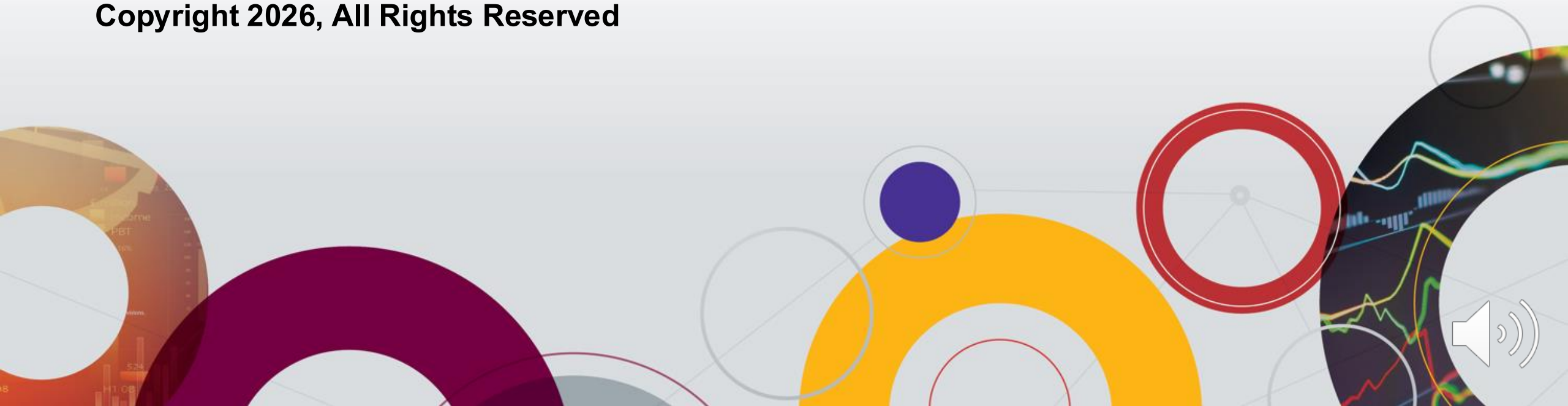


5.5b Trustworthy AI

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Trustworthy AI

Methodological concerns

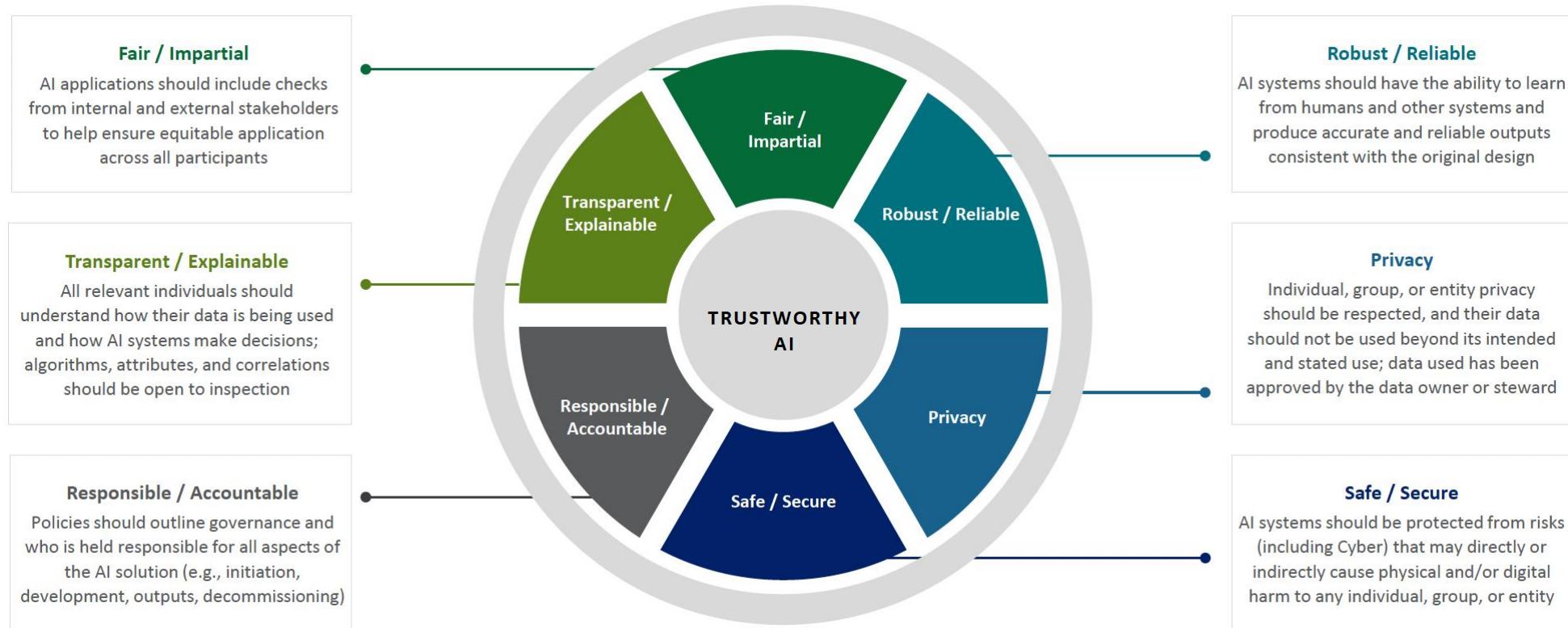
Critical appraisal

Trustworthy AI

Education and training

Toward the future

Issues and approaches to ensure trustworthy AI (Ammanath, 2022)



HHS Trustworthy AI Playbook (2021)

Trust-related issues

Data ownership

- Who owns patient photographs? (Davis, 2020; Davis, 2022)
- De-identified data can be re-identified with other data; who should own resulting data? (Ross, 2022)
- Data should be “democratized” (Allen, 2019), especially for disabled (Newman-Griffis, 2022) and historically underserved groups (Gottschalk, 2022; Hossain, 2023)

Defining, measuring, and mitigating bias – Algorithmic Bias Playbook (Obermeyer, 2021)

Use of contrastive learning to reduce bias

- Performance disparities across demographic groups in pathology AI systems (Lin, 2025)
- Bias in patient matching for clinical trials and medical question-answering for clinical decision support (Ji, 2025)

Safety and certification

Ensure safety of AI in real-world clinical care (Sittig, 2024)

Clinical Laboratory Improvement Amendments (CLIA) as a model (Jackson, 2024)

Clinical certification pathway for generalist medical AI systems (Rajpurkar, 2025)

Red teaming (Chang, 2025)

Full clinical certification

Autonomous operation with continuous recertification cycles

Supervised clinical practice

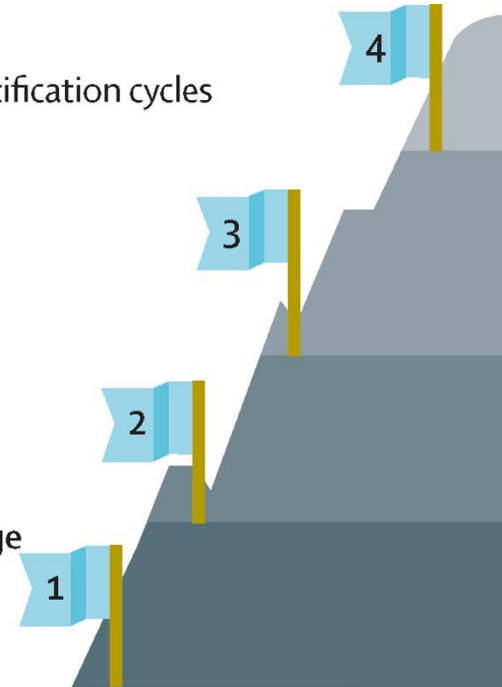
Safe integration into health-care teams under supervision

Specialty task performance

Demonstrating reliable specialty-specific capabilities with strict monitoring

Foundation stage: master medical knowledge

Baseline competency through standardised testing and scenario analysis



Pursuing equity at each AI development life cycle stage (Johnson, 2025)

Model Development

Assemble a multiparty stakeholder team with diverse representations and backgrounds.

Understand available AI technologies and training data for building and tuning models.

Converge on working definitions of fairness and equity of use of each AI-based intervention.

Perform AI capabilities and needs analysis, with explicit multistakeholder attention to implications for equity.

Develop model and review measures of model performance on representative workloads.

Document and address biases and gaps in data and performance with fairness tools.

Revisit key questions on equity in the context of demonstrated model performance.

Health System Integration

Seek input from local multistakeholder team on workflow vision leveraging integration of model into local workflow.

Ensure transparency and disclosure about operating characteristics of end-to-end system, potential biases, recommended uses, caveats on risk to equity.

Engage in limited testing of prototype with end-user representatives on realistic scenarios.

Iteratively refine system based on feedback from stakeholders.

Revisit key questions on equity in context of prototype end-to-end system.

Model Deployment

Engage in phase of pragmatic trials with reporting to multistakeholder team.

Iteratively refine system, testing new variants with trials.

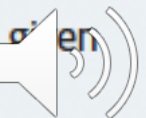
Log usage for analytics.

Provide clinician and patient feedback tools and opportunities to collect qualitative input on AI capabilities in practice.

Monitor over time for drift of capabilities with changing frequency and nature of usage, demographics of population, short- and longer-term impact on population.

Revisit key questions on equity in context of prototype end-to-end system.

Periodically revisit key questions on equity, open logs and experiences.



Accountability

Even when AI employed fairly and transparently, must also address bias in hands of “imperfect users” (Kostick-Quenet, 2022) and structural challenges in healthcare (London, 2022)

Concerns over control by tech companies; models should be open-source (Toma, 2023)

Models should be validated locally and recurrently (Youssef, 2023)

Safe harbors for red-teaming and evaluation (Longpre, 2024)

Priorities – National Academy of Medicine (NAM)

Priorities (Matheny, 2025)

- Ensuring safe, effective, and trustworthy use of AI
- Promotion and development of an AI-competent healthcare workforce
- Investing in AI research to support the science, practice, and delivery of health and healthcare
- Promotion of policies and procedures to clarify AI liability and responsibilities

Code of Conduct (Adams, 2025)

- Advance humanity
- Ensure equity
- Engage impacted individuals
- Improve workforce well-being
- Monitor performance
- Innovate and learn

Coalition for Health AI (CHAI) – blueprint for healthy AI

Objectives

Building trust

- Ensuring that AI in healthcare is reliable and widely accepted and understood by patients, clinicians, and organizations.

Broad representation

- Engaging diverse stakeholders, including underserved communities, to ensure AI tools are developed and tested across various environments and data sets.

Public/private collaboration

- Working with both the public and private sectors, leveraging the strengths of each to develop effective AI standards and tools.

Transparency

- Promoting transparency in AI development and implementation, addressing concerns about the “black box” nature of AI decision-making.

Supporting innovation

- Balancing the need for regulation with the need to foster innovation by developing consensus-driven standards to ultimately increase adoption of safe and effective AI.

Deliverables

Certification of Assurance Labs

- Supporting the development and certification of a nationwide network of quality assurance labs to evaluate and validate AI models using community best practices and standards.

Continuous evaluation

- Providing tools and frameworks for ongoing testing and monitoring of AI models to ensure they remain safe, effective, and unbiased over time.

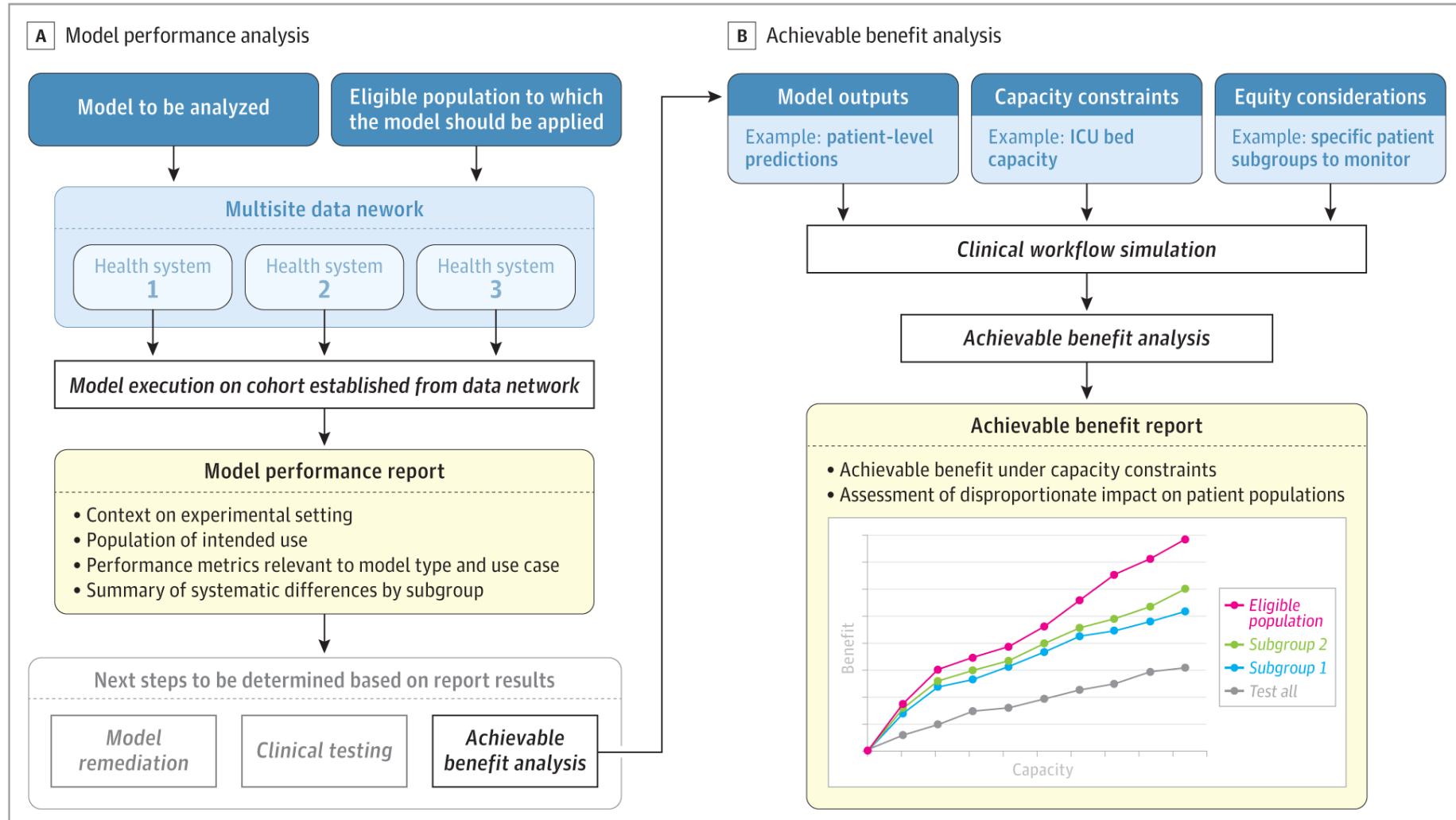
Best Practice Guidance

- With the input of experts that have real-world experience, developing best practice guidance for developing and deploying health AI technologies demonstrated use case by use case.

National Registry

- To support greater transparency, CHAI will publish model report cards on a publicly-available registry.

Nationwide network of health AI assurance laboratories (Shah, 2024)



Applied model cards and guidelines for responsible use (Palmieri, 2026)



Applied Model Card

Name: [Response goes here]		For inquires or to report an issue contact: abc@abc.com or +1 (999) 999 9999	
Release: [Response goes here]			
Summary [Response goes here]		Uses and Directions [Response goes here]	
AI System Facts [Response goes here]			
Warnings [Response goes here]			

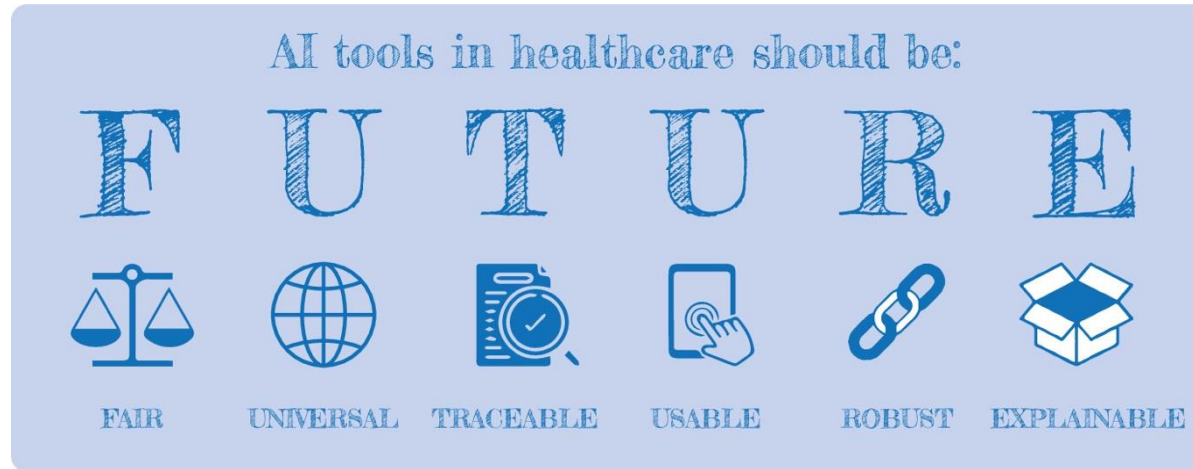
METRIC: Usefulness, Usability, and Efficacy		METRIC: Fairness and Equity		METRIC: Safety and Reliability	
Goal of the metric [Response goes here]		Goal of the metric [Response goes here]		Goal of the metric [Response goes here]	
Result [Response goes here]	Interpretation of the result [Response goes here]	Result [Response goes here]	Interpretation of the result [Response goes here]	Result [Response goes here]	Interpretation of the result [Response goes here]
Test type: [Response goes here]		Test type: [Response goes here]		Test type: [Response goes here]	
Testing Data Description: [Response goes here]		Testing Data Description: [Response goes here]		Testing Data Description: [Response goes here]	
Validation Process & Justification: Links to a method description		Validation Process & Justification: Links to a method description		Validation Process & Justification: Links to a method description	

Other information: [Response goes here]

Quoted guideline heading	Paraphrased key requirements
AI policies and governance structures	Require expertise and cross-functional representation (leadership, IT, compliance, clinicians, safety, privacy) Oversee selection, life cycle, compliance, and risk management of AI tools Report AI activities and adverse events to the governing board
Patient privacy and transparency	Establish data access, use, and protection policies consistent with HIPAA and state law Disclose and educate patients on AI use affecting their care; obtain consent when relevant
Data security and data use protections	Encrypt data in transit and at rest Apply strict access controls and audit logs Perform regular security assessments and maintain an incident response plan
Ongoing quality monitoring	Require vendors to supply validation and bias-testing data Continuously test, validate, and monitor AI performance and bias after deployment Scale monitoring frequency to risk proximity to patient care
Voluntary, blinded reporting of AI safety-related events	Implement confidential, deidentified reporting of AI safety events to independent entities Incorporate AI incidents into existing safety reporting systems
Risk and bias assessment	Identify, document, and mitigate risks and biases before and after deployment Request vendor bias data and model cards Use representative training and validation datasets; test on local data Perform regular bias audits
Education and training	Provide role-specific AI training for clinicians and staff Ensure documentation and guidance on each AI tool



FUTURE-AI: international consensus guideline for trustworthy AI in healthcare (Lekadir, 2025)



Clusters of requirements	Core principles
1. Fairness, diversity, inclusiveness, non-discrimination, unbiased AI, equity	Fairness
2. Generalisability, adaptability, interoperability, applicability, universality	Universality
3. Traceability, monitoring, continuous learning, auditing, accountability	Traceability
4. Human centred AI, user engagement, usability, accessibility, efficiency	Usability
5. Robustness, reliability, resilience, safety, security	Robustness
6. Transparency, explainability, interpretability, understandability	Explainability

Trustworthy and Responsible AI Network (TRAIN) (Embi, 2025)

Government agencies

US Food and Drug Administration
Assistant Secretary for Technology
Policy/Office of the National Coordinator
for Health Information Technology
Centers for Medicare & Medicaid Services

Nongovernmental organizations

Joint Commission
American Hospital Association

AI guidance organizations

Coalition for Health AI
National Academy of Medicine
World Health Organization

Practical health AI collaboratives

Health AI Partnership,
VALID AI

Trustworthy and Responsible AI Network (TRAIN)

TRAIN

Member organizations

Health systems (eg, academic health centers and individual and multihospital systems)
AI technology enablers (eg, AI developers and large tech vendors)
Stakeholder partners (eg, specialty societies and patient groups)

Services

Sharing lessons, strategies, and best practices
Federated AI performance testing and assurance
Education and expert consultation

Tools

AI registry platform
AI monitoring platform for algorithm vigilance
AI evaluation resources
AI maturity assessment and improvement

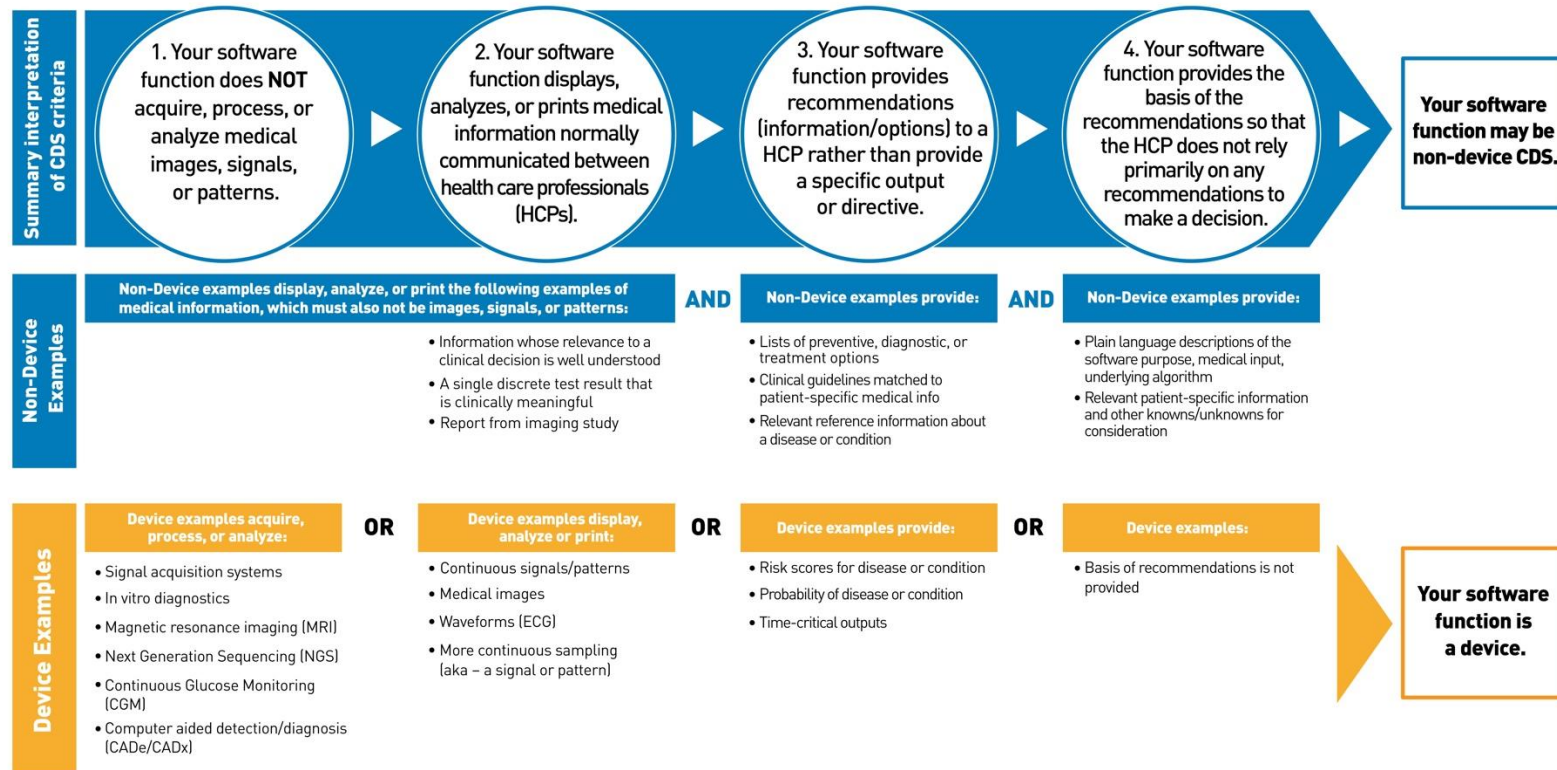
FDA guidelines for when software is or is not medical device

Your Clinical Decision Support Software: Is It a Device?



The FDA issued a guidance, Clinical Decision Support Software, to describe the FDA's regulatory approach to Clinical Decision Support (CDS) software functions. This graphic gives a general and summary overview of the guidance and is for illustrative purposes only. Consult the guidance for the complete discussion and examples. Other software functions that are not listed may also be device software functions. *

Your software function must meet all four criteria to be Non-Device CDS.



Education and training of clinicians and informaticians

AI should build on competencies in clinical informatics (Hersh, 2014; Hersh 2020; Hersh, 2023)

Others note

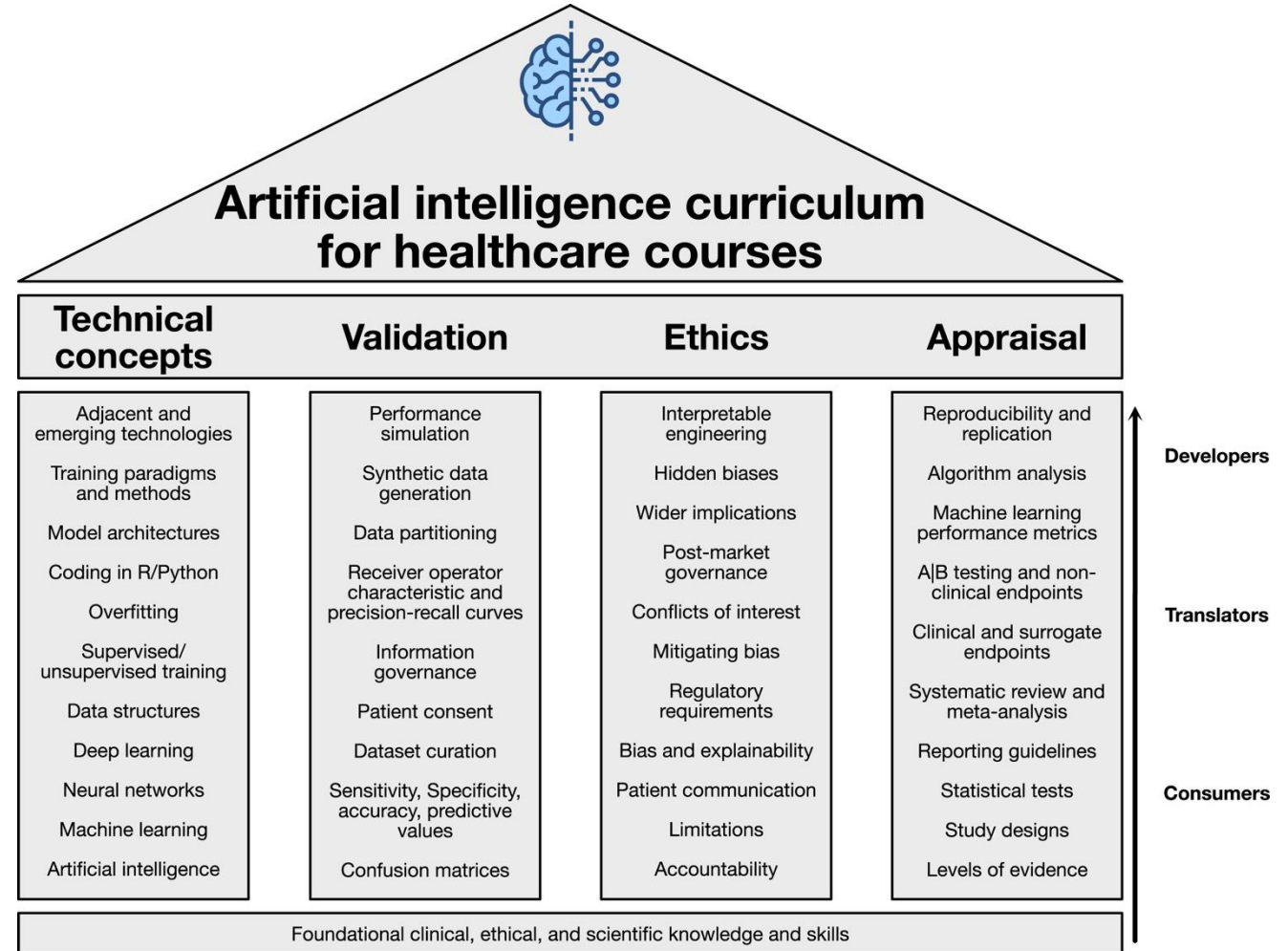
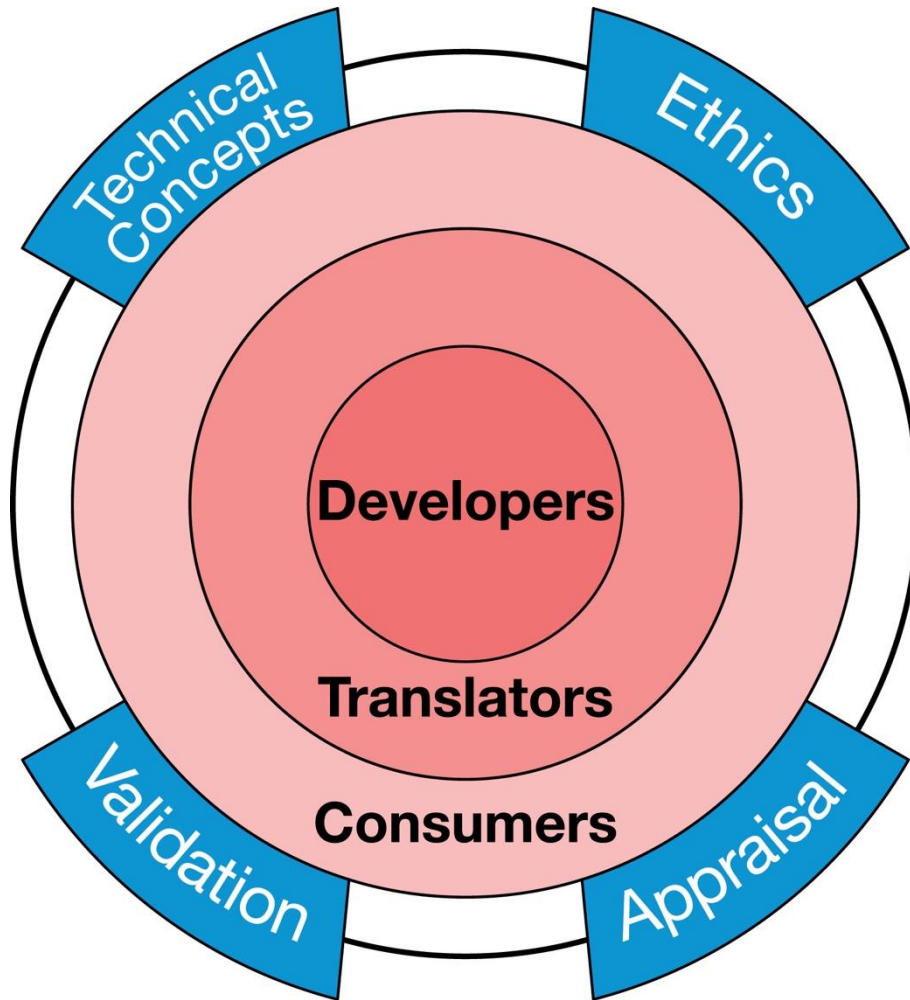
- Clinicians must be prepared to practice in a world of AI (James, 2022)
- Medical schools face dual challenges of needing to teach about AI in practice but also adapt to its use by learners and faculty (Cooper, 2023)
- AI education for clinicians has basic, proficient, and expert skill levels (Schubert, 2025)

AI-specific competency frameworks for professionals (Russell, 2023), primary care clinicians (Liaw, 2023) and educators (Pylman, 2025)

1. Find, search, and apply knowledge-based information to patient care and other clinical tasks
2. Effectively read from, and write to, the electronic health record (EHR) for patient care and other clinical activities
3. Use and guide implementation of clinical decision support (CDS)
4. Provide care using population health management approaches
5. Protect patient privacy and security
6. Use information technology to improve patient safety
7. Engage in quality measurement selection and improvement
8. Use health information exchange (HIE) to identify and access patient information across clinical settings
9. Engage patients to improve their health and care delivery through personal health records and patient portals
10. Maintain professionalism in use of information technology tools, including social media
11. Provide clinical care via telemedicine and refer patients as indicated
12. Apply personalized/precision medicine
13. Participate in practice-based clinical and translational research
14. Use and critique artificial intelligence (AI) applications in clinical care



AI education needs for developers, translators, and consumers (Ng, 2023)



Competencies for use of AI-based tools by healthcare professionals (Russell, 2023)

Domains	Details
Basic knowledge of AI	Explain what AI is and describe its healthcare applications
Social and ethical implications of AI	Explain how social, economic, and political systems influence AI-based tools and how these relationships impact justice, equity, and ethics
AI-enhanced clinical encounters	Carry out AI-enhanced clinical encounters that integrate diverse sources of information in creating patient-centered care plans
Evidence-based evaluation of AI-based tools	Evaluate the quality, accuracy, safety, contextual appropriateness, and biases of AI-based tools and their underlying datasets in providing care to patients and populations
Workflow analysis for AI-based tools	Analyze and adapt to changes in teams, roles, responsibilities, and workflows resulting from implementation of AI-based tools
Practice-based learning and improvement regarding AI-based tools	Participate in continuing professional development and practice-based improvement activities related to use of AI tools in healthcare



Competencies for use of AI in primary care (Liaw, 2023)

Domain	Bottom Line	Competency
Foundational knowledge	What is this tool?	Clinicians will explain the fundamentals of AI, how AI-based tools are created and evaluated, the critical regulatory and socio-legal issues of the AI-based tools, and the current and emerging roles of AI in health care.
Critical appraisal	Should I use this tool?	Clinicians will appraise the evidence behind AI-based tools and assess their appropriate uses via validated evaluation frameworks for health care AI.
Medical decision making	When should I use this tool?	Clinicians will identify the appropriate indications for and incorporate the outputs of AI-based tools into medical decision making such that effectiveness, value, equity, fairness, and justice are enhanced.
Technical use	How do I use this tool?	Clinicians will execute the tasks needed to operate AI-based tools in a manner that supports efficiency and builds mastery.
Patient communication	How should I communicate with patients regarding the use of the tool?	Clinicians will communicate what the tool is and why it is being used, answer questions about privacy and confidentiality, and engage in shared decision making, in a manner that preserves or augments the clinician-patient relationship.
Unintended consequences (cross-cutting)	What are the “side effects” of this tool?	Clinicians will anticipate and recognize the potential adverse effects of AI-based tools and take appropriate actions to mitigate or address unintended consequences.



AI in medical education

Competencies for medical educators (Pylman, AAMC, 2025)

- Understanding AI (What is AI and how does it work?)
- Working with AI
- Critical appraisal of AI outputs
- Ethical use of AI
- AI possibilities in medical education
- AI enhanced clinical encounters (curriculum for students)
- Using AI in research and scholarship
- Continuous professional development (CPD) in AI

Recommendations (Macy Foundation, 2025)

- Promote the implementation of safe and effective AI in medical education
- Implement coproduced, AI-enhanced teaching, learning, assessment, and evaluation approaches in medical education
- Develop and leverage robust data ecosystems to facilitate AI implementation in medical education
- Facilitate ethical AI governance frameworks and responsible utilization of AI tools and technologies in medical education
- Advance scholarship on AI in medical education

Patient/consumer education important as well

Health AI literacy also important skill for patient empowerment (Campos, 2025)

Guidelines for responding to disagreements with patients having AI-generated answers (Bisarya, 2025)

Educational strategies

Clinical education (Abdulnour, 2025)

- Need to be cognizant of de-skilling, never-skilling, and mis-skilling
- Must be “deft” – discussion, evidence, feedback, teaching

Need to figure out “opposite of cheating” (Gallant, 2025)

- Cheating results from students not valuing educational activities and/or being motivated by external needs (e.g., getting into medical school)
- Students construct, not receive knowledge
- Students learn from making errors
- Students must emerge with ethics and integrity in application of knowledge
- Faculty should apply explicit objectives with meaningful and secure assessments

Toward the future

Consensus strategies (Angus, 2025)

- Multistakeholder engagement throughout total product life cycle
- Measurement tools for evaluation and monitoring developed and disseminated
- Creation of nationally representative data infrastructure and learning environment to support generation of generalizable knowledge about health effects of AI tools across different settings
- Incentive structure should be promoted, using market forces and policy levers, to drive changes

Other requirements

- AI-capable organizations (Novak, 2023)
- Sustainable deployment – model maintenance (Davis, 2024)
- Interoperable digital eco-system (Mandl, 2024)
- Translational AI from basic to clinical science (Hersh, 2024)
- Commercial forces should not undermine value (Mandl, 2025)
- Changes in healthcare to give best incentives for use (Razavian, 2025; Gottlieb, 2025)

Toward the future

Real-world use not yet to plateau of productivity

Need “algorithmovigilance” (Embi, 2021) and human vigilance (Adler-Milstein, 2024)

“AI won’t replace radiologists, but radiologists who use AI will replace radiologists who don’t,” (Langlotz, 2019)

- Radiology a “dying profession” (Hinton, 2016; quoted by Wood, 2024)? Unlikely (Lohr, 2025; Mousa, 2025)
- AI likely to reduce radiologist work time by 33%, counterbalanced by growing use of imaging and static workforce (Langlotz, 2025)
- Probably same for other health professions

