA’s regulation on the “Meaning of intended uses” offers insight into the range of direct and circumstantial evidence the agency can consider when assessing objective intent (FDA, 2017b [21 C.F.R. § 801.4]). Yet how the agency will apply those principles in the specific context of AI/ML software tools is not clear.

Without greater clarity on these matters, clinicians lack a sense of whether a given type of AI-DDS tool usually is, or usually is not, subject to FDA oversight or what FDA’s oversight process entails. Almost six years after the Cures Act, FDA’s approach for regulating AI/ML CDS/DDS software remains a work in progress, leaving uncertainties that can erode clinicians’ confidence when using these tools. Through two rounds of draft guidance (in 2017 and 2019), the FDA solicited public comments to clarify its approach to regulating CDS/DDS tools. A final guidance on Clinical Decision Support Software appears on the list of “prioritized device guidance documents the FDA intends to publish during FY2022” (October 1, 2021 – September 30, 2022) (FDA, 2021c). As this paper went to press in mid-September 2022, the final guidance was not yet available, but the authors hope it may clarify these and other unresolved questions around the regulation of CDS/DDS tools.

Unfortunately, guidance documents—whether draft or final—have no binding legal effect and do not establish clear, enforceable legal rights and duties on which software developers, clinicians, state regulators, and members of the public can rely. There is fairly wide scholarly agreement that the use of guidance as a regulatory tool can be appropriate for emerging technologies where knowledge is rapidly evolving and flexibility is warranted, but there can be long-term costs when agencies choose to rely on guidance and voluntary compliance instead of promulgating enforceable regulations (Wu, 2011; Cortez, 2014). FDA’s Digital Innovation Action Plan (FDA, 2017a; Gottlieb, 2017) and its Digital Health Software Precertification (Pre-Cert) Program (FDA, 2021b) both acknowledge

that its traditional premarket review process for moderate and higher-risk devices is not well suited for “the faster iterative design, development, and type of validation used for software-based medical technologies” (FDA, 2017a). The FDA’s 2021 AI/ML Action Plan envisions incorporating ongoing post-marketing monitoring and updating of software tools after they enter clinical use (FDA, 2021a). This may leave health care providers in the uncomfortable position of using tools that may be modified even after the FDA clears them for clinical use and potentially facing liability if patient injuries occur. Also, it implies that vendors and developers of AI/ML tools will need access to real-world clinical health care data to support ongoing monitoring of how the tools perform in actual clinical use.

Future reliance on post-marketing monitoring offers an example of why regulating via non-binding guidance documents can create long-term problems. The HIPAA Privacy Rule contains an exception that lets HIPAA-covered health care providers, such as hospitals, share data with device manufacturers to help them meet their FDA regulatory compliance obligations (for example, to help manufacturers comply with the FDA’s adverse-event reporting requirements) (HHS, 2003). Unfortunately, when FDA regulates manufacturers by means of guidance documents and other non-mandatory programs, this important HIPAA pathway for accessing data may be unavailable, because guidance documents create no enforceable legal obligations. To maximize software developers’ access to real-world evidence for post-marketing monitoring and updating of AI/DDS tools, the FDA will ultimately need to set binding regulatory requirements (for example, for developers to monitor for racial, gender, or

other biases in the post market period). Related concerns surround the future development of state law, including both state regulations and tort law. Safe clinical use of AI/DDS tools will ultimately require state-level medical practice regulations and common law addressing issues such as appropriate staffing for, and use of, AI/DDS tools in clinical settings. To foster optimal development of state law, it is helpful to have federal regulations providing a stable demarcation between the FDA's role versus that of the states. Federal guidance documents, due to their non-binding nature and ease of revision, may not meet this need. The FDA's current heavy reliance on guidance documents and voluntary measures may be appropriate in the early years as AI/DDS tools emerge as a new technology, but the agency should stay mindful of the need to promulgate regulations whenever appropriate and feasible.

Apart from the regulatory framework, another mechanism to instill trust is through increased and consistent collaboration among developers, ethicists, and clinical diagnosticians during various phases of the AI lifecycle. Early innovation in the process of AI pre-market design, testing, clinical application, and post-market oversight resulted in fragmented and siloed professional stakeholder groups with different goals, expertise, ethical frameworks, and paradigms of professionalism and professional accountability. While a great deal of health care professional ethical attention, input, and engagement has been integrated into AI use and application in the post market phase, there has been an important gap in full integration of professional end-user partnership within the AI tool development

process needed to build trustworthy AI tools.

Numerous AI and digital health ethical frameworks have been published as part of the concerted effort to build trustworthy human-AI partnerships. For example, the European Commission's *Ethics Guidelines for Trustworthy AI* is a foundational work on the topic, with seven key requirements:

1. Human agency and oversight,
2. Technical robustness and safety,
3. Privacy and data governance,
4. Transparency,
5. Diversity, non-discrimination and fairness,
6. Environmental and societal well-being, and
7. Accountability (European Commission, 2019).

Additionally, over 40 different U.S. technology companies and venture capital firms have signed on to a *Responsible Innovations Charter*, with similar key principles:

1. Innovating intentionally,
2. Operating with accountability and transparency,
3. Advancing inclusive prosperity,
4. Building sustainably,
5. Respecting people,
6. Championing diversity, and
7. Promoting healthy societies (Responsible Innovation Labs, 2022).

The American Medical Association has developed policies and frameworks for practicing diagnosticians to govern and assess AI integration into clinical practice (Crigger et al., 2022). Essentially, the structured assessment aids the clinician in ascertaining: whether a tool is beneficial to patient outcomes; whether a tool appears

to work; and whether a tool appears to work for their patients. These guidelines, along with several global government-produced assessments for organizational leaders, provide a systematic and structured assessment for providers to select and utilize trustworthy and beneficial AI for their practice.

3 Ensuring and Promoting Health Equity in the Deployment of AI-Assisted Diagnostic Tools

In addition to facilitating uptake and overcoming barriers to the adoption of AI-DDS tools elucidated in this review, being cognizant of the implications for equity throughout the life cycle of these tools and making a consistent effort to address past, current, and potential equity issues are critical to preventing widening disparities in health care delivery. While there is excitement and demonstrated benefits to bringing AI-DDS tools into clinical practice, poor data quality, prevalent biases in health care, and a lack of structural supports available to end users jeopardize progress toward achieving health equity and fuel ongoing uncertainties and hesitations about adopting these tools.

AI/ML algorithms are often developed using limited data samples that may not represent the people they are meant to impact (Zou and Schiebner, 2021). Furthermore, social determinants of health data are generally not well captured in data sets used to train these algorithms. Data elements derived from diverse sources that could help provide a more holistic view of the patient may not be available to certain care settings due to the limitations of EHR systems, data privacy concerns, a lack of data standardization, and financial constraints on the part of health systems to obtain large data sets (Zusterzeel et al., 2022; Alami et al., 2020). Inaccurate representation in training, testing, and validation data sets also results in the development of flawed models. Models not accurately trained in the context that they are intended for may also have difficulty performing when there is a shift in

population demographics (Singh et al., 2020).

AI tools rely on human interaction from their inception to deployment, and AI algorithms can replicate explicit and implicit biases in human decision making in health care settings (Char et al., 2018). Inherent discrimination occurring within care delivery can be challenging to predict and uncover, and biases could easily transfer over into the design and use of AI algorithms (Leslie et al., 2021; Char et al., 2018). For example, the biases of developers, researchers, and designers can manifest early in the development phase if they choose target variables and proxies for those variables without considering upstream social determinants of health and related confounders (Leslie et al., 2021). Along with the data collection issues summarized above, other data extraction and measurement errors due to biases built into physical devices can negatively influence care decisions and perpetuate inequities (Leslie et al., 2021; Zou and Schiebner, 2021). In the case of the pulse oximeter, this medical device uses infrared and red light signaling that interacts with skin pigmentation to read the oxygen saturation in the patient's blood and shows varying results based on skin color (Zou and Schiebner, 2021). Previous studies have shown how patients with darker skin received inaccurate oxygen readings compared to White patients (Leslie et al., 2021; Zou and Schiebner, 2021). This data is fed into algorithms to assist with decision making, and clinicians may unintentionally accept results and act on flawed recommendations, affecting the ability of

patients to acquire needed care, such as supplementary oxygen (Zou and Schiebner, 2021; Rajkomar et al., 2018).

In addition to the adverse effects of incorrect data usage and biases, the absence of infrastructure to support equitable AI in developing and deploying AI-DDS tools will ultimately widen disparities. The digital gap perpetuates inequities through many social factors that may intertwine, including a lack of broadband internet access across regions and an inability to purchase up-to-date and well-equipped devices (Ramsetty and Adams, 2020). For example, AI tools extracting data from EHR systems may be more prevalent in larger health care organizations in well-resourced cities than small rural hospitals or physician practices, which have fewer resources and expertise readily available (Goldfarb and Teodoridis, 2022; Reisman,

2017). The associated financial costs for EHR implementation continue to be a primary barrier to the adoption of AI-DDS tools (Goldfarb and Teodoridis, 2022). AI algorithms applied to clinical settings that disproportionately serve populations that experience a form of privilege (i.e., wealthy populations) marginalize groups that do not actively seek care in the same settings (DeCamp and Lindvall, 2020; Rajkomar et al., 2018). Nevertheless, data collection issues persist in settings with EHR systems due to the lack of compatibility between these systems and certain providers serving different hospitals and health care facilities, further contributing to data silos and insufficiently informed AI tools (Goldfarb and Teodoridis, 2022).

4 Path Forward – Policy Implications and Action Priorities

Fostering provider adoption of novel AI-DDS systems will require broad infrastructural support, beginning with robust tool evaluations by health systems and payers, clear commitments from health systems and developers to regular monitoring and updating of algorithms, and training care teams to effectively interpret and implement changes based on AI-DDS outputs. Developers, payers, health systems, and providers are becoming increasingly aware of potential biases in AI algorithms and their deployment. Data representativeness and robust model training must be a top priority in algorithm development to increase trust and adoption among all relevant stakeholders. Data integrity and reliability are at the very core of sound algorithm development, yielding better prospects for provider adoption of those algorithms. Therefore, collaborative efforts aimed at curating rich and multimodal patient data—including crucial social determinants information—will be paramount. Such efforts need to be coupled with robust and consistent standards for data access, sharing, harmonization, and interoperability, while simultaneously prioritizing data privacy and security to ultimately drive excellent model development. In a similar vein, boosting provider comfort and adoption may also depend on model transparency. Providing health care teams with key parameters driving an AI-DDS output that can serve as modifiable targets for patient outcome improvement may facilitate greater adoption. To conclude, this paper presents key action priorities in each of the four domains related to provider adoption of AI-DDS tools outlined in this paper:

Domain 1: Reason to Use

- Establishing clear impetus to incorporate novel AI-DDS tools into health systems is contingent on a given tool’s clinical efficacy, specifically as it relates to a health system’s target population, and affordability, both to the health system and patient. Developers, payers, health systems, and providers are becoming increasingly aware of potential biases in AI algorithms and their deployment. Data representativeness and robust model training and testing must be the top priority in algorithm development in efforts to increase trust and adoption among all relevant stakeholders.
- Collaborative efforts among multiple health care systems aimed at curating rich and multimodal patient data—including essential social determinants information—will be paramount. Such efforts need to be coupled with robust and consistent standards for data access, sharing, and interoperability, while simultaneously prioritizing data privacy and security, to ultimately drive excellent model development.
- In addition to ensuring robust clinical utility, algorithm developers must design AI-DDS tools to integrate seamlessly into existing care team infrastructures, ensuring that their product value is not diminished by logistical inefficiency and cognitive burden.

Domain 2: Means to Use

- Policy makers and payers should consider promoting sustainability through reimbursement to create a sustainable environment for the adoption and continual use of AI-DDS tools and to further promote capital infrastructure investments by health systems to facilitate this goal.
- If consensus-based standards do not emerge, ensuring interoperability could require a “top-down” regulatory approach. For instance, the United States Office of the National Coordinator for Health Information Technology (ONC) could develop health IT certification criteria that assess the ability of EHR systems to support data lifecycles. However, given the nascent understanding of ideal workflows and life cycles, standardization at this time is likely premature.
- Policy makers and payers should consider using incentives to encourage the use of evidence-based AI-DDS in clinical practice. As per prior payment models, if adoption is sufficient and the evidence of improved processes and outcomes becomes established, AI-DDS tools may become the standard of care in specific clinical scenarios.

Domain 3: Method to Use

- Public and private research funders should increase focus and funding opportunities to advance the still nascent implementation science of AI-DDS, for example, through RFPs that focus on integrating AI-DDS into clinical workflows and health IT systems and its

impact on the behaviors of clinical teams.

- Institutions of medical education and accreditation organizations should review emerging competencies for the use of AI-DDS and consider how to integrate these into the current training and certification ecosystem to adapt to the rapidly changing needs of the clinical front line.
- Professional societies, trade associations, and health care quality organizations should identify diagnostic centers of excellence that specialize in AI-DDS to facilitate the surfacing and effective diffusion of best practices through interdisciplinary learning networks and capacity-building programs.
- Software and algorithm designers of point-of-care AI-DDS for providers and patients at home should leverage the public SMART on FHIR and SMART/HL7 Bulk FHIR APIs regulated under the ONC 21st Century Cures Act Rule, so that algorithms can be widely and uniformly integrated into care across EHR vendor products and other IT tools.
- Regulators should monitor, for example through the 21st Century Cures Act EHR Reporting Program, EHR vendor implementation of public FHIR APIs to ensure their turnkey use by apps made accessible at the point of care.

Domain 4: Desire to Use

- Professional societies, trade associations, and health care quality organizations should center AI-related efforts to promote clinician well-being

through human-centered design in AI technology, aligned with the work-life balance of health care professionals outlined in the Quintuple Aim. The FDA should offer guidance and/or other communications, specifically tailored to health care providers tasked with using AI/DDS tools, to aid their understanding of the types of software are – and are not – likely to receive FDA oversight under 21 U.S.C. § 360j(o)(1)(E). Specifically, it will be imperative to clarify how broadly the agency construes the saving clause for “software that processes signals...”, and the agency’s approach for assessing whether software is “intended ... for the purpose ... of enabling” a health care professional to independently review the basis of its recommendations. Encouraging clinicians to trust these tools may require helping them develop an intuitive grasp of the FDA’s role and its jurisdictional limits.

- The FDA should continue to explore the special considerations affecting design, validation review, market authorization, and post marketing oversight for AI-DDS tools, offering timely guidance while recognizing that, over the long term, notice-and-comment rulemaking may offer advantages over the continued use of guidance documents – for example – to enhance developers’ access to HIPAA-protected real-world data for use in regulatory compliance activities, and to provide needed clarity and stability to foster development of state regulations and common law addressing clinical use of AI-DDS systems.

- Professional medical, nursing, and other health care societies should develop clinical practice guidelines for AI system applications.
- The FDA, CDC, and ONC should ensure transparency and publicly accessible reporting for flaws and safety incidents related to AI-DDS tools, malfunctions, and patient harm.
- Software developers should integrate human clinical diagnosticians at all phases of software development, design, validation, implementation, and iterative improvements.

AI-DDS systems are becoming increasingly prevalent, sophisticated, and reliable. Across medical specialties, these tools demonstrate potential to make the clinical diagnostic process more efficient and accurate, ultimately improving patient outcomes. Focused efforts to create equitable and robust AI-DDS algorithms, streamline integration of new AI-DDS tools into clinical workflows, and train health care providers to effectively use such tools—coupled with strong regulatory oversight and financial incentives—will optimize the likelihood that innovative, clinically impactful AI-DDS systems are adopted and used responsibly by health care providers to the ultimate benefit of their patients.

References

1. 21 U.S. Code § 360j. 2017. *General provisions respecting control of devices intended for human use*. Available at: <https://www.law.cornell.edu/uscode/text/21/360j> (accessed July 26, 2022).
2. 100th Congress. 1988. *Public Law 100-578, 102 STAT. 2903*. Available at: <https://www.govinfo.gov/content/pkg/STATUTE-102/pdf/STATUTE-102-Pg2903.pdf> (accessed July 27, 2022).
3. 114th Congress. 2016a. *H.R. 34 – 21st Century Cures Act*. Available at: <https://www.congress.gov/bill/114th-congress/house-bill/34/text> (accessed July 26, 2022).
4. 114th Congress. 2016b. *S.524 – Comprehensive Addiction and Recovery Act of 2016*. Available at: <https://www.congress.gov/bill/114th-congress/senate-bill/524/text> (accessed July 27, 2022).
5. Abbas, H., F. Garberson, S. Liu-Mayo, E. Glover, and D. P. Wall. 2020. Multi-modular AI Approach to Streamline Autism Diagnosis in Young Children. *Scientific Reports* 10(5014). <https://doi.org/10.1038/s41598-020-61213-w>.
6. Abdulkareem, M., and S. E. Petersen. 2021. The Promise of AI in Detection, Diagnosis, and Epidemiology for Combating COVID-19: Beyond the Hype. *Frontiers in Artificial Intelligence*. <https://doi.org/10.3389/frai.2021.652669>.
7. Adiekum, A., A. Blasimme, and E. Vayena. 2018. Elements of Trust in Digital Health Systems: Scoping Review. *J Med Internet Res* 20(12):e11254. doi:10.2196/11254.
8. Aggarwal, N., M. Ahmed, S. Basu, J. J. Curtin, B. J. Evans, M. E. Matheny, S. Nundy, M. P. Sendak, C. Shachar, R. U. Shah, and S. Thadaney-Israni. 2020. Advancing Artificial Intelligence in Health Settings Outside the Hospital and Clinic. *NAM Perspectives*. Discussion Paper, National Academy of Medicine, Washington, DC. <https://doi.org/10.31478/202011f>.
9. Ajzen, I. 1985. From Intentions to Actions: A Theory of Planned Behavior. In *Action Control*, edited by J. Kuhl and J. Beckmann. Berlin: Springer. Pp. 11-39.
10. Ajzen, I. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50(2):179-211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T).
11. Alami, H., P. Lehoux, Y. Auclair, M. de Guise, M. P. Gagnon, J. Shaw, D. Roy, R. Fleet, M. A. Ag Ahmed, and J. P. Fortin. 2020. Artificial Intelligence and Health Technology Assessment: Anticipating a New Level of Complexity. *Journal of Medical Internet Research* 22(7). <https://doi.org/10.2196/17707>.
12. Anumana. 2022. *About Us*. Available at: <https://www.anumana.ai/aboutus> (accessed May 12, 2022).
13. Ardila, D., A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* (6):954-961. <https://doi.org/10.1038/s41591-019-0447-x>.

14. Barker, W., and C. Johnson. 2021. The Ecosystem of Apps and Software Integrated with Certified Health Information Technology. *Journal of the American Medical Informatics Association* 28(11):2379-2384. <https://doi.org/10.1093/jamia/ocab17>.
15. Benjamins, S., P. Dhunoo, and B. Meskó. 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine* 3(118). <https://doi.org/10.1038/s41746-020-00324-0>.
16. Berger, D. 1999. A brief history of medical diagnosis and the birth of the clinical laboratory. Part 1—Ancient times through the 19th century. *MLO: Medical Laboratory Observer* 31(7):28-30, 32, 34-40. Available at: <https://pubmed.ncbi.nlm.nih.gov/10539661/> (accessed July 26, 2022).
17. Benjamins, R. 2021. A choices framework for the responsible use of AI. *AI and Ethics* 1(1):49-53. <https://doi.org/10.1007/s43681-020-00012-5>.
18. Bitterman, D. S., H. J. W. L. Aerts, and R. H. Mak. 2020. Approaching autonomy in medical artificial intelligence. *The Lancet Digital Health* 2(9):e447-e449. [https://doi.org/10.1016/S2589-7500\(20\)30187-4](https://doi.org/10.1016/S2589-7500(20)30187-4).
19. Brajer, N., B. Cozzi, M. Gao, M. Nichols, M. Revoir, S. Balu, J. Futoma, J. Bae, N. Setji, A. Hernandez, and M. Sendak. 2020. Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. *JAMA Network Open* 3(2):e1920733. <https://doi.org/10.1001/jamanetworkopen.2019.20733>.
20. Brown, S. H., and R. A. Miller. 2014. Legal and regulatory issues related to the use of clinical software in health care delivery. In *Clinical Decision Support*, 2nd edition, edited by R. A. Greenes. New York: Elsevier.
21. CaseText. 2008. *Riegel v. Medtronic, Inc.* Available at: <https://casetext.com/case/riegel-v-medtronic-inc-3> (accessed September 16, 2022).
22. CaseText. 2009a. *Mracek v. Bryn Mawr Hospital*. 610 F Supp 2d, 401 (ED Pa 2009). Available at: <https://casetext.com/case/mracek-v-bryn-mawr-hosp-2> (accessed July 26, 2022).
23. CaseText. 2009b. *Singh v. Edwards Lifesciences*. Available at: <https://casetext.com/case/singh-v-edwards-lifesciences> (accessed July 27, 2022).
24. Char, D., N. Shah, and D. Magnus. 2018. Implementing machine learning in health care – addressing ethical challenges. *New England Journal of Medicine* 378(11):981-983. <https://doi.org/10.1056/NEJMp1714229>.
25. Chen, M. M., L. P. Golding, and G. N. Nicola. 2021. Who Will Pay for AI? *Radiology: Artificial Intelligence* 3(3). <https://doi.org/10.1148/ryai.2021210030>.

26. Clemens, J., and J. D. Gottlieb. 2017. In the Shadow of a Giant: Medicare's Influence on Private Physician Payments. *Journal of Political Economy* 125(1):1-39. <https://www.journals.uchicago.edu/doi/10.1086/689772>.
27. Cortez, N. 2014. REGULATING DISRUPTIVE INNOVATION. *Berkeley Technology Law Journal* 29:175-218. <http://dx.doi.org/10.2139/ssrn.2436065>.
28. Crigger, E., K. Reinbold, C. Hanson, A. Kao, K. Blake, and M. Irons. 2022. Trustworthy Augmented Intelligence in Health Care. *Journal of Medical Systems* 46(12). <https://doi.org/10.1007/s10916-021-01790-z>.
29. Curnutte, M. A., K. L. Frumovitz, J. M. Bollinger, A. L. McGuire, and D. J. Kaufman. 2014. Development of the clinical next-generation sequencing industry in a shifting policy climate. *Nature Biotechnology* 32(10):980-982. <https://doi.org/10.1038/nbt.3030>.
30. DeCamp, M., and C. Lindvall. 2020. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association* 27(12):2020-2023. <https://doi.org/10.1093/jamia/ocaa094>.
31. Deverka, P. A., and J. C. Dreyfus. 2014. Clinical Integration of Next Generation Sequencing: Coverage and Reimbursement Challenges. *Journal of Law, Medicine & Ethics* 42:22-41. doi: 10.1111/jlme.12160.
32. Digital Diagnostics. 2022. *IDx-DR*. Available at: <https://www.digitaldiagnostics.com/products/eye-disease/idx-dr/> (accessed July 26, 2022).
33. Digital Diagnostics. 2019. *Autonomous AI diagnostics launch in retail health clinics*. November 19. Available at <https://www.digitaldiagnostics.com/newsroom/autonomous-ai-diagnostics-launch-in-retail-health-clinics/> (accessed on May 11, 2022).
34. Duffy, G., P. P. Cheng, N. Yuan, B. He, A. C. Kwan, M. J. Shun-Shin, K. M. Alexander, J. Ebinger, M. P. Lundgren, F. Rader, D. H. Liang, I. Schnittger, E. A. Ashley, J. Y. You, J. Patel, R. Witteles, S. Cheng, and D. Ouyang. 2022. High-Throughput Precision Phenotyping of Left Ventricular Hypertrophy with Cardiovascular Deep Learning. *Journal of the American Medical Association, Cardiology* 7(4):386–395. <https://doi.org/10.1001/jamacardio.2021.6059>.
35. Escobar, G. J., V. X. Liu, A. Schuler, B. Lawson, J. D. Greene, and P. Kipnis. 2020. Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration. *New England Journal of Medicine* 383(20):1951-1960. <https://doi.org/10.1056/NEJMsa2001090>.
36. European Commission. 2019. *Ethics Guidelines for Trustworthy AI*. High-level expert group on Artificial Intelligence. Available at: <https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf> (accessed on March 10, 2022).

37. Evans, B. and F. Pasquale. 2022. Product Liability Suits for AI/ML Software 22-46, in *The Future of Medical Device Regulation: Innovation and Protection*, edited by I. G. Cohen, T. Minsin, W. N. Price II, C. Robinson, and C. Shachar. London: Cambridge University Press.
38. Evans, B., and P. Ossorio. 2018. The Challenge of Regulating Clinical Decision Support Software After 21st Century Cures. *American Journal of Law & Medicine* 44(2-3):237-251. <https://doi.org/10.1177/0098858818789418>.
39. Evans, B. J., G. Javitt, R. Hall, M. Robertson, P. Ossorio, S. M. Wolf, T. Morgan, and E. W. Clayton. 2020. How Can Law and Policy Advance Genomic Analysis and Interpretation for Clinical Care? *Journal of Law, Medicine, and Ethics* 48 (Supp 1):44-68. <https://doi.org/10.1177/1073110520916995>.
40. Fenton, J. J., S. H. Taplin, P. A. Carney, L. Abraham, E. A. Sickles, C. D'Orsi, E. A. Berns, G. Cutter, E. Hendrick, W. E. Barlow, and J. G. Elmore. 2007. Influence of Computer-Aided Detection on Performance of Screening Mammography. *N Engl J Med* 365:1399-1409. DOI: 10.1056/NEJMoa066099.
41. U.S. Food and Drug Administration (FDA). 2021a. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. Available at: <https://www.fda.gov/media/145022/download> (accessed on May 11, 2022).
42. FDA. 2021b. *Digital Health Software Precertification (Pre-Cert) Program*. Available at: <https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program> (accessed on May 11, 2022).
43. FDA. 2021c. CDRH Proposed Guidances for Fiscal Year 2022 (FY2022). Available at: <https://www.fda.gov/medical-devices/guidance-documents-medical-devices-and-radiation-emitting-products/cdrh-proposed-guidances-fiscal-year-2022-fy2022> (accessed on September 14, 2022).
44. FDA. 2019a. *Clinical Decision Support Software: Draft Guidance for Industry and Food and Drug Administration Staff*. Pp. 28. Available at: <https://www.fda.gov/media/109618/download> (accessed on May 11, 2022).
45. FDA. 2019b. *Clinical Decision Support Software: Draft Guidance for Industry and Food and Drug Administration Staff*. Available at: <https://www.fda.gov/media/109618/download> (accessed on May 11, 2022).
46. FDA. 2017a. *Digital Health Innovation Action Plan*. Available at: <https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/UCM568735.pdf> (accessed on May 11, 2022).
47. FDA. 2017b. *CFR - Code of Federal Regulations Title 21*. Available at: <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=801.4> (accessed September 16, 2022).
48. FDA. 2016. *Step 3: Pathway to Approval*. Available at: <https://www.fda.gov/patients/device-development-process/step-3-pathway-approval> (accessed May 15, 2022).

49. GlobeNewswire. 2020. *Anaconda Releases 2020 State of Data Science Survey Results*. Available at <https://www.globenewswire.com/news-release/2020/06/30/2055578/0/en/Anaconda-Releases-2020-State-of-Data-Science-Survey-Results.html> (accessed May 11, 2022).
50. Goh, K. H., L. Wang, A. Y. K. Yeow, H. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan. 2021. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun* 12. <https://doi.org/10.1038/s41467-021-20910-4>.
51. Goldfarb, A., and F. Teodoridis. 2022. Why is AI adoption in health care lagging? *Brookings*, March 9. Available at: <https://www.brookings.edu/research/why-is-ai-adoption-in-health-care-lagging/> (accessed May 17, 2022).
52. Goldhahn J., V. Rampton, and G. A. A. Spinaz. 2018. Could artificial intelligence make doctors obsolete? *BMJ* 363:k4563. <https://doi.org/10.1136/bmj.k4563>.
53. Gottlieb, S. 2019. *FDA Announces New Steps to Empower Consumers and Advance Digital Healthcare*. Available at <https://www.fda.gov/news-events/fda-voices/fda-announces-new-steps-empower-consumers-and-advance-digital-healthcare> (accessed on May 11, 2022).
54. He, J., S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine* 25:30-36. <https://doi.org/10.1038/s41591-018-0307-0>.
55. U.S. Department of Health and Human Services (HHS). 2003. *Disclosures for Public Health Activities*. Available at: <https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/disclosures-public-health-activities/index.html> (accessed September 16, 2022).
56. U.S. Department of Health and Human Services (HHS). 2020. *21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program*. Available at: <https://www.federalregister.gov/d/2020-07419> (accessed July 26, 2022).
57. Heartflow. 2014. *Heartflow Secures De Novo Clearance from the U.S. Food and Drug Administration for Breakthrough FFR_{CT} Technology*. Available at: <https://www.heartflow.com/newsroom/heartflow-secures-de-novo-clearance/> (accessed March 15, 2022).
58. High-Level Expert Group on AI (AI HLEG). 2019. *Ethics Guidelines for Trustworthy AI*.
59. Hinton, G. 2016. *On Radiology*. Available at: <https://www.youtube.com/watch?v=2HMPRXstSvQ> (accessed May 15, 2022).
60. Kaufman Hall & Associates. 2022. *National Hospital Flash Report*. Available at: <https://www.kaufmanhall.com/sites/default/files/2022-03/National-Hospital-Flash-Report-March-2022.pdf> (accessed May 25, 2022).

61. Kawamoto, K., P. V. Kukhareva, C. Weir, M. C. Flynn, C. J. Nanjo, D. K. Martin, P. B. Warner, D. E. Shields, S. Rodriguez-Loya, R. L. Bradshaw, R. C. Cornia, T. J. Reese, H. S. Kramer, T. Taft, R. L. Curran, K. L. Morgan, D. Borbolla, M. Hightower, W. J. Turnbull, M. B. Strong, W. W. Chapman, T. Gregory, C. H. Stipelman, J. H. Shakib, R. Hess, J. P. Boltax, J. P. Habboushe, F. Sakaguchi, K. M. Turner, S. P. Narus, S. Tarumi, W. Takeuchi, H. Ban, D. W. Wetter, C. Lam, T. J. Caverly, A. Fagerlin, C. Norlin, D. C. Malone, K. A. Kaphingst, W. K. Kohlmann, B. S. Brooke, and G. Del Fiol. 2021. Establishing a multidisciplinary initiative for interoperable electronic health record innovations at an academic medical center. *JAMIA Open* 4(3). <https://doi.org/10.1093/jamiaopen/ooab041>.
62. Kawamoto, K. 2005. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 330(7494):765. <https://doi.org/10.1136/bmj.38398.500764.8F>.
63. Kellogg, K. C., M. Sendak, and S. Balu, 2022. AI on the Frontlines. *MIT Sloan Management Review*. Available at: <https://sloanreview.mit.edu/article/ai-on-the-front-lines/> (accessed May 11, 2022).
64. Kensaku, K., P. Kukhareva, C. Weir, M. Flynn, C. Nanjo, D. Martin, P. B. Warner, D. E. Shields, S. Rodriguez-Loya, R. L. Bradshaw, R. C. Cornia, T. J. Reese, H. S. Kramer, T. Taft, R. L. Curran, K. L. Morgan, D. Borbolla, M. Hightower, W. J. Turnbull, M. B. Strong, W. W. Chapman, T. Gregory, C. H. Stipelman, J. H. Shakib, R. Hess, J. P. Boltax, J. P. Habboushe, F. Sakaguchi, K. M. Turner, S. P. Narus, S. Tarumi, W. Takeuchi, H. Ban, D. W. Wetter, C. Lam, T. J. Caverly, A. Fagerlin, C. Norlin, D. C. Malone, K. A. Kaphingst, W. K. Kohlmann, B. S. Brooke, and G. Del Fiol. 2021. Establishing a multidisciplinary initiative for interoperable electronic health record innovations at an academic medical center. *JAMIA Open* 4(3). <https://doi.org/10.1093/jamiaopen/ooab041>.
65. Khalifa, A., C. Mason, H. Garvin, M. Williams, G. Del Fiol, B. Jackson, S. Bleyl, G. Alterovitz, and S. Huff. 2021. Interoperable Genetic Lab Test Reports: Mapping Key Data Elements to HL7 FHIR Specifications and Professional Reporting Guidelines. *Journal of the American Medical Informatics Association* 28(12):2617–25. <https://doi.org/10.1093/jamia/ocab201>.
66. Khan, N. S., M. S. Ghani, and G. Anjum. 2021. ADAM-sense: Anxiety-displaying activities recognition by motion sensors. *Pervasive and Mobile Computing* 78(21). <https://doi.org/10.1016/j.pmcj.2021.101485>.
67. Krueger, L. 2022. *Clinical decision-making bias in darker skin types: a prospective survey study identifying diagnostic bias in decision to biopsy*. Abstract presented at 18th Skin of Color Society Scientific Symposium, March 24, 2022. Boston, MA.
68. Lankton, N. K., D. H. McKnight, and J. Tripp. 2015. Technology, Humanness, and Trust: Rethinking Trust in Technology. *Journal of the Association for Information Systems* 16(10):880-918. DOI: 10.17705/1jais.00411.

69. Lee, P., A. Abernethy, D. Shaywitz, A. V. Gundlapalli, J. Weinstein, P. M. Doraiswamy, K. Schulman, and S. Madhavan. 2022. Digital Health COVID-19 Impact Assessment: Lessons Learned and Compelling Needs. *NAM Perspectives*. Discussion Paper, National Academy of Medicine, Washington, DC. <https://doi.org/10.31478/202201c>.
70. Lee, Y., Y. S. Kim, D.-I. Lee, S. Jeong, G.-H. Kang, Y. S. Jang, W. Kim, H. Y. Choi, J. G. Kim, and S.-H. Choi. 2022. The application of a deep learning system developed to reduce the time for RT-PCR in COVID-19 detection. *Science Reports* 12(1234). <https://doi.org/10.1038/s41598-022-05069-2>.
71. Leslie, D., A. Mazumder, A. Peppin, M. K. Wolters, and A. Hagerty. 2021. Does “AI” stand for augmenting inequality in the era of covid-19 healthcare? *BMJ* 372:1-5. <https://doi.org/10.1136/bmj.n304>.
72. Lin, D., T. Nazreen, T. Rutowski, Y. Lu, A. Harati, E. Shriberg, P. Chlebek, and M. Aratow. 2022. Feasibility of a Machine Learning-Based Smartphone Application in Detecting Depression and Anxiety in a Generally Senior Population. *Frontiers in Psychology* 13. <https://doi.org/10.3389/fpsyg.2022.811517>.
73. Luzniak, K. 2021. “What’s the cost of artificial intelligence in healthcare?” *Neoteric*, December 16. Available at: <https://neoteric.eu/blog/whats-the-cost-of-artificial-intelligence-in-healthcare/> (accessed May 9, 2022).
74. Mäkelä, K., M. I. Mäyränpää, H. K. Sihvo, P. Bergman, E. Sutinen, H. Ollila, R. Kaarteenaho, and M. Myllärniemi. 2021. Artificial intelligence identifies inflammation and confirms fibroblast foci as prognostic tissue biomarkers in idiopathic pulmonary fibrosis. *Human Pathology* (107):58-68. <https://doi.org/10.1016/j.humpath.2020.10.008>.
75. Maliha G., S. Gerke, I. G. Cohen, and R. B. Parikh. 2021. Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation. *Milbank Quarterly* 99(3):629-647. <https://doi.org/10.1111/1468-0009.12504>.
76. Mandel, J. C., D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni. 2016. SMART on FHIR: A Standards-Based, Interoperable Apps Platform for Electronic Health Records. *Journal of the American Medical Informatics Association* 23(5):899–908. <https://doi.org/10.1093/jamia/ocv189>.
77. Mandl K.D., and F. T. Bourgeois. 2017. The Evolution of Patient Diagnosis: From Art to Digital Data-Driven Science. *Journal of American Medical Association* 318(19):1859–1860. <https://doi.org/10.1001/jama.2017.15028>.
78. Mandl, K. D., and I. S. Kohane. 2012. Escaping the EHR Trap - The Future of Health IT. *New England Journal of Medicine* 366(24):2240-2242. <https://doi.org/10.1056/NEJMp1203102>.
79. Mandl, K. D., and I. S. Kohane. 2017. A 21st-Century Health IT System - Creating a Real-World Information Economy. *New England Journal of Medicine* 376(20):1905-1907. <https://doi.org/10.1056/NEJMp1700235>.

80. Mandl, K. D., J. C. Mandel, S. N. Murphy, E. V. Bernstam, R. L. Ramoni, D. A. Kreda, J. M. McCoy, B. Adida, and I. S. Kohane. 2012. The SMART Platform: Early Experience Enabling Substitutable Applications for Electronic Health Records. *Journal of the American Medical Informatics Association* 19(4):597-603. <https://doi.org/10.1136/amiajnl-2011-000622>.
81. Marmar, C. R., A. D. Brown, M. Qian, E. Laska, C. Siegel, M. Li, D. Abu-Amara, A. Tsiartas, C. Richey, J. Smith, B. Knoth, and D. Vergyri. 2019. Speech-based markers for posttraumatic stress disorder in US veterans. *Depression and Anxiety* (36):607-616. <https://doi.org/10.1002/da.22890>.
82. Matheny, M., S. Thadaney Israni, M. Ahmed, and D. Whicher, Editors. 2019. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. NAM Special Publication, National Academy of Medicine, Washington, DC.
83. Melnick, E. R., L. N. Dyrbye, C. A. Sinsky, M. Trockel, C. P. West, L. Nedelec, M. A. Tutty, and T. Shanafelt. 2020. The association between perceived electronic health record usability and professional burnout among US physicians. *Mayo Clinic Proceedings* 95(3):476-487. <https://doi.org/10.1016/j.mayocp.2019.09.024>.
84. Miller, A. R. 1994. Medical Diagnostic Decision Support Systems – Past, Present, and Future: A Threaded Bibliography and Brief Commentary. *Journal of Medical Informatics* 1(1):8-27. <https://doi.org/10.1136/jamia.1994.95236141>.
85. Miller, A. R., and A. Geissbuhler. 2007. Diagnostic Decision Support Systems. *Clinical Decision Support Systems: Theory and Practice*. 2nd edition, edited by K. J. Hannah and M. J. Ball. New York, NY: Springer Science.
86. Nakahara, H. K. Namba, A. Fukami, R. Watanabe, M. Mizutani, T. Matsu, S. Nishimura, S. Jinnouchi, S. Nagamachi, T. Ohnishi, S. Futami, L. G. Flores, M. Nakahara, and S. Tamura. 1998. Computer-Aided Diagnosis (CAD) for Mammography: Preliminary Results. *Breast Cancer* 5:401-405. <https://doi.org/10.1007/BF02967438>.
87. North Carolina State Health Plan and John Hopkins Bloomberg School of Public Health. 2021. *North Carolina Hospitals: Charity Care Case Report*. Available at: <https://s3.documentcloud.org/documents/21094171/download-1.pdf> (accessed May 11, 2022).
88. Office of the National Coordinator for Health Information Technology (ONC). 2018. *Clinical Decision Support*. Available at: <https://www.healthit.gov/topic/safety/clinical-decision-support> (accessed September 14, 2022).
89. Ommaya, A. K., P. F. Cipriano, D. B. Hoyt, K. A. Horvath, P. Tang, H. L. Paz, M. S. DeFrancesco, S. T. Hingle, S. Butler, and C. A. Sinsky. 2018. Care-Centered Clinical Documentation in the Digital Environment: Solutions to Alleviate Burnout. *NAM Perspectives*. Discussion Paper, National Academy of Medicine, Washington, DC. <https://doi.org/10.31478/201801c>.

90. Parakh, A., H. Lee, J. H. Lee, B. H. Eisiner, D. V. Sahani, and S. Do. 2019. Urinary stone detection on CT images using deep convolutional neural networks: Evaluation of model performance and generalization. *Radiology: Artificial Intelligence* 1(4). <https://doi.org/10.1148/ryai.2019180066>.
91. Parikh, R. B., and L. A. Helmchen. 2022. Paying for artificial intelligence in medicine. *npj Digital Medicine* 5(63):1-5. <https://doi.org/10.1038/s41746-022-00609-6>.
92. Price, W. N., S. Gerke, and I. G. Cohen. 2019. Potential liability for physicians using artificial intelligence. *JAMA* 322(18):1765. <https://doi.org/10.1001/jama.2019.15064>.
93. Rajkomar, A., M. Hardt, M. Howell, G. Corrado, and M. Chin. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine* 169:866-872. <https://doi.org/10.7326/M18-1990>.
94. Ramsetty, A., and C. Adams. 2020. Impact of the digital divide in the age of COVID-19. *Journal of American Medical Informatics Association* 27(7):1147-1148. <https://doi.org/10.1093/jamia/ocaa078>
95. Ray, A., A. Gupta, and A. Al. 2020. Skin Lesion Classification with Deep Convolutional Neural Network: Process Development and Validation. *Journal of Medical Internet Research Dermatology* (1):e18438. <https://doi.org/10.2196/18438>.
96. Reisman, M. 2017. EHRs: The Challenge of Making Electronic Data Usable and Interoperable. *P&T* 42(9):572-575. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5565131/> (accessed May 23, 2022).
97. Responsible Innovation Labs. 2022. *Charter*. Available at: <https://www.rilabs.org/charter> (accessed on June 30, 2022).
98. Ridgely, M. S., and M. D. Greenberg. 2012. Too many alerts, too much liability: sorting through the malpractice implications of drug-drug interaction clinical support. *Saint Louis University Journal of Health Law & Policy* 5(2):257-295. Available at: <https://scholarship.law.slu.edu/jhlp/vol5/iss2/4> (accessed July 27, 2022).
99. Rodin, J., and S. Madsbjerg. 2021. *Making Money Moral : How a New Wave of Visionaries Is Linking Purpose and Profit*. Wharton School Press.
100. Sandhu, S., A. L. Lin, N. Brajer, J. Sperling, W. Ratliff, A. D. Bedoya, S. Balu, C. O'Brien, and M. P. Sendak. 2020. Integrating a Machine Learning System into Clinical Workflows: Qualitative Study. *JMIR* 22(11). <https://doi.org/10.2196/22421>.
101. Sanyal, S. 2021. How much does artificial intelligence cost in 2021? *Analytics Insights*. Available at: <https://www.analyticsinsight.net/how-much-does-artificial-intelligence-cost-in-2021/> (accessed May 11, 2022).

102. Sendak, M. P., J. D'Arcy, S. Kashyap, M. Gao, M. Nichols, K. Corey, W. Ratliff, and S. Balu. 2020a. A Path for Translation of Machine Learning Products into Healthcare Delivery. *European Medical Journal Innovations*. <https://doi.org/10.33590/emjinnov/19-00172>.
103. Sendak, M. P., W. Ratliff, D. Sarro, E. Alderton, J. Futoma, M. Gao, M. Nichols, M. Revoir, F. Yashar, C. Miller, K. Kester, S. Sandhu, K. Corey, N. Brajer, C. Tan, A. Lin, T. Brown, S. Engelbosch, K. Anstrom, M. C. Elish, K. Heller, R. Donohoe, J. Theiling, E. Poon, S. Balu, A. Bedoya, and C. O'Brien. 2020b. Real-World Integration of a Sepsis Deep Learning Technology into Routine Clinical Care: Implementation Study. *JMIR Medical Informatics* 8(7): e15182. <https://doi.org/10.2196/15182>.
104. Sendak, M. P., M. Gao, N. Brajer, and S. Balu. 2020c. Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine* 3(4). <https://doi.org/10.1038/s41746-020-0253-3>.
105. Sendak, M. P., S. Balu, and K. A. Schulman. 2017. Barriers to Achieving Economies of Scale in Analysis of EHR Data: A Cautionary Tale. *Applied Clinical Informatics* 8(3):826-831. <https://doi.org/10.4338/ACI-2017-03-CR-0046>.
106. Shen, Y., F. E. Shamout, J. R. Oliver, J. Witowski, K. Kannan, J. Park, N. Wu, C. Huddleston, S. Wolfson, A. Millet, R. Ehrenpreis, D. Awal, C. Tyma, N. Samreen, Y. Gao, C. Chhor, S. Gandhi, C. Lee, S. Kumari-Subaiya, C. Leonard, R. Mohammed, C. Moczuski, J. Altabet, J. Babb, A. Lewin, B. Reig, L. Moy, L. Heacock, and K. J. Geras. 2021. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nature Communication* 12(5645). <https://doi.org/10.1038/s41467-021-26023-2>.
107. Signaevsky, M., B. Marami, M. Prastawa, N. Tabish, M. A. Iida, X. F. Zhang, M. Sawyer, I. Duran, D. G. Koenigsberg, C. H. Bryce, L. M. Chahine, B. Mollenhauer, S. Mosovsky, L. Riley, K. D. Dave, J. Eberling, C. S. Coffey, C. H. Adler, G. E. Serrano, C. L. White III, J. Koll, G. Fernandez, J. Zeineh, C. Cordon-Cardo, T. G. Beach, and J. F. Cray. 2022. Antemortem detection of Parkinson's disease pathology in peripheral biopsies using artificial intelligence. *Acta Neuropathological Communications* 10(21). <https://doi.org/10.1186/s40478-022-01318-7>.
108. Singh, R. P., G. L. Hom, M. D. Abramoff, J. P. Campbell, and M. F. Chiang. 2020. Current Challenges and Barriers to Real-World Artificial Intelligence Adoption for the Healthcare System, Provider, and the Patient. *Translational Vision Science & Technology* 9(2):1-6. <https://doi.org/10.1167/tvst.9.2.45>.
109. SPOT: How HCA Healthcare is "sniffing out" sepsis early. 2018. *HCA Healthcare Today*. Available at: <https://hcahealthcaretoday.com/2018/09/10/spot-how-hca-is-sniffing-out-sepsis-early/>. (accessed May 1, 2022).

110. Sutton, R. T., D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine* 3(17). <https://doi.org/10.1038/s41746-020-0221-y>.
111. Syrowatka, A., M. Kuznetsova, A. Alsubai, A. L. Beckman, P. A. Bain, K. J. Thomas Craig, J. Hu, G. P. Jackson, K. Rhee, and D. W. Bates. 2021. Leveraging artificial intelligence for pandemic preparedness and response: a scoping review to identify key use cases. *npj Digital Medicine* 4(96). <https://doi.org/10.1038/s41746-021-00459-8>.
112. Tadavarthi, Y. B. V., E. Krupinski, A. Prater, J. Gichoya, N. Safdar, and H. Trivedi. 2020. The State of Radiology AI: Considerations for Purchase Decisions and Current Market Offerings. *Radiology: Artificial Intelligence* 2(6). <https://doi.org/10.1148/ryai.2020200004>.
113. Unsworth, H., V. Wolfram, B. Dillon, M. Salmon, F. Greaves, X. Liu, T. MacDonald, A. K. Denniston, V. Sounderajah, H. Ashrafian, A. Darzi, C. Ashurst, C. Holmes, and A. Weller. 2022. Building an evidence standards framework for artificial intelligence-enabled digital health technologies. *The Lancet Digital Health* 4(4):e216-e217. [https://doi.org/10.1016/S2589-7500\(22\)00030-9](https://doi.org/10.1016/S2589-7500(22)00030-9).
114. Vinson, D. R., S. D. Casey, P. L. Vuong, J. Huang, D. W. Ballard, and M. E. Reed. 2022. Sustainability of a Clinical Decision Support Intervention for Outpatient Care for Emergency Department Patients With Acute Pulmonary Embolism. *JAMA Netw Open* 5(5):e2212340. doi:10.1001/jamanetworkopen.2022.12340.
115. Walker, H. K. 1990. The Origins of the History and Physical Examination, *Clinical Methods: The History, Physical, and Laboratory Examinations*. 3rd edition, edited by W.D. Hall and J.W. Hurst. Boston, MA.
116. Wiens, J., S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P. N. Ossorio, S. Thadaney Israni, and A. Goldenberg. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine* 25(9):1337-1340. <https://doi.org/10.1038/s41591-019-0548-6>.
117. Wolff, J., J. Pauling, A. Keck, and J. Baumbach. 2020. The Economic Impact of Artificial Intelligence in Health Care: Systematic Review. *Journal of Medical Internet Research* 22(2):e16866. <https://doi.org/10.2196/16866>.
118. Wong, A., E. Otlis, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penzoza, M. Ghous, and K. Singh. 2021. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*. 181(8):1065-1070. <https://doi.org/10.1001/jamainternmed.2021.2626>.

119. Wu, A. C., C. Graif, S. G. Mitchell, J. Meurer, and K. D. Mandl. 2021. Creative Approaches for Assessing Long-term Outcomes in Children. *Pediatrics* 148(Suppl 1):s25-s32. <https://doi.org/10.1542/peds.2021-050693F>.
120. Wu, T. 2011. Agency Threats. *Duke Law Journal* 60(8):1841-1857.
121. Wynants, L., B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. J. Bonten, D. L. Dahly, J. A. Damen, T. P. A. Debray, V. M. T. de Jong, M. De Vos, P. Dhiman, M. C. Haller, M. O. Harhay, L. Henckaerts, P. Heus, M. Kammer, N. Kreuzberger, A. Lohmann, K. Luijken, J. Ma, G. P. Martin, D. J. McLernon, C. L. Andaur Navarro, J. B. Reitsma, J. C. Sergeant, C. Shi, N. Skoetz, L. J. M. Smits, K. I. E. Snell, M. Sperrin, R. Spijker, E. W. Steyerberg, T. Takada, I. Tzoulaki, S. M. J. van Kuijk, B. C. T. van Bussel, I. C. C. van der Horst, F. S. van Royen, J. Y. Verbakel, C. Wallisch, J. Wilkinson, R. Wolff, L. Hooft, K. G. M. Moons, and M. van Smeden. 2020. Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. *BMJ* 269:m1328. <https://doi.org/10.1136/bmj.m1328>.
122. Yala, A., C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay. 2019. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology* 292(1):60-66. <https://doi.org/10.1148/radiol.2019182716>.
123. Yang, Z., C. Silcox, M. Sendak, S. Rose, D. Rehkopf, R. Phillips, L. Peterson, M. Marino, J. Maier, S. Lin, W. Liaw, I. A. Kakadiaris, J. Heintzman, I. Chu, and A. Bazemore. 2022. Advancing primary care with Artificial Intelligence and Machine Learning. *Healthcare (Amsterdam, Netherlands)* 10(1). <https://doi.org/10.1016/j.hjdsi.2021.100594>.
124. Yu, K., A. L. Beam, and I. S. Kohane. 2018. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2:719–731. <https://doi.org/10.1038/s41551-018-0305-z>.
125. Zou, J. and L. Schiebinger. 2021. Ensuring that biomedical AI benefits diverse populations. *eBioMedicine* 67:1-6. <https://doi.org/10.1016/j.ebiom.2021.103358>.

DOI

<https://doi.org/10.31478/202209c>

Suggested Citation

Adler-Milstein, J., N. Aggarwal, M. Ahmed, J. Castner, B. Evans, A. Gonzalez, C. A., James, S. Lin, K. Mandl, M. Matheny, M. Sendak, C. Shachar, and A. Williams. 2022. Meeting the Moment: Reducing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis. NAM Perspectives. Discussion Paper, Washington, DC.

<https://doi.org/10.31478/202209c>.

Author Information

Julia Adler-Milstein, PhD, is Professor of Medicine and Director of the Center for Clinical Informatics and Improvement Research (CLIR) at the University of California-San Francisco. **Nakul Aggarwal, BS**, is an MD-PhD candidate at the University of Wisconsin-Madison. **Mahnoor Ahmed, MEng**, is an Associate Program Officer at the National Academy of Medicine. **Jessica Castner, PhD, RN-BC**, is the 2021-2022 National Academy of Medicine Nurse Scholar-in-Residence, President of Castner Incorporated, and Editor-in-Chief of the Journal of Emergency Nursing. **Barbara J. Evans, PhD, JD**, is Professor of Law and Stephen C. O'Connell Chair at the University of Florida. **Andrew A. Gonzalez, MD, JD, MPH**, is Associate Director for Data Science and Research Scientist at Regenstrief Institute. **Cornelius A. James, MD**, is a Clinical Assistant Professor in the Departments of Internal Medicine, Pediatrics, and Learning Health Sciences at the University of Michigan.

Steven Lin, MD, is Clinical Associate Professor of Medicine at Stanford University. **Kenneth D. Mandl, MD, MPH**, is Director of the Computational Health Informatics Program (CHIP) at Boston Children's Hospital. **Michael E. Matheny, MD, MS, MPH**, is Co-Director of the Center for Improving the Public's Health through Informatics at Vanderbilt University. **Mark P. Sendak, MD, MPP**, is Population Health and Data Science Lead at the Duke Institute for Health Innovation at Duke University. **Carmel Shachar, JD, MPH**, is Executive Director of the Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School. **Asia Williams, MPH**, is an Associate Program Officer at the National Academy of Medicine.

Acknowledgments

This paper benefitted from the insights of **Matthew Diamond**, U.S. Food and Drug Administration; **Maryellen Giger**, University of Chicago; **Brian Gurbaxani**, Centers for Disease Control and Prevention; and **Christina Silcox**, Duke University.

Sections of the paper were developed based on the thoughtful input of **Clifford Goodman, PhD**, Lewin Group; **Vivian Lee, MD, PhD, MBA**, Verily; and **Suzanne Tamang, PhD**, Stanford University and Veterans Affairs.

Conflict-of-Interest Disclosures

Jessica Castner discloses receiving grants and fees from the National Institutes of

Health, fees from the Emergency Nurses Association, and serving as co-chair of the American Thoracic Society's Health Policy Committee on Terrorism and Inhalation Disasters section. Barbara Evans discloses receiving grants from the National Institutes of Health. Steven Lin discloses serving as VP of Health Sciences for Codex Health, where he is a paid consultant; and receiving grants administered through Stanford University from Amazon, American Academy of Family Physicians, American Board of Family Medicine, Center for Professionalism and Value in Health Care, DeepScribe, Google Health, Omada Health, Predicta Med, Quadrant Technologies, Soap Health, Society of Teachers of Family Medicine, UCSF, and Verily. Kenneth Mandl discloses that his laboratory receives sponsored research funding from Quest Diagnostics; and that Boston Children's Hospital receives corporate philanthropic support for his laboratory from SMART Advisory Committee members, which include the American Medical Association, BMJ Group, Eli Lilly and Company, Google Cloud, Hospital Corporation of America, Microsoft, Optum, Cambia Health Solutions, Quest Diagnostics, and Humana. Mark Sendak discloses that he is co-inventor of technology licensed from Duke University to Cohere Med, Inc and Clinetic, Inc.; and that he holds equity in Clinetic, Inc. Carmel Shachar discloses that she is a member of Advarra's Institutional Research Board.

Andrew Gonzalez is the current NAM Gilbert S. Omenn fellow; Steven Lin is the current NAM James C. Puffer, M.D./American Board of Family Medicine fellow; and Julia Adler-Milstein and Kenneth D. Mandl are members of the National Academy of Medicine.

Correspondence

Questions or comments about this paper should be directed to namedicine@nas.edu.

Disclaimer

The views expressed in this paper are those of the authors and not necessarily of the authors' organizations, the National Academy of Medicine (NAM), or the National Academies of Sciences, Engineering, and Medicine (the National Academies). The paper is intended to help inform and stimulate discussion. It is not a report of the NAM or the National Academies. Copyright by the National Academy of Sciences. All rights reserved.

Appendix I: Objectives, Scope, and Methodology

We describe our scope and methodology for addressing the four objectives outlined below:

Objectives

1. What machine learning (ML) technologies are currently available for the medical diagnosis of diseases such as cancer and heart disease in the U.S.?
2. What ML technologies are emerging for medical diagnosis of diseases such as cancer and heart disease?
3. What challenges affect the development or use of ML technologies for medical diagnosis?
4. What policy options could help address these challenges, and what are the potential opportunities and considerations?

Scope and methodology

To address all four research objectives, we assessed available and emerging ML technologies for medical diagnosis as well as the benefits and challenges associated with their use. To do so, we reviewed reports and scientific literature describing current and emerging ML technologies; interviewed a variety of stakeholders, including agency officials, industry members, and academic researchers; and conducted an expert meeting in conjunction with the National Academy of Medicine.

Limitations to scope

We focused our review on five select diseases, and the available and emerging ML technologies designed to render a diagnosis or directly support a medical professional's diagnosis for these diseases. We excluded AI methods using expert or rules-based systems, and focused on AI methods relying on statistical learning using observed or simulated data. ML techniques discussed are examples and not an exhaustive list of all ML techniques available, or in development, for medical diagnosis purposes.

Literature Search

In the course of our work we conducted two literature searches. To establish background and identify appropriate technologies, we reviewed articles from scientific literature. Our second search targeted survey and review articles on machine learning using the same search terms adding the terms "review" and "meta-analysis". We filtered the search by journals with the highest count of studies, reviewed the abstracts, and selected the most relevant articles for further review based on our objectives. Another source of literature we reviewed were recommendations from interviewees and the National Academies. Throughout our work, we monitored literature to identify new articles appropriate to addressing our objectives.

Interviews

We interviewed key stakeholders in the field of ML medical diagnostic technologies, including:

- relevant federal agencies including the Federal Trade Commission, the Department of Energy, the Department of Veterans Affairs including two Veterans Affairs Medical Centers, the National Institute of Health, and the Food and Drug Administration;
- seven private companies focused on developing machine learning based medical diagnostics and three industry/professional organizations; and
- five academic researchers.

Because this is a small and non-generalizable sample of the stakeholders involved in using ML medical diagnostic technologies, the results of our interviews are illustrative and represent important perspectives, but are not generalizable.

Expert Meeting

We collaborated with the National Academy of Medicine to convene a meeting of 16 experts over three days on available and emerging ML technologies for medical diagnostics. We worked with National Academy of Medicine staff to identify experts from a range of stakeholder groups including federal agencies, academia, industry, and legal scholars, with expertise covering all significant areas of our review, including individuals with research or operational expertise in using ML technologies for medical diagnosis.⁶² We evaluated the

experts for any conflicts of interest. A conflict of interest was considered to be any current financial or other interest (such as an organizational position) that might conflict with the service of an individual because it could (1) impair objectivity or (2) create an unfair competitive advantage for any person or organization. The 16 experts were determined to be free of reported conflicts of interest, except those that were outside the scope of the forum or where the overall design of our panel and methodology was sufficient to address them, and the group as a whole was determined to not have any inappropriate biases.⁶³ (See Appendix II for a list of these experts and their affiliations).

We divided the meeting into six moderated discussion sessions: (1) existing AI/ML tools and technologies in medical diagnostics; (2) emerging AI/ML tools and technologies in medical diagnostics; (3) challenges to development; (4) challenges to adoption; and (5) policy options to address challenges to development; (6) policy options to address challenges to adoption. Each session featured an open discussion among all meeting participants based on key questions we provided. The meeting was transcribed to ensure that we accurately captured the experts' statements. After the meeting, we reviewed the transcripts to characterize their responses and to inform our understanding. Following the meeting, we continued to seek the experts' advice to clarify and expand on what we had heard. We provided our draft

⁶²This meeting of experts was planned and convened with the assistance of the National Academy of Medicine to better ensure that a breadth of expertise was brought to bear in its preparation, however all final decisions regarding meeting substance and expert participation were the responsibility of GAO. Any conclusions and recommendations in GAO reports are solely those of GAO.

⁶³For example, one expert had equity interest in companies developing ML technologies for medical diagnostics. We determined the expert's relationship did not prevent them from serving on the panel, as the discussion was not planned to revolve around any specific technology or vested interest.

report to the experts for their technical review, consistent with previous technology assessment methodologies.

Policy Options

We intend policy options to provide policymakers with a broader base of information for decision-making.⁶⁴ The options are neither recommendations to federal agencies nor matters for congressional consideration. They are also not listed in any specific rank or order. We are not suggesting that they be done individually or combined in any particular fashion. Additionally, we did not conduct work to assess how effective the options may be, and express no view regarding the extent to which legal changes would be needed to implement them.

We present three policy options in response to the challenges identified during our work and discuss potential opportunities and considerations of each. While we present options to address the major factors we

identified, the options are not inclusive of all potential policy options. The policy options and analyses were supported by the above evidence. Policy ideas, identified from the evidence above, were: (1) adapted into policy options by combining similar ideas that were duplicative, (2) grouped into a higher-level policy option, (3) examples of how to implement a policy option, or (4) did not fit into our scope.

We conducted our work from November 2020 through September 2022 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement to obtain sufficient and appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

⁶⁴ Policymakers is a broad term including, for example, Congress, federal agencies, state and local governments, academic and research intuitions, and industry.

Appendix II: Expert Participation

We collaborated with the National Academies of Science, Engineering, and Medicine to convene a meeting of experts over three days to inform our work on artificial intelligence, particularly machine learning, in medical diagnostics. The meeting was held virtually on June 2, 3, and 8, 2021. Experts who participated in this meeting are listed below. We corresponded with experts for additional assistance throughout our work. We provided our draft report to the experts for their technical review, consistent with previous technology assessment methodologies.

Hugo Aerts

Director of Artificial Intelligence in Medicine
Program
Mass General Brigham

Eric Horvitz

Chief Scientific Officer
Microsoft

Pat Baird

Sr. Regulatory Specialist
Philips

Constance Lehman

Chief of Breast Imaging
Massachusetts General Hospital

Barbara Evans

Professor of Law and Engineering
University of Florida

Mia Levy

Director of Cancer Center
Rush University

Richard Frank

Chief Medical Officer
Siemens Heathineers

Anil Parwani

Professor of Pathology; Vice Chair and
Director of Anatomical Pathology
Ohio State University

Maryellen Giger

A.N. Pritzker Distinguished Service Professor
of Radiology
University of Chicago

Lily Peng

Physician-Scientist and Product Manager
Google

Kenneth Goodman

Director, Institute for Bioethics and Health
Policy
University of Miami

Bruce Pyenson

Principal and Actuary
Milliman

John Halamka

President
Mayo Clinic Platform

Berkman Sahiner

Senior Biomedical Research Scientist and
Biomedical Product Assessment Service
Expert
Center for Devices and Radiological Health,
U.S. Food and Drug Administration

Robert Sparrow

Professor of Philosophy
Monash University

Eric Topol

Professor of Molecular Medicine and
Executive Vice-President
Scripps Research
Founder and Director
Scripps Research Translational Institute

Appendix III: GAO Contacts and Staff Acknowledgments

GAO contact

Karen L. Howard, PhD, (202) 512-6888 or howardk@gao.gov

Staff acknowledgments

In addition to the contacts named above, Hayden Huang (Assistant Director), Matthew Hunter (Analyst-in-Charge), Jenny Chanley, Jehan Chase, Jonathan Felbinger, Nathan Hanks, Eric Larson, Eric Lee, Anika McMillon, Yesook Merrill, Ben Shouse, and Michael Steinberg made key contributions to this report. George Bogart, Robert Copeland, Leia Dickerson, Kaitlin Farquharson, Charlotte Hinkle, Neelaxi Lakhmani, Monica Perez-Nelson, Britney Tsao, Walter Vance, and Wesley Wilhem also contributed to this report.

(104629)

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's website (<https://www.gao.gov>). Each weekday afternoon, GAO posts on its website newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to <https://www.gao.gov> and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, <https://www.gao.gov/ordering.htm>.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

Connect with GAO

Connect with GAO on [Facebook](#), [Flickr](#), [Twitter](#), and [YouTube](#).

Subscribe to our [RSS Feeds](#) or [E-mail Updates](#).

Listen to our [Podcasts](#) and read [The Watchblog](#).

Visit GAO on the web at <https://www.gao.gov>.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact: Website: <https://www.gao.gov/fraudnet/fraudnet.htm>

Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

A. Nicole Clowers, Managing Director, ClowersA@gao.gov, (202) 512-4400,
U.S. Government Accountability Office, 441 G Street NW, Room 7125, Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, YoungC1@gao.gov, (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149, Washington, DC 20548

Strategic Planning and External Liaison

Stephen Sanford, Managing Director, spel@gao.gov, (202) 512-9715
U.S. Government Accountability Office, 441 G Street NW, Room 7B37N, Washington, DC 20548