BLUEPRINT FOR TRUSTWORTHY AI IMPLEMENTATION GUIDANCE AND ASSURANCE FOR HEALTHCARE

COALITION FOR HEALTH AI

VERSION 1.0 _ APRIL 04, 2023

ACKNOWLEDGEMENTS

Information presented is a summary of information from discussions from a group of subject matter experts from healthcare and other industries in a series of virtual workgroup meetings and convenings funded by the Gordon and Betty Moore Foundation and conducted in collaboration with the Partnership on AI (PAI), and the Coalition for Health AI (CHAI) Steering Committee, notably including leadership from: University of California Berkeley; University of California Health; University of California San Francisco; University of California Office of the President; Duke Health; Google; Johns Hopkins University; Mayo Clinic; Microsoft; MITRE; Optum; SAS; Data Science Team, Stanford Medicine; Vanderbilt University Medical Center as well as US Federal Observers from the Agency for Healthcare Research and Quality (AHRQ); Centers for Medicare & Medicaid Services (CMS); US Food and Drug Administration (FDA); the Office of the National Coordinator in Health Information Technology (ONC); the National Institutes of Health (NIH); and the White House Office of Science and Technology Policy (OSTP). We would like to thank all contributors and participants in these ongoing discussions for their thoughtful input and feedback.

The following are copyrights and recognitions with regards to the Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare.

Discussion and refinement toward a blueprint for an implementation guidance for health artificial intelligence (AI) are ongoing and a full landscape analysis will be performed in a subsequent phase of the Coalition for Health AI's work.

TABLE OF CONTENTS

PURPOSE

The use of artificial intelligence (AI) in healthcare offers enormous potential for accelerating clinical research and improving the quality and efficiency of healthcare delivery. However, a growing body of evidence demonstrates that the adoption of AI and the subset of AI known as machine learning (ML) may also increase the risks of negative outcomes for patients and introduce or worsen bias. There is therefore an urgent need for a framework focusing on health impact, fairness, ethics, and equity principles to ensure that AI in healthcare benefits all populations, including groups from underserved and under-represented communities. When standard guidelines are not harmonized or are poorly understood, the potential for distrust of AI is increased among both healthcare providers and patients. Moreover, there is currently an inability to easily assess the robustness of algorithms on relevant data and evaluate the process for health systems developing and deploying AI and machine learning.

This report is the result of convening experts from multiple institutions representing healthcare systems, academia, government, and industry, through the Coalition for Health AI (CHAI), to identify and propose solutions to issues that must be addressed in order to enable trustworthy AI in healthcare. Specifically, this work summarizes collective recommendations as a step toward a blueprint for assurance standards on trustworthy AI in healthcare. This blueprint will enable health AI, harmonizing standards and reporting, and educate end users on how to evaluate AI technologies in ways that can drive their responsible adoption. Furthermore, the goal of this blueprint is to facilitate guidelines regarding an ever-evolving landscape of health AI tools to ensure high-quality care, increase trustworthiness among the healthcare community, and meet the needs of patients and providers.

BACKGROUND

Because healthcare applications can have a critical impact on patient outcomes and well-being, AI in healthcare must meet high standards for safety, efficacy, equity, and usability. However, some recently published analyses of AI-based algorithms have raised concerns. For example, a study that assessed 415 published deep learning and ML models designed to diagnose COVID-19 and predict patient risk from medical images such as chest x-rays and chest computed tomography scans found that none were meeting their intended purpose (1). Further, a "living review" of 232 diagnostic and prognostic algorithms for COVID-19 found that all of the models had either a "high or unclear" risk of bias (2).

Failure to provide information about AI system characteristics, behavior, efficacy, and equity can limit trust, acceptance, and ethical and proper use of these systems. This information is critical to the safe and responsible use of AI systems. In recent years, multiple general-use resources that characterize ML systems have been published, including FactSheets (3,4), Model Cards (5), and ML Test Score (6). Several resources for characterizing AI-based clinical systems are intended to support assessment of clinical trial protocols, such as SPIRIT-AI (7) and CONSORT-AI (8), or assessment of published studies, such as TRIPOD (9), CHARMS (10), PROBAST (11), STARD (12), and DECIDE-AI (13). Still others are intended to assist researchers and/or developers in determining the appropriateness of models for incorporation into biomedical or clinical applications, including MI-CLAIM (14), Algorithm-Based Clinical Decision Support Oversight (15), Risk Prediction Model (16), bias checklists (17,18), Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research (19), and reporting standards such as MINIMAR (20). However, there are few guides that offer a holistic approach to assessments of AI-based clinical systems for health systems, consumers, and end users.

This work has brought together a collaboration across institutions with expertise in different areas relevant to this effort to attain sufficiently broad coverage. The goal was to ensure applicability to a wide range of clinical AI-based systems and thus facilitate widespread adoption. Although there are current efforts to develop core components for AI/ML for specific medical applications (21), the clinical AI/ML community would benefit from an approach that is applicable to AI-based clinical algorithms for various uses (e.g., diagnostic, prognostic) and clinical subdomains (e.g., oncology, cardiology, etc.).

Experience suggests that it is difficult to build ecosystems when multiple approaches are left to bloom in the wild without a consensus-based standardization. Thus, it is important to assemble a guiding coalition that can agree on a canonical structure for health AI assurance standards for throughout the application's lifecycle. We recognize the importance of an iterative process for developing guidance. Over time, assurance standards can change as needed. Our goal is to have a group that builds this consensus together while avoiding disparate, conflicting approaches that prevent developers and others from knowing what AI applications or technologies to adopt and how to implement AI in a clinical setting. Such a group, as well as the processes used and guidelines developed, should include input from stakeholder groups such as those listed below. By summarizing the culmination of a year of work via industry, academia, and government participants, this work explores the parameters for the guidance, guardrails, best practices, and governance needed to help ensure trustworthy AI.

Stakeholder Groups	**Stakeholders**
Data Science	Data Scientists
Informatics	Informaticists, Software Engineers, Vendors
End users	Providers, Clinicians, Nurses, including trainees
	Health Care Operations
	Insurers, Payors
Patients	Patient Advisory Groups
	Patient Advisory Boards
Regulatory and Policy	Legal
	Ethics
	Government/Policy

	Professional Societies that publish and review clinical practice guidelines
Health Care Administration	Health Care Leadership
Research	Translational and Implementation Science
	Research Funders
Trainees	Educators, computer science students, medical, nursing, and public health informatics students, continuing education.

KEY ELEMENTS OF TRUSTWORTHY AI IN HEALTHCARE

In alignment with the NIST AI risk management framework, we have structured this section to parallel NIST definitions and extend/view them with respect to healthcare. These concepts build upon foundations of validation and reliability. To ensure trustworthiness, the NIST AI risk management framework describes four key functions (map, measure, manage, and govern). **MAP** establishes the context for framing risks related to an AI system. **MEASURE** employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts. **MANAGE** function entails allocating risk resources to mapped and measured risks on a regular basis and as defined by **GOVERN**, which is a cross-cutting function infused throughout AI risk management that enables the other functions of the process (22). The expectation is that organizational practices are carried out in accord with "professional responsibility," defined by ISO as an approach that "aims to ensure that professionals who design, develop, or deploy AI systems and applications or AI-based products or systems, recognize their unique position to exert influence on people, society, and the future of AI" (23). Each element described therein contributes to the deploy/withdraw decision that is made and periodically *re-evaluated* by organizations for specific use cases.

Moreover, NIST refers to **social responsibility** as the organization's responsibility "for the impacts of its decisions and activities on society and the environment through transparent and ethical behavior" (24), and **sustainability** as the "state of the global system, including environmental, social, and economic aspects, in which the needs of the present are met without compromising the ability of future generations to meet their own needs" (22,23).

3.1 Useful

For an algorithm to be *useful* (25), the algorithm must provide a specific **benefit** to patients and/or health care delivery and be **usable,** beyond being **valid** and **reliable**. For example, an algorithm with **intended benefit** in a pilot population with poor usability may not achieve an impact on clinical outcomes. Measurement of utility can also be streamlined using technological methods. Here, we are using the term usefulness (relevant to the impact on society and patients) to describe algorithms that are:

- Valid with respect to accuracy, operability and meeting intended purpose and benefit (clinical validation)
- Reliable
- Testable
- Usable
- Beneficial

3.1.1 Valid and Reliable

NIST defines **validation** as the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (24). Certain AI/ML-enabled technologies may be considered Software as Medical Device (SaMD) and thus subject to regulation (26). In cases where these technologies are regulated as SaMD, process validation is mandated prior to use (27,28). Software validation typically includes installation qualification, operational qualification, and performance qualification (29,30). When deployed, health AI systems shall be valid for accuracy, operability, and meeting their intended purpose to provide benefit. These systems should also be monitored, with controls established, and re-validated.

The concept of **reliability**, as defined in ISO standards, captures the ability of any item (in this case, an AI model/tool) to perform its required function without failure, under stated conditions and over a defined time interval (31). In the application to healthcare, we further contextualize this into **reliability** and **reproducibility. Monitoring** is then essential to understanding reliability and reproducibility.

Key facets of reliability include failure prevention, workflow integration, and robustness under dataset shifts. The goal of failure prevention is to minimize the likelihood of **failure**, defined as "the termination of the ability of an item to perform a required function" (32). One reason that reliability is important comes from differences in or changes to the environment in which the tool is used. Specification of a tool's intended use is heavily affected by such dataset shifts. In addition, how the model is integrated into other systems can affect its reliability. Intended use, known error cases, and measurements of reliability should capture the role the tool plays in the broader, human-centric clinical workflow. This workflow includes actions taken by the clinical team as a result of this model.

Reproducibility is an important related factor for ensuring that outcomes are consistent across sites (e.g., between hospitals or even between units) and thus for reliability of the entire health system in question. AI/ML is particularly sensitive to variations in hardware and software versions. As with other variables (e.g., testability), reliability should be considered across the model's life cycle. This may even be influenced by concept drift and changing data (e.g., lab instruments changing over time). As models evolve, their ability to reliably influence clinical workflows may be affected by no-longer-accurate mental models
retained from previous model iterations or shared by colleagues. Thus, an assurance

standards guide should capture information such as metrics for reliability, embedded workflow, versioning, expected datasets, and guardrails for drift (33).

Monitoring involves the ongoing surveillance of an AI tool to raise an alarm when shifts in the input data, tool outputs, or user behavior are detected. In monitoring, it is important to identify failures and vulnerabilities quickly so that negative effects are minimized. This includes central reporting, which allows all sites to learn from the experiences of others. This is critical in rare incidents, in which individual sites might not recognize a pattern but combining information across sites can enable faster event detection. When monitoring, it is important to consider backward compatibility. Model updates should not reduce the quality of human-AI collaboration. Currently, most monitoring is manual but new tools are being developed for automated real-time audits of individual predictions (34). Such tools will be vital for improving oversight and governance.

An assurance standards guide can help define the type of information relevant for inclusion in the specification for an ML model and its encompassing tool. When monitoring, it is useful to predefine what actions will be taken based on the monitoring results. Having a predefined protocol can be useful when unintended model behaviors arise, especially in real-time, high-volume cases where decisions that could affect many end users must be made quickly. In addition to shutting down a system, there may be a continuum of possibilities such as Bayesian learning, stepping back temporarily, etc.

Monitoring should be surveyed across various settings. Metrics may be monitored upon the live deployment of a system, but also focus on monitoring algorithm-level issues and workflow-level reliability. Guidance is also needed regarding how often models should be updated and systems maintained.

3.1.2 Testable

In this report, **testability** refers to the extent to which an AI algorithm's performance can be verified as satisfactory in terms of meeting *all* standards for trustworthy AI, including topics covered in this brief such as robustness, safety, bias mitigation, fairness, and equity in both development and evaluation. Testability requires a strong contextual understanding of the model and its intended use, including *where, why,* and *how*. Consequently, testability must be evaluated at levels ranging from (inter)national regulatory bodies (e.g., FDA, EMA) to institutions, and even down to individual healthcare units.

The model's entire life cycle from conception through deployment—not just training and development phases—must be considered as part of testing. It is not only important to look at the training phase or when in deployment. In fact, testing is integrated into the model lifecycle and is not a single step or even a set of discrete individual steps but should instead be understood as a continuous process. An assurance standards guide can detail the various elements that need to be assessed during the different stages of the health AI lifecycle (see the monitoring section above for additional discussion).

3.1.3 Usable

Usability is defined here as denoting the quality of the user's experience, including effectiveness, efficiency, and satisfaction, when using an algorithm's output. Usability comprises several key factors. The first is context: usability is heavily dependent on a model's context, such as emergent situations versus continuous surveillance for patients admitted to the hospital. The second is the end-user and/or patient or stakeholder perspective. Patient perspectives should be incorporated as early as the design phase of the health technology. It is important that end users be able to contribute to the assessment of usability. Simplicity is another key variable. Other tradeoffs aside, a simpler model is often easier to use, and excess complexity decreases usability.

Workflow considerations are also important for usability. For non-emergency notifications, non-intrusive alerts are preferable because they do not interrupt workflows and can be evaluated together at the appropriate time, thereby reducing alert fatigue. At times, explainability may detract from usability, depending on how it is implemented.

In thinking about an assurance standards guide, there are some key items to address. These include delineating how usability is measured and by whom. Usability is typically determined by end-users. Another area to explore is defining how patient perspectives can best be incorporated into usability design and the workflow itself. Finally, there is a need for designs that help end users who may not have data science training to understand a model's output.

3.1.4 Beneficial

The ***benefit*** of an algorithm should be measured by the algorithm's impact on its intended outcomes (effectiveness) and overall health through its intended (and unintended) use, weighed against deleterious effects and risks.

When testing, it is important to understand the current state of the workflow to which the AI technology will be introduced in order to determine the effectiveness of the technology as it is integrated into the workflow. The status quo, against which the model is compared, may be difficult to fully define; however, working to define one can help to capture the value of the model and potential return on investment (ROI) and return on health (ROH), thereby increasing the adoption rate.

Introducing a new health AI technology can be quantified and its significance can be demonstrated statistically using different study design methodologies such as randomized clinical trials. However, although randomized controlled studies are the preferred standard for validating new clinical interventions, the approach was conceived to deal with conventional medical treatments (usually new drugs or devices) and is not always easily matched to the evaluation of algorithms or AI-based decision-supposed tools. Study designs that can demonstrate the effects of an algorithm on patient outcomes will require a preliminary, staged set of evaluations designed to demonstrate trustworthy health AI. Establishing a pathway or guide for implementation also has the potential to set standards for the evaluation of such systems.

An assurance standards guide (35,36) could include notes on the level of evidence and types of study designs used to assess model effectiveness in terms of validity, acceptability, fairness, equity, transparency, and health impact. Furthermore, documentation, reproducible methods, and accessible code are important for multisite testing and can be embedded as pointers in a schema for the relevant resources.

During each phase of the lifecycle, different common issues arise in testing. An assurance standards guide should call these out for different phases of the lifecycle and enable capturing performance metrics, provenance, and other information to ensure testing results can be examined.

An assurance standards guide can also help capture relevant information about the testing and its results across the lifecycle. In addition, there are several policy issues to be defined including what type of AI tool could be testable, who is responsible for testing it, and how to incentivize and/or enforce routine testing in the model lifecycle. Finally, it would be helpful to have guidance on strategies to address health systems' responses if a model fails testing, as well as standard procedures that should be done as potential next steps. There may be impacts other than the intended effects, such as those on society.

3.2 Safe

In the context of safety, NIST's AI Risk Management Framework (22) notes that **safe** AI systems are ones in which "human life, health, property, or the environment" are not put at risk of harm (31). In health care, this refers to preventing worse outcomes for the patient, provider, or health system from occurring as a result of the use of an ML algorithm. An AI-enabled device can become unsafe for many reasons (37). A lack of oversight on ensuring fairness, addressing, or mitigating bias, and ensuring accountability can make any model unsafe. Looking at the potential role of and appropriateness of outcome proxies is also important. Using a proxy for a desired outcome, instead of the desired outcome itself, can create additional risk. Models that are aligned with a proxy of a desired outcome can potentially lead to unintended and unsafe consequences. Further, model performance may deteriorate or change in unexpected ways in response to underlying shifts in data, rendering the model unsafe to use (38). There may also be downstream impacts that may not be readily known or available in the development process for the model.

As a baseline, safe AI models should not create worse outcomes than the status quo. A known safety risk of algorithmic technologies is automation bias (i.e., uncritical acceptance of an automated suggestion). As with testing, considering the entire lifecycle and considering unintended, downstream consequences of AI deployment is vital. An assurance standards guide should define metrics and provenance information, including how safety is measured and by whom this information is captured. It should define how safety events caused by AI could be identified and reported. Furthermore, it should define and enable the parties that provide data (e.g., hospital EHRs, patient-generated health data) on roles and responsibilities for maintaining safe AI.

Finally, it should offer opportunities to reevaluate the status quo—as we discover new sub-populations, we can re-evaluate them for noninferior outcomes.

3.3 Accountable and Transparent

Accountability describes the responsibility of individuals involved in the development, deployment, and maintenance of AI systems to maintain auditability, minimize harm, report negative impact, and communicate design tradeoffs and opportunities for redress. The concept of **transparency**, meanwhile, reflects the extent to which individuals interacting with an AI system or whose data are input into an AI system have access to information about that system and its outputs, regardless of whether they are aware that they are interacting with an AI (34).

In healthcare, the **transparency** of an AI model implies traceability. For a model to be transparent there must be precise communication from the time of dataset curation and model design to the model's final output, encompassing performance, confidence level, and generalizability. The type of information reported must be adapted to each stakeholder's perspective and needs.

Enabling transparency is not a one-time process. To maintain transparency, the model must be continuously evaluated and addressed throughout the AI system lifecycle. Transparency is enabled when criteria involving the selection and curation of underlying datasets, the validation and reliability of the models, and the engagement of stakeholders, patients, and end-users are considered.

For datasets, there should be a standardized process and policies in place for curation. Each dataset should include relevant metadata. Furthermore, the collection process must be specified. Inclusion and exclusion criteria, demographic information with diversity details, and device characteristics should be included. The provenance and limitations of the data will need to be specified.

For models, the motivation and intended use of each model should be disclosed and the decisions used to design a model should also be made public. There should be transparency regarding the data used to train the model. There should be robust external evaluation to guarantee generalization before deployment in healthcare settings. It is important to have disclosure of a model's performance and level of confidence for each output. The model should be continuously evaluated throughout its lifecycle and be adaptable in response to feedback.

For stakeholders, considerations regarding the audience are critical. For example, different types of information should be provided for technical versus nontechnical audiences. There should be clear communication regarding tradeoffs made by the model. As stated in the bias and equity section, diverse multidisciplinary teams including stakeholders, patients, and end users should be involved or engaged throughout the model lifecycle.

An assurance standards guide can help address transparency when multiple datasets and/or models are combined. In some cases, data used for training may not be public and algorithms themselves may be proprietary. It may be helpful to define approaches for further transparency in these cases. In terms of transparency for end users, model cards (21) have been used for this purpose. Similar to a nutrition label, model cards can be designed to provide specific information to increase transparency based on the technical knowledge of the end user. There are questions around policy for models already deployed and datasets already in use. For example, they could be exempted from requirements, given certain time to follow proposed policies, retired automatically, or given guardrails to follow. A framework for transparency in datasets and models would be the next step upon which a certification process could be built as well.

3.4 Explainable and Interpretable

Explainability, according to definitions offered by NIST, refers to a representation of the mechanisms underlying AI systems' operation, whereas **interpretability** refers to the meaning of AI systems' output in the context of their designed functions. Notably, this differs from NIST's definition of **transparency**, whose scope "spans from design decisions and training data to model training, the structure of the model, its intended use cases, and how and when deployment, post-deployment, or end user decisions were made and by whom"(22). Both explainability and interpretability are critical to building user trust in health AI.

Explainability without interpretability may lead to physicians understanding the computing principles behind a "black box" model without being able to debug the results for the current patient. Interpretability without explainability may offer insight into why a model result is produced without helping physicians understand how to adapt their mental model for other patients.

3.5 Fair – with Harmful Bias Managed

In this report, **bias** refers to disparate performance or outcomes for selected groups defined by protected attributes such as race and ethnicity, and, in this paper, differences that are perpetuated and/or exacerbated by AI models and their use. Bias, equity, and fairness are interrelated. In **equity**, the goal is to ensure that everyone has the opportunity to achieve their full health potential, regardless of a specific group membership. With regard to health AI, this means ensuring that AI, through action or inaction, does not increase a specific group's risk for bias or adverse fairness outcomes - similar to noninferiority studies in pharmacology, defined by the NCI as "a study [that] tests whether a new treatment is not worse than an active treatment it is being compared to." Algorithmic **fairness** refers to the multidisciplinary field of study that seeks to define, measure, and address fairness as it relates to algorithms used for decision-making. There are several key aspects to consider for algorithmic fairness: better design of new algorithms being built (39); audits of performance and consequences of currently used algorithms (40); and examination of the consequences of algorithm use on a regular cadence.

Leveraging health equity by design involves looking with intention at the goal of promoting health equity (41). This entails defining equity goals, potentially as part of an institution's overall quality program. As part of this process, all stakeholders and community members should be included throughout the entire lifecycle of the AI tool (42). This involves everything from data collection to deployment, as well as behavioral considerations for algorithm/user interaction (see later elements on testability and usability). In addition to health equity, there are often multiple variables that are being optimized at the same time (performance, fixed costs, profit, value, etc.). The key is to make an informed decision considering the inherent tradeoffs with other goals, thereby ensuring that the various factors are ultimately explicitly weighed as desired by the corresponding organization/user.

There are processes and measures that can help evaluate AI for potential bias, equity, and fairness. However, it is not possible to completely predefine the set of measures and processes that are required for specific settings. Establishing frameworks and checklists can help guide decisions (17,18). Overall, there should be multiple checkpoints for every stage in the AI design, development, and implementation lifecycle and at different points during the stages of evaluation and continual monitoring. This is needed to account for AI system performance against historical data, data generated in current settings to access dynamic socio-demographic changes in population, practice patterns, user behavior, and updates to scientific data and clinical evidence. For example, this requires not just examining the algorithm itself and its output, but also evaluating how it works and its impact. The algorithm may use proxies that are correlated with variables such as race, which might not be known unless carefully considered together. Monitoring structures need to be set up as an iterative process that includes multiple checkpoints. These should be placed before and during model training as well as before and after deployment. This helps ensure that there is no unseen data shift or other issues that may have degraded performance or introduced new biases in the model and associated workflow.

There are several approaches that can help mitigate algorithmic bias in health AI and promote health equity (17,18). Better incentives are needed to promote health equity by design. This includes incentives to fix data at the collection step instead of only focusing on phases involving model development and deployment. Regular fairness audits may need to be conducted (40).

3.5.1 Systemic Bias

Systemic bias can be present in AI datasets; in the organizational norms, practices, and processes across the AI lifecycle; and throughout the broader society that uses AI systems (22). In healthcare, one situation occurs when an algorithm is known to work well in a certain population but not in another one. In some cases, such algorithms may not be used at all or may only be applied to the subset of the population where high performance is seen.

- **Measurement bias** is introduced when there are differences in quality or ways that features are selected and calculated across groups.
- **Missing validation bias** occurs when there is a lack of

validation studies to examine performance in subgroups before deployment.

- Model definition and design biases include **label bias** and **modeling bias**.
- **Label bias** occurs when biased proxy variables are used in place of ideal predictive variables during model training.
- **Modeling bias** occurs when a model's design yields inequitable outcomes.

The concept of "algorithmically underserved" helps illustrate several aspects of bias, equity, and fairness and illustrates health equity by design and the associated processes that may be important to apply (43). Careful work is needed to ensure each of these aspects is considered. One example of a program in this area where some guidelines are being developed is the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) Program (33).

3.5.2 Computational and Statistical Biases

Computational and statistical biases may occur in datasets used to train AI systems and may also be present in the resulting algorithmic processes. Such bias often stems from systematic errors due to non-representative samples (34). In healthcare settings, these biases may result in some patients being underserved because they do not have data recorded/available, possibly because some/all of their records are not available electronically or available on platforms that support algorithmic/clinical decision support apps such as SMART-on-FHIR or CDS-Hooks-capable systems (44,45). It may also be that the patient explicitly decided to decline to make their data available or simply choose not to complete forms/information fully.

Another aspect is **population bias**, in which some patients represent populations for whom insufficient data are available to evaluate the performance of models with confidence. For example, an American Samoan patient may be algorithmically underserved when there is too small of a sample size available in the training set.

3.5.3 Human-Cognitive Biases

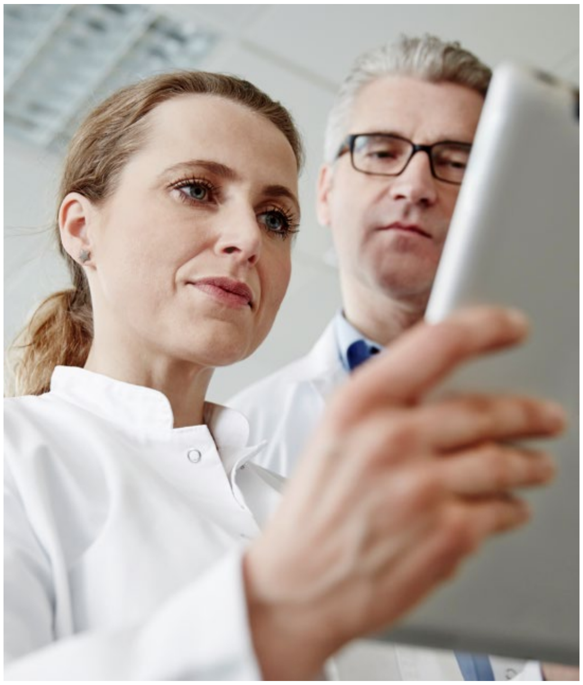
Human-cognitive biases, as defined by NIST, are those that relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about the purposes and functions of an AI system (22). Governance is key to fairness-affirming strategies and to overseeing bias mitigation. Establishing who governs and how governing occurs in standardized ways can help mitigate risks. This requires a multidisciplinary team to establish processes and measures for bias.

3.6 Secure and Resilient

NIST notes that AI systems, as well as the ecosystems in which they are deployed, may be considered **resilient** if they are able to withstand unexpected adverse events or unexpected changes in their environment or use, or if they can maintain their functions and structure in the face of internal and external change, degrading safely and gracefully when necessary (31). Furthermore, AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be **secure** (22). In the context of healthcare, these definitions still apply. Because ensuring a high degree of availability for health applications is paramount, safe and graceful degradation is a crucial component for redundancy and resilience.

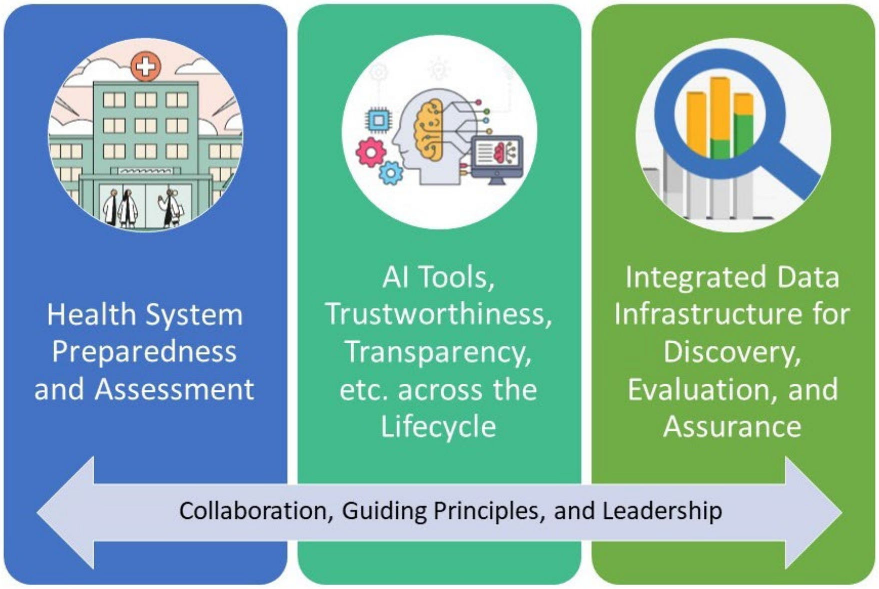
3.7 Privacy-Enhanced

Although NIST's definition of **privacy** refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity (22), healthcare has unique standards for privacy that already prevail. For example, in the United States, the 1996 Health information Portability and Accountability Act (HIPAA) establishes rules governing the collection, handling, transmission, storage, and disclosure of protected health information (PHI) and offers standards for deidentification of PHI. These standards evolved over time to encompass additional protections, as in the 2008 Genetic Information Nondiscrimination Act (GINA), while other jurisdictions may be bound by rules that require a different set of protections, such as the European Union's General Data Protection Regulation (GDPR).



4. Next Steps

Each healthcare institution may use different kinds of AI tools. However, there is a need to use a common, agreed-upon set of principles to build them and facilitate their use. Through an assurance lab, health systems as well as tool developers and vendors can submit processes and tools for evaluation to ensure readiness to employ AI tools in a way that benefits patients, is equitable, and promotes the ethical use of AI. In large medical centers, resources to support such measures may already exist; however, this may not be true of other small, rural, and/or resource-constrained health systems. There may thus be a need for an advisory body to advance the field with these entities as well and ensure equity so that, for a given patient, access to trustworthy health AI would not depend on where they live or with which health system they are interacting.


Health System Preparedness and Assessment
AI Tools, Trustworthiness, Transparency, etc. across the Lifecycle
Integrated Data Infrastructure for Discovery, Evaluation, and Assurance
Collaboration, Guiding Principles, and Leadership

Below are the pillars for how an AI assurance, evaluation, and discovery lab can help achieve results through health system preparedness and assessment, AI tool trustworthiness and assessment, and integrated data infrastructure for enabling trustworthy AI.

4.1 Setting up Assurance Lab and Advisory Service Infrastructure

Interdependent assurance labs and associated consulting services will help in creating an ecosystem that has at minimum four infrastructure components: a **shared definition of value** and components such as **registries** of tools, templates of **legal agreements** as well as **sandbox environments** for testing tools.

4.1.1 Identifying and Articulating Value

Because negative financial margins are common for many health systems, it is important to ensure a clear value proposition for the patient and the organization for deploying AI solutions, and beginning with a value proposition evaluation is recommended. Demonstrating the value framework can engage the enthusiasm of decision-makers, and then the other elements can be done to lead to better patient outcomes and return on investment (ROI). Governance requirements, bureaucratic processes, and best practices come secondary to value in terms of securing initial buy-in. Thus, one goal for assurance standards, including potential consultation services, would be to serve as an enabler of value for health systems and their patients, which also includes ensuring that policies do not deplete that value. For example, there is a risk of overburdening our health systems with excessive reporting or regulatory requirements.

On the other hand, initial processes to understand the value proposition are just the beginning. There is a need for a structured intake process for candidate use cases (in which a model would drive a clinical care workflow) based on virtual model deployments that calculate achievable utility via simulating several days of care workflows (36). Enabling such analyses will require structured checklists that require the submitter to think through how the potential use of the AI tool would impact the workload in the organization and what tasks would need to be done, and by whom, to realize the value.

Value also needs to be demonstrated for the patient, for the healthcare delivery system, and for society. This includes incentivizing developers to participate and ensuring that value is derived for promoting transparency and ethical oversight throughout the entire process.

Finally, a maturity model can be developed and applied both to health systems and on the tools used. Several maturity models exist (46,47) but await further development for adoption, spread, and scale within healthcare.

By understanding the level of maturity of an organization, the next steps needed in the consultation process will become apparent to enable the value proposition. The other approach is to establish maturity models for the developers of the AI models or the models themselves, as in the Food and Drug Administration (FDA) precertification pilot (48) that sought to establish criteria for the industry developers rather than the device/tool itself. For models, establishing guardrails with potential intervention points may be another option, as describe in the FDA's Guidance for Industry addressing AI software as a medical device (49).

4.1.2 Registries

One approach to empower patients is to create a registry for AI tools, analogous to the ClinicalTrials.gov clinical trials registry. Registries could also be established at an institutional level. The key is the local implementation of a uniform national framework. Patients could look up what is available in their own facility and see the tools. Care providers and AI tool developers could compare algorithms and analytic options by reviewing the registry and would be able to examine model cards and other publications that propose nutrition-like labels for AI models.

Health care providers with access to information about a patient's clinical history, phenotype, genotype, etc. can interact with such registries to see if a particular algorithm is likely to perform well. Ideally, the algorithm can be downloaded from the registry and interact with the patient's data and provide results to clinicians. Like clinical trials, AI tools are created in different institutions using different populations. This information could be captured as metadata in a registry and used to help determine when the underlying algorithms may be suitable for a particular patient, thus facilitating precision medicine.

To build such a registry, technology and policies should be developed to enable it to be used as part of an ecosystem. An assurance lab can help ensure that the information on such registries is trustworthy. There can be thousands of data sources that are integrated. The registry of tools can help increase transparency and provide a platform for evaluation rubrics that can inform data and model validation and other aspects necessary for an ecosystem to flourish.

4.1.3 Standardization and Sandboxes

To establish an assurance lab and associated technical assistance service, there should be agreement on a set of reporting criteria necessary to perform such an evaluation on an algorithm. It also necessitates willing institutions. Several existing organizations already have **sandboxes** for testing models locally. While not all models can be built on local data, the validation should be done locally (or at least with local data/workflow conditions). An evaluation and monitoring sandbox platform that includes a data standards-based federated repository can help ensure long-term reliability of new AI algorithms as well by enabling evaluation and ongoing monitoring to identify bias, detect performance degradation due to data shift, and assess usefulness of algorithms.

Standardization can enable a marketplace in which data providers and algorithm developers can collaboratively contribute to validation. This includes creating a template-based, checkbox legal agreement approach for the participation of the data providers and the algorithm developers for validation. There are some existing exemplars of such legal agreements for data sharing/testing, including from two-party agreements to multiple-party industry-based datasets.

With standardized templates for agreements, much of the time currently consumed by legal negotiations can be saved. Furthermore, having standard schemata for data will accelerate the process so that data can easily be processed by an assurance lab and technical assistance service. Approaches to creating such agreements on sharing data and creating sandboxes have already been demonstrated on a smaller scale. Convening a group would enable scaling to expand that to more parties, with more use cases, and with more data types as the technology and policy allow.

4.1.4 Independence

One requirement for an assurance lab is its independence, which is needed for building trust among potential stakeholders and users and enabling collaborative work in the precompetitive space. Without having conflicts of interest, the assurance lab can work to set up a minimum set of assurance requirements (which may not be mutually exclusive) rather than picking "winners" or "losers" where different approaches exist. The goal is to be collectively exhaustive to ensure all elements of a minimum standard are captured.

There is also no need to reinvent the wheel. Rather, it is a matter of finding pieces already out there and assembling them, filling in gaps where necessary. It is important to orchestrate the sequence of processes to get to the result, namely an assurance lab with various ecosystem components in place with standardization.

4.1.5 Process and Engagement

One challenge for an assurance lab is getting tools and datasets into the same analytical environment. There could be thousands of data providers, each of which has metadata that describes caveats about their datasets. Legal templates can facilitate the process. Furthermore, privacy-preserving AI technologies offer possibilities in which neither the data provider nor the algorithm provider needs to share data or intellectual property. Different test platforms may be required to accommodate nuances in different medical record systems and underlying data representation. For all of these, engagement in creating standard processes will be critical. The result can be a standard set of reports, potentially via data/model card, so that the user knows that every time one receives the tool from any entity, one will receive a report with a consistent format. This would enable information (such as that contained in a model card) to be entered into a registry. Various levels of information can be provided, and there can be different levels of transparency regarding the results obtained. For example, certain proprietary pieces of information as well as certain performance metrics, especially in the initial stages, may be made available only to certain users. This standardization is useful for incentivizing an ecosystem because commercial providers know what to expect. The key is to engage various stakeholders on the type of information needed, potential metadata to share, and potential users of the information generated by an assurance lab.

4.2 Institutionalizing Trustworthy AI Systems

There are several prerequisite components for institutionalizing trustworthy AI systems, and Assurance labs and associated technical assistance services can help with these tasks. The relevant prerequisites are seen in a number of frameworks such as Trustworthy AI Executive Order (EO) 13960 (50), the U.S. White House Office of Science and Technology Policy Blueprint for an AI Bill of Rights (51), the World Health Organization's Ethics and Governance of Artificial Intelligence for Health (52), the Organization for Economic Co-operation and Development (OECD) Tools for Trustworthy AI (53), frameworks developed from the perspective of insurers (54), industry and academic-derived principles, and U.S. state-level efforts (55).

There are several common themes. The first is to create an inventory or registry of various models/tools in the system. The second is to define which types of models from the inventory are subject to which guidelines (automated algorithms with higher levels of autonomy typically have more stringent monitoring) (56,57). The third theme is to define organizational structures, such as who is responsible for overseeing trustworthy AI systems, and for responding to requests in governance processes. Currently, little standardization exists. An assurance standards guide could help define successful oversight and governance.

Once organizational structures and oversight processes are established, then there is a basis for creating an established set of maturity levels against which health systems can be evaluated. In this context, there needs to be a floor or a minimum level of functionality that health systems should be able to perform toward enabling trustworthy AI. With a predictive model, there should be a person who is responsible to evaluate and ensure that tools do not have a disparate impact (e.g., the minimum standard set by the California Attorney General). In the Trustworthy AI EO, federal agencies are called on to certify that all applications meet a minimum set of nine principles or retire the application (50).

To ensure that the AI tools used by health systems possess these elements, an opportunity exists to specify who tests and when they test. Therefore, in addition to assurance standards, there may be a need for adjudicating bodies, and such tests may represent something that is certifiable, thus promoting confidence in such tools. The result is ongoing monitoring to ensure continued trustworthy AI, facilitated by testing, evaluation, and/or assurance bodies.

4.3 Energizing a Coalition of the Willing

There are several actions that can help move toward an assurance standards guide and beyond it to a roadmap with timelines, which can identify priorities and catalyze action. It also helps bring together a critical mass of the willing and creates a "fear of missing out" atmosphere. In designing the roadmap and timeline, it is important not to instill or exacerbate existing digital divides.

There exists a potential opportunity between CHAI, this "coalition of the willing", and the National Academy of Medicine to collaborate. This can be done by both codifying best practices and the corresponding "code of conduct" for AI. A consensus publication will certainly help move the field forward, ideally driven by public comment periods in which people can reflect and comment on the commentary paper that is produced. There is also a need to go beyond papers to actual practical code and software.

To foster an environment where an assurance standards guide and tools are deployed, we must examine various incentive structures and policies surrounding these. Incentives shape behavior, sometimes implicitly. A compelling business case for putting in the effort to build and coalesce around a national standard is needed. Such a standard should not be rigid, but rather one that is living and updated over time as new technologies and situations arise.

Finally, engagement from the beginning is key, from the design level through the release level. The assurance standards guide should allow the end users to better comprehend what is being disseminated to them as well as provide auxiliary information via a registry of tools and evaluation rubrics. Education of the community of stakeholders would include generating documentation, and other materials to inform, maintain, and receive feedback constantly from those tools that are being deployed. These could not only include purpose-designed healthcare AI tools but leverage ingenuity from outside healthcare as well.

Moving forward requires getting beyond the idea of one-way monitoring to ensure that the community can collectively learn and then change practice quickly. This will involve a national, cohesive community of leaders and people who can move the field forward. Through convening of stakeholders, CHAI can help move the field forward toward an assurance standards guide and associated frameworks to foster a community that adopts it.

References

1. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence. 2021 Mar;3(3):199–217.
2. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020 Apr;369:m1328.
3. Arnold M, Bellamy RKE, Hind M, Houde S, Mehta S, Mojsilović A, et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM J Res Dev. 2019 Jul;63(4/5):6:1-6:13.
4. Richards J, Piorkowski D, Hind M, Houde S, Mojsilović A. A Methodology for Creating AI FactSheets. 2020. Available at: https://arxiv.org/pdf/2006.13796.pdf
5. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2019. p. 220–9. (FAT* '19).
6. Breck E, Cai S, Nielsen E, Salib M, Sculley D. The ML test score: A rubric for ML production readiness and technical debt reduction. In: 2017 IEEE International Conference on Big Data (Big Data). 2017. p. 1123–32.
7. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020 Sep;26(9):1351–63.
8. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep;26(9):1364–74.
9. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015 Jan;162(1):W1-73.
10. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014 Oct;11(10):e1001744.
11. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019 Jan;170(1):51–8.
12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015 Oct;351:h5527.
13. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022 May;28(5):924–33.

14. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020 Sep;26(9):1320–4.
15. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, Young A, Jelovsek JE, O'Brien C, Parrish AB, Elengold S, Lytle K, Balu S, Huang E, Poon EG, Pencina MJ. A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc. 2022 Aug 16;29(9):1631-1636. doi: 10.1093/jamia/ocac078.
16. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012 May;98(9):683–90.
17. Wang HE, Landers M, Adams R, Subbaswamy A, Kharrazi H, Gaskin DJ, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. J Am Med Inform Assoc. 2022 Jul;29(8):1323–33.
18. Dankwa-Mullan I, Scheufele EL, Matheny ME, Quintana Y, Chapman WW, Jackson G, et al. A Proposed Framework on Integrating Health Equity and Racial Justice into the Artificial Intelligence Development Lifecycle. J Health Care Poor Underserved. 2021;32(2):300–17.
19. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016 Dec;18(12):e323.
20. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020 Dec;27(12):2011–5.
21. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med. 2020 Mar;3:41.
22. Tabassi E. AI Risk Management Framework. Gaithersburg, MD: National Institute of Standards and Technology; 2023.
23. 14:00-17:00. ISO/IEC TR 24368:2022 [Internet]. ISO. [cited 2023 Mar 7]. Available from: https://www.iso.org/standard/78507.html
24. 14:00-17:00. ISO 26000:2010 [Internet]. ISO. 2021 [cited 2023 Mar 7]. Available from: https://www.iso.org/standard/42546.html
25. Wornow M, Gyang Ross E, Callahan A, Shah NH. APLUS: A Python library for usefulness simulations of machine learning models in healthcare. J Biomed Inform. 2023 Mar;139:104319.
26. Global Approach to Software as a Medical Device. FDA [Internet]. 2022 Sep 27 [cited 2023 Mar 16]; Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/global-approach-software-medical-device
27. 14:00-17:00. ISO 9000:2015 [Internet]. ISO. 2020 [cited 2023 Mar 7]. Available from: https://www.iso.org/standard/45481.html
28. Hojo T. Quality Management Systems - Process Validation Guidance [Internet]. Global Harmonization Task Force; 2004. Available from: https://www.imdrf.org/sites/default/files/docs/ghtf/final/sg3/technical-docs/ghtf-sg3-n99-10-2004-qms-process-guidance-04010.pdf
29. Tartal J. Quality System Regulation Process Validation [Internet]. FDA; 2015. Available from: https://www.fda.gov/media/94074/download

30. General Principles of Software Validation; Final Guidance for Industry and FDA Staff.
31. 14:00-17:00. ISO/IEC TS 5723:2022 [Internet]. ISO. [cited 2023 Mar 7]. Available from: https://www.iso.org/standard/81608.html
32. 14:00-17:00. ISO 14224:2016 [Internet]. ISO. [cited 2023 Mar 22]. Available from: https://www.iso.org/standard/64076.html
33. Aim-ahead [Internet]. Available from: https://datascience.nih.gov/artificial-intelligence/aim-ahead
34. Schulam P, Saria S. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. In: Chaudhuri K, Sugiyama M, editors. Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. PMLR; 2019. p. 1022–31. (Proceedings of Machine Learning Research; vol. 89).
35. 35. AI assurance guide [Internet]. AI assurance guide. [cited 2023 Mar 22]. Available from: https://cdeiuk.github.io/ai-assurance-guide
36. Center for Devices and Radiological Health. Artificial Intelligence and Machine Learning in Software as a Medical Device [Internet]. U.S. Food and Drug Administration. FDA; Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device
37. Saria S, Subbaswamy A. Tutorial: Safe and Reliable Machine Learning [Internet]. ACM FAccT Conference; 2019 Feb 22. Available from: https://www.dropbox.com/s/sdu26h96bc0f4l7/FAT19-AI-Reliability-Final.pdf?dl=0
38. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S. The Clinician and Dataset Shift in Artificial Intelligence. N Engl J Med. 2021 Jul 15;385(3):283-286. doi: 10.1056/NEJMc2104626.
39. Playbook [Internet]. The University of Chicago Booth School of Business. [cited 2023 Mar 22]. Available from: https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias/playbook
40. Lu J, Sattler A, Wang S, Khaki AR, Callahan A, Fleming S, et al. Considerations in the reliability and fairness audits of predictive models for advance care planning. Front Digit Health. 2022 Sep;4:943768.
41. Silcox C, Dentzer S, Bates DW. AI-enabled clinical decision support software: A "trust and value checklist" for clinicians. NEJM Catal Innov Care Deliv. 2020 Nov;1(6).
42. Rojas JC, Fahrenbach J, Makhni S, Cook SC, Williams JS, Umscheid CA, et al. Framework for Integrating Equity Into Machine Learning Models: A Case Study. Chest. 2022 Jun;161(6):1621–7.
43. Roski J, Maier EJ, Vigilante K, Kane EA, Matheny ME. Enhancing trust in AI through industry self-governance. J Am Med Inform Assoc. 2021 Jul;28(7):1582–90.
44. Halamka J. The algorithmically underserved need our attention [Internet]. Mayo Clinic Platform. 2022. Available from: https://www.mayoclinicplatform.org/2022/11/15/the-algorithmically-underserved-need-our-attention/
45. HL7.FHIR.UV.SMART-APP-LAUNCH\textbackslashbackend services - FHIR v4.0.1 [Internet]. Available from: https://www.hl7.org/fhir/smart-app-launch/backend-services.html
46. Jung K, Kashyap S, Avati A, Harman S, Shaw H, Li R, et al. A framework for making predictive models useful in practice. J Am Med Inform Assoc. 2021 Jun;28(6):1149–58.

47. IBM Garage Methodology [Internet]. Use an IT maturity model. Available from: https://www.ibm.com/garage/method/practices/think/it-maturity-model/
48. How we measure AI readiness [Internet]. Available from: https://coe.gsa.gov/2020/10/28/ai-update-2.html
49. Center for Devices and Radiological Health. Digital Health Software Precertification (Pre-Cert) Pilot Program [Internet]. U.S. Food and Drug Administration. FDA; Available from: https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-pilot-program
50. Executive Office of the President. Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. Vol. 85, Federal Register. 2020. p. 78939–43.
51. White House. Blueprint for an AI bill of rights: Making automated systems work for the American people. Nimble Books; 2022.
52. Ethics and governance of artificial intelligence for health [Internet]. World Health Organization; 2021. Available from: https://www.who.int/publications/i/item/9789240029200
53. OECD. Tools for trustworthy AI. Organisation for Economic Co-Operation and Development (OECD); 2021 Jun. (OECD Digital Economy Papers).
54. Smith LT, Bochanski SJ. Avoiding Unfair Bias in Insurance Applications of AI Models. 2022;
55. Attorney general bonta launches inquiry into racial and ethnic bias in healthcare algorithms [Internet]. State of California - Department of Justice - Office of the Attorney General. 2022. Available from: https://oag.ca.gov/news/press-releases/attorney-general-bonta-launches-inquiry-racial-and-ethnic-bias-healthcare
56. Automated vehicles for safety [Internet]. NHTSA. Available from: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety
57. Five levels that will define the future of autonomous enterprises [Internet]. 2021. Available from: https://www.ibm.com/cloud/blog/five-levels-that-will-define-the-future-of-autonomous-enterprises