BLUEPRINT FOR TRUSTWORTHY AI IMPLEMENTATION GUIDANCE AND ASSURANCE FOR HEALTHCARE

COALITION FOR HEALTH AI

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: UC Berkeley, Duke Health, Google, Johns Hopkins University, Mayo Clinic, MITRE, Microsoft, Optum, Stanford Medicine, SAS, UC San Francisco, UC Berkeley as well as US Federal Observers from the 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.

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.

CHAI
Coalition for Health AI

TABLE OF CONTENTS

PURPOSE

The use of artificial intelligence (AI) in healthcare offers enormous potential for accelerating clinical research and improving the quality and delivery of healthcare. However, a growing body of evidence shows that such tools can perpetuate and increase harmful outcomes and bias absent a framework focusing on health impact, fairness, and equity across all populations, including those underserved and under-represented. When standard guidelines are unharmonized or poorly understood, the potential for distrust for AI is cultivated within the community. Moreover, there is currently an inability to easily judge the robustness of algorithms on relevant data and evaluate health systems developing as well as deploying these tools.

This report is the result of bringing together experts across over a dozen institutions, through The Coalition for Health AI (CHAI), to work through issues that would need to be addressed for enabling trustworthy AI in healthcare. Specifically, this work summarizes those discussions as a step toward a blueprint for implementation guidance on trustworthy AI in healthcare. This blueprint seeks to enable AI that harmonizes standards and reporting for health AI and educate end users on how to evaluate these technologies to drive their adoption. Furthermore, the goal of this blueprint is to enable guidelines regarding an ever-evolving landscape of health AI tools to ensure high quality care, increase credibility within the community, and meet healthcare needs.

BACKGROUND

Since healthcare applications can have a critical impact on patient well-being and outcomes, AI in healthcare must meet a higher standard than AI in other fields. However, some recently published analyses expressing concerns on some AI-based algorithms have raised questions about their safety, efficacy, and equity. For example, a study that assessed 415 published deep-learning and machine learning (ML) models for diagnosing COVID-19 and predicting patient risk from medical images, such as chest x-rays and chest computed tomography scans found that none of them were fit for clinical use [1]. Further, a review of 232 diagnostic and prognostic algorithms for COVID-19 found none of them were fit for clinical use [2, 3].

Failure to provide information about AI system characteristics, behavior, efficacy, and equity can limit trust, acceptance, and proper use of these systems. This information is critical to the safe use of AI systems. In recent years, multiple resources have been published for general use that characterize ML systems including FactSheets [4, 5], Model cards [6], and ML Test Score [7]. Several resources designed to characterize clinical AI-based systems are intended to support assessment of clinical trial protocols, such as SPIRIT-AI [8] and CONSORT-AI [9], or assessment of published studies such as TRIPOD [10], CHARMS [11], PROBAST [12], STARD [13], and DECIDE-AI [14, 15]. Still others are intended to assist researchers and/or developers determining the appropriateness of models for incorporation into biomedical or clinical applications including MI-CLAIM [16], ABCDS [17], Guidelines [18], Risk Prediction Model [19, 20] and Bias

Checklist [21, 22]. However, few independent, scalable guides exist on assessments of AI-based clinical systems for health systems, consumers, and end users.

This work has brought together a collaboration across a number of 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 more widespread adoption. Some of the institutions that have already published guidelines (i.e., Duke, Stanford, Johns Hopkins) are part of this work [23, 24]. While there are current efforts to develop core ingredients for AI/ML for specific medical applications like cardiac software and medical devices [25], the clinical AI/ML community would benefit from an approach that could be applied to AI-based clinical algorithms for various uses (e.g., diagnostic, prognostic) and clinical subdomains (e.g., ontology, cardiology).

From past experiences, it has become evident that it is hard to build ecosystems when multiple approaches are left in the wild to bloom without at least minimal consensus-based standardization. Thus, it is important to assemble a guiding coalition and that they agree on the canonical structure for implementation guidance. It is not to say that the guidance needs to be inflexible. Over time, implementation guidance can change as needed. The key is to have a group that builds this consensus together, so there are not hundreds of approaches around that prevent developers and others from knowing what to adopt and how. Such a group, process, and guidelines developed should include input from various stakeholder groups, such as those listed below. This report is the beginning of that effort. By summarizing the culmination of about a year of work via industry, academia and government participants, this work explores the parameters for the guidance, the guardrails, the best practices, the governance to help ensure ethical 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.

CHAI
Coalition for Health AI

KEY ELEMENTS FOR TRUSTWORTHY AI IN HEALTHCARE

In working toward implementation guidance for ethical health AI, several elements were defined and explored including bias, equity, fairness, testability, usability, safety, transparency, reliability, and monitoring, described as follows.

Bias, Equity, and Fairness

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 can achieve their health potential regardless of specific group membership. With regards to health AI, this means ensuring that AI is not disadvantageous to a specific group. 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.

Leveraging health equity by design involves looking with intentionality at promoting health equity [26]. This means there is a need to explicitly define equity goals. As part of this process, there is a need to include all the various stakeholders and community members, throughout the entire lifecycle of the AI tool [27]. This involves everything from data collection to deployment as well as behavioral considerations for algorithmic-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.). So, the key is to make an informed decision considering the inherent tradeoffs with other goals. This ensures 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. Yet, it is not possible to completely pre-define the set of measures and processes that are required for specific settings. Establishing frameworks and checklists can help guide decisions. Overall, there should be multiple checkpoints for different stages of evaluation and continual monitoring to account for dynamic changes in the population, user behavior, and algorithmic performance. For example, it is not just about looking at the algorithm itself and what it is doing but also evaluating how it works. The algorithm may use proxies that are correlated with certain variables such as race, which might not be known unless carefully considered together. Monitoring structures need to be set up that include 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. 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.

Governance is also key to overseeing mitigation strategies. Establishing who governs and how governing occurs, in a standardized way, can help mitigate risks. This requires a multidisciplinary team to establish processes and measures for bias.

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 [28]. There are three aspects to this idea. First, some patients may be underserved because they do not have data recorded/available. This may be 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 [29, 30]. It may also be that the patient explicitly decided to decline making their data available or simply choose not to complete forms/information fully. The second aspect is that some patients may be from populations without enough data to evaluate performance of models with confidence. For example, an American Samoan patient may be algorithmically underserved when there is too small of sample size available in the training set. The final 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. 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 [31].

Testability

In this report, testability refers to the extent to which an ML algorithm's performance can be verified as satisfactory. The context of how the AI is being used plays an important role in determining testing needed. Testability requires strong contextual understanding of the model and its intended use including where, why, and how.

The model's life cycle needs to be looked at when doing testing. It is not only important to look at the training phase or when in deployment. Testing needs to be done throughout the lifecycle, including from conception to after deployment. In fact, testability as part of the model lifecycle is not a single step or even discrete individual steps but should be part of a continuous process. (See the monitoring section as well.)

When testing, it is important to establish a reasonable baseline on which the AI system will be involved. The status quo, against which the model is compared, may be

difficult to fully define. However, working to define one can help to capture value of the model and potential return on investment (ROI), thereby increasing adoption rate. It should be recognized that certain study designs may have more weight. Randomized controlled studies are considered the "gold standard" but are not always possible. Thus, capturing study design used, performance, and sample size information may be useful to assess level of evidence, and an implementation guide could include notes on the level of evidence and type of study designs used to assess model impact. Furthermore, capturing documentation, reproducible methods, and accessible code are important for multi-site testing- and can also be embedded as pointers in a schema for the relevant resources.

During each phase of the life cycle, there are different common issues that arise in testing. An implementation 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 implementation guide can help capture relevant information about the testing and its results across the lifecycle. In addition, there are a number of policy issues to be defined including what type of AI tool could be testable, who is reasonable for testing it, and how to incentivize/enforce regular testing in the model lifecycle. Finally, it would be helpful to have guidance on health systems' responses if a model fails testing as well as standard procedures that should be done as potential next steps.

Usability

Usability is another key area. It is defined here as denoting the quality of the user's experience, including effectiveness, efficiency, and satisfaction, when using an algorithm's output. Usability has several important factors that need to be considered. The first is context: usability is heavily dependent on a model's context. The second is end user and/or patient perspective. There need to be optimal ways for patient perspectives to be incorporated into usability design. It is important that end users be involved as contributors to the assessment of usability. Simplicity is another variable. Excessive complexity decreases usability. Other tradeoffs aside, a simpler model is often easier to use.

Workflow considerations are important for usability. For non-emergency notifications, non-intrusive alerts are preferable as 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 implementation guide, there are some key items to address. This includes delineating how usability is measured and by whom. Another area to explore is defining how patient perspectives can best be incorporated into usability design and process itself. Finally, there is a need to empower designs that help end users who may not have data science training and understand a model's output.

Safety

Safety is a vital factor. Here, this refers to preventing worse outcomes for the patient, provider, or health system from accruing as a result of use of an ML algorithm. A lack of oversight can make any model unsafe. Looking at potential role 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. There may also be downstream impact that may not be readily known or available in the development process for the AI model.

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

Transparency

The transparency of an AI model implies interoperability, traceability, and explainability. 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 has to be continuously evaluated and addressed throughout the AI system lifecycle. Transparency is enabled when criteria involving the underlying datasets, models, and stakeholders 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, 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, 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 that the model has been trained on. 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 the AI system lifecycle and be able to adapt to feedback.

For stakeholders, considerations around the audience are critical. For example, certain types of information should be provided for technical versus non-technical audiences. There should be clear communication regarding tradeoffs made by the model. As stated in the bias and equity section, diverse multidisciplinary teams should be involved throughout the model lifecycle.

An implementation guide can help address transparency when there are multiple datasets and/or models that 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 uses, model cards have been used for this purpose. Like a nutrition label, model cards can be designed to provide specific information to increase transparency based 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 grandfathered, given certain time to follow proposed policies, retired automatically, or given certain 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.

Reliability

Reliability captures the ability of an AI model/tool to perform its required function under stated conditions. Key facets of reliability include failure prevention, dataset shifts, and workflow integration. The goal in fail prevention is to minimize the likelihood of failure. One of the reasons why 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, and measurements of reliability, should capture the role the tool plays in the broader, human-centric clinical workflow. Reproducibility is an important related factor for ensuring outcomes are consistent across sites and thus reliability for the entire heath system in question. AI/ML is particularly sensitive to variations in hardware and software versions. Like other variables (e.g., see testability), reliability is a factor that needs to be considered across the continuum of the life cycle. Thus, an implementation guide should capture information such as metrics for reliability, its embedded workflow, versioning, expected datasets, and guardrails for drift [32].

Monitoring

Monitoring involves the ongoing surveillance of an AI tool with the goal of raising an alarm when shifts in the input data, tool outputs, or use behavior are detected. In monitoring, it is important to identify failures and vulnerabilities quickly so that negative effects are minimized. Central reporting is important so that all sites can learn from the experiences of others. This is critical in rare incidents: where individual sites may not realize a

pattern but combining information across sites can enable faster event detection. When monitoring, it is important to consider backwards compatibility. Model updates should not reduce the quality of human-AI collaboration.

An implementation guide can help define the type of information that is relevant to include in the specification for a machine learning model and its encompassing tool. When monitoring, it is useful to pre-define what actions will be taken based on the monitoring results. Having a predefined protocol can be useful for when unintended model behaviors arise, especially in real-time, high-volume cases where decisions that could affect many end users need to 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.

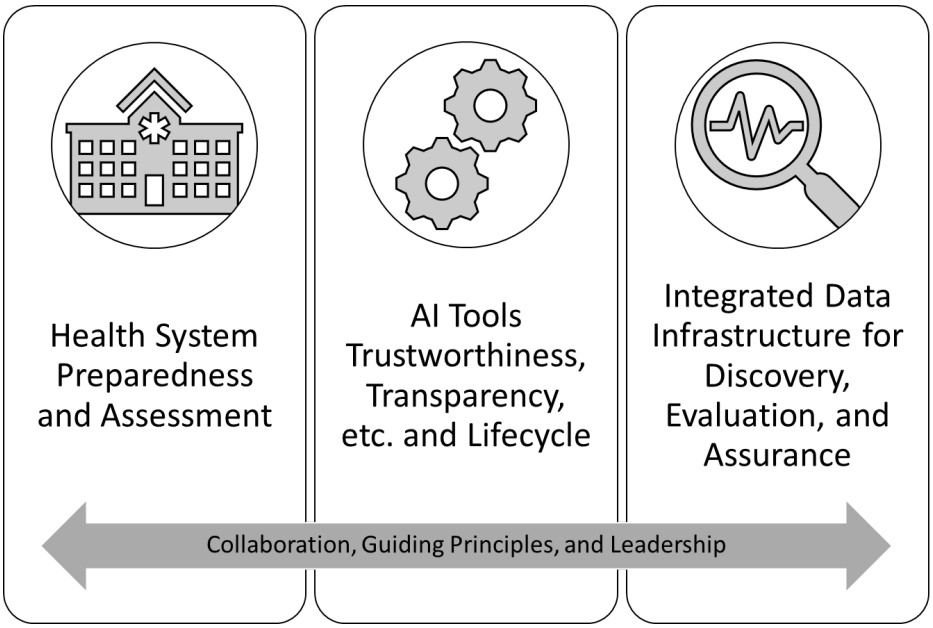
Monitoring is a factor that needs to be looked at across various settings. Metrics may be monitored upon the live deployment of a system as a whole, but also focus on monitoring algorithm-level issues and workflow-level reliability. Guidance is also needed for how often models need to be updated and systems maintained.

NEXT STEPS

Every institution can have different flavors of AI tools. Yet, there is a need to use the same principles to build them and facilitate their use. Through an assurance accreditation 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, there may be the resources to make this happen now. Other small and rural resource-constrained health systems may not have the resources to do it on their own. So, there may also be a need for an advisory body to move the field forward with these entities as well and ensure equity so that, for a given patient, ethical AI would not depend on where they live or with which health system they are interacting.

Below are the key pillars for how an AI assurance, evaluation, and discovery lab can help achieve results through health system preparation, AI tool use, and infrastructure for enabling trustworthy AI.



Setting up an Assurance Lab and Advisory Services

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

A Shared Definition of Value

With negative margins being commonplace for many health systems, it is important to ensure a clear value proposition for the patient and the organization for deploying AI solutions. Thus, it is recommended to begin with the value proposition evaluation. Demonstrating the value framework can get the decision makers excited, and then the other elements can be done to lead to better patient outcomes and ROI. Governance requirements, bureaucratic processes, and best practices come secondary to value in terms of getting initial buy in. Thus, one goal for implementation guidance, including potential consultation services, would be to serve as an enabler of value for health systems and their patients, which also includes ensuring policies do not deplete that value. For example, there is 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 potential projects based on potentially virtual model deployments with the workflow. This can include structured checklists that require the submitter to think through how the potential solution would impact the organization and how it would generate value.

Initial attempts to understand the value proposition are just the beginning. There is a need for a structured intake process for candidate use-cases (where a model would drive a clinical care workflow) based on virtual model deployments that calculate achievable utility via simulating several days of care workflows [ref: https://doi.org/10.1093/jamia/ocaa318].

Enabling such analyses will require structured checklists that require the submitter to think through how the potential use of the AI tool would impact workload in the organization and who would need to do what to realize the value.

Value also needs to be throughout the proposed ecosystem and for its development. This includes incentivizing developers to participate and make sure there is value 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 utilized. Several maturity models exist [33, 34] but need 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, similar to the Food and rug Administration (FDA) pre-certification pilot [35] that sought to establish criteria for the industry developers, rather than the device/tool itself. For models, setting up guardrails with potential intervention points may be another option, similar to the FDA AI software as a medical device (SaMD) guidance [36].

Registries

One approach to empower patients is to set up a registry, where AI tools are registered. This may be like clinicaltrials.gov for clinical trials, but for predictive algorithms. It could be done at an institutional level as well. The key is local implantation of a uniform national framework. Patients could look up what is available in their own facility and see the tools. Care providers and AI tools developers can compare algorithms and analytic options by reviewing the registry. It would be useful to look at all the model cards and other publications that propose nutrition-like labels for AI models.

Providers with access to patient 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 pulled down 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 can be captured as metadata in a registry. This metadata could be used to help determine when the underlying algorithms may be suitable for a particular patient, facilitating precision medicine.

To build such a registry, the technology and policies need to be developed to enable it to be used as part of an ecosystem. An assurance accreditation lab can help ensure 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.

Legal Agreements and Sandboxes

To set up such an assurance accreditation lab and associated technical assistance service, there needs to be agreement on a set of criteria necessary to be reported on an algorithm to perform such an evaluation. It also necessitates willing institutions. A number of 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 platform can help ensure long-term reliability as well.

There needs to be standardization to enable a marketplace where data providers and algorithm developers can come together for validation. This includes creating a template-based, check box 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 taken up by legal negotiations can be saved. Furthermore, having standard schemas for data will speed up the process so that data can easily be processed by an assurance accreditation lab and technical assistance service. Approaches to setting up such agreements on sharing the data and creating sandboxes have already been done 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 allows.

Independence

One of the requirements for having an assurance accreditation lab is ensuring its independence. Doing so can build trust among potential stakeholders and users and enable collaborative work in the pre-competitive space. Without having conflicts of interest, the assurance accreditation 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 together, filling in gaps where necessary. It is important to orchestrate the sequence of processes to get to the end result, namely an assurance accreditation lab with various ecosystem components in place with standardization.

Process and Engagement

One of the challenges for an assurance accreditation lab is getting tools and datasets into the same analytical environment. There could be thousands of data providers, each of which have metadata that describes the caveats about their data sets. Legal templates can help facilitate the process. Furthermore, privacy preserving AI technologies offer possibilities where neither data provider nor algorithm provider need to share data/IP. There may also need to be different test platforms for 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 is going to get back a similar kind of report. This standardization is useful to incentivize an ecosystem as then commercial providers know what to expect. This would enable information like a model card to be entered into a registry. Various levels of information can be provided, and there can be different levels of transparency on the results obtained. For example, certain proprietary pieces of information as well as certain performance metrics, especially in

initial stages, may be made available only to certain users. 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 accreditation lab.

Institutionalizing Trustworthy AI Systems

An accreditation lab and associating technical assistance service can help in institutionalizing trustworthy AI systems. For systems to institutionalize trustworthy AI systems, there are several prerequisite components. These are seen in a number of frameworks such as Trustworthy AI Executive Order (EO) 13960 [37], U.S. AI Bill of Rights [38], the World Health Organization (WHO) 39], Organisation for Economic Co-operation and Development (OECD) AI [40], industry and academic-derived principles, and states' efforts [41].

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 [42, 43]. The third theme is to define organizational structures such as who is responsible for overseeing the trustworthy AI systems and for responding to requests in governance processes. Currently, there is little standardization. An implementation 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 disparate impact. (e.g., 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 [44].

To ensure health systems' AI tools possess these elements, an opportunity exists to specify who tests and when they test. So, in addition to implementation guidance, there may be a need to have adjudicating bodies that decide who is testing and how, and such tests may represent something that is certifiable, thus promoting confidence in such tools. The result is ongoing monitoring to ensure continued ethical AI, facilitated by testing, evaluation, and/or accrediting bodies.

Energizing a Coalition of the Willing

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

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

To foster an environment where an implementation guide and tools are deployed, there is a need to look at various incentive structures and policies surrounding these. Incentives shape behavior, sometimes implicitly. There needs to be a compelling business case for putting in effort to build and coalesce around a national standard. Such a standard should not a be rigid standard, 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 as well as at the release level. The implementation guide should allow the end users to better comprehend what is being disseminated to them as well as providing auxiliary information via a registry of tools and evaluation rubrics. Education of the community of stakeholders would include generating documentation, and other materials to both inform, maintain, and receive feedback constantly from those tools that are being deployed.

Moving forward requires getting beyond the idea of one-way monitoring to ensuring 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 implementation guide and associated frameworks to foster a community that adopts it.

CHAI
Coalition for Health AI

REFERENCES

1. Roberts, M., Driggs, D., Thorpe, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0
2. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal *BMJ* (Clinical Research ed.). 2020 Apr;369:m1328. DOI: 10.1136/bmj.m1328. PMID: 32265220; PMCID: PMC7222643.
3. M. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020 Apr 7;369:m1328. doi: 10.1136/bmj.m1328. Update in: BMJ. 2021 Feb 3;372:n236. Erratum in: BMJ. 2020 Jun 3;369:m2204. PMID: 32265220; PMCID: PMC7222643
4. M. Arnold et al., "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," in *IBM Journal of Research and Development,* vol. 63, no. 4/5, pp. 6:1-6:13, 1 July-Sept. 2019, doi: 10.1147/JRD.2019.2942288
5. Richards, J., Piorkowski, D., Hind, M., Houde, S., Mojsilović, A. A Methodology for Creating AI FactSheets. *arXiv*:2006.13796, June 28, 2020
6. Mitchell, M. et al. Model cards for model reporting. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency 220–229 (2019) doi:10.1145/3287560.3287596
7. 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)*. 2017:1123-1132. doi:10.1109/BigData.2017.8258038
8. Cruz Rivera, S., Liu, X., Chan, AW. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 26, 1351–1363 (2020). https://doi.org/10.1038/s41591-020-1037-7
9. Liu X, The SPIRIT-AI and CONSORT-AI Working Group, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x
10. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015 Jan 6;162(1):W1-73. doi: 10.7326/M14-0698. PMID: 25560730.
11. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. (2014) Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med* 11(10): e1001744. https://doi.org/10.1371/journal.pmed.1001744
12. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S; PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med*. 2019 Jan 1;170(1):51-58. doi: 10.7326/M18-1376. PMID: 30596875

13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HC, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015 Oct 28;351:h5527. doi: 10.1136/bmj.h5527. PMID: 26511519; PMCID: PMC4623764.
14. Vasey, B., Nagendran, M., Campbell, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 28, 924–933 (2022). https://doi.org/10.1038/s41591-022-01772-9
15. Vasey B, Nagendran M, Campbell B, Clifton D A, Collins G S, Denaxas S et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI *BMJ* 2022; 377 :e070904 doi:10.1136/bmj-2022-070904
16. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z, Yu B, Butte AJ. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020 Sep;26(9):1320-1324. doi: 10.1038/s41591-020-1041-y. PMID: 32908275; PMCID: PMC7538196
17. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207
18. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res*. 2016 Dec 16;18(12):e323. doi: 10.2196/jmir.5870. PMID: 27986644; PMCID: PMC5238707
19. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012 May;98(9):683-90. doi: 10.1136/heartjnl-2011-301246. Epub 2012 Mar 7. PMID: 22397945.
20. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012 May;98(9):691-8. doi: 10.1136/heartjnl-2011-301247. Epub 2012 Mar 7. PMID: 22397946.
21. H Echo Wang, Matthew Landers, Roy Adams, Adarsh Subbaswamy, Hadi Kharrazi, Darrell J Gaskin, Suchi Saria, A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models, *Journal of the American Medical Informatics Association*, Volume 29, Issue 8, August 2022, Pages 1323–1333, https://doi.org/10.1093/jamia/ocac065
22. Correction: A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models, *Journal of the American Medical Informatics Association*, Volume 29, Issue 9, September 2022, Page 1656, https://doi.org/10.1093/jamia/ocac102
23. Tina Hernandez-Boussard, Selen Bozkurt, John P A Ioannidis, Nigam H Shah, MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care, *Journal of the American Medical Informatics*

Association, Volume 27, Issue 12, December 2020, Pages 2011–2015, https://doi.org/10.1093/jamia/ocaa088

24. Sendak, M.P., Gao, M., Brajer, N. et al. Presenting machine learning model information to clinical end users with model facts labels. *npj Digit. Med*. 3, 41 (2020). https://doi.org/10.1038/s41746-020-0253-3
25. Silcox, C, Dentzer, S., Bates, D.W. AI-Enabled Clinical Decision Support Software: A "Trust and Value Checklist" for Clinicians. *NEJM Catalyst*. 2021/09/09. doi: 10.1056/CAT.20.0212
26. Rojas JC, Fahrenbach J, Makhni S, Cook SC, Williams JS, Umscheid CA, Chin MH. Framework for Integrating Equity Into Machine Learning Models: A Case Study. *Chest*. 2022 Jun;161(6):1621-1627. doi: 10.1016/j.chest.2022.02.001. Epub 2022 Feb 7. PMID: 35143823; PMCID: PMC9424327.
27. Joachim Roski, Ezekiel J Maier, Kevin Vigilante, Elizabeth A Kane, Michael E Matheny, Enhancing trust in AI through industry self-governance, *Journal of the American Medical Informatics Association*, Volume 28, Issue 7, July 2021, Pages 1582–1590, https://doi.org/10.1093/jamia/ocab065
28. The Algorithmically Underserved Need Our Attention, *Mayo Clinic Platform Blog*, November 15, 2022. Available at: https://www.mayoclinicplatform.org/2022/11/15/the-algorithmically-underserved-need-our-attention/
29. Backend Services, Smart App Launch Implementation Guide, (v2.0.0: STU 2) based on FHIR R4. Available at: https://www.hl7.org/fhir/smart-app-launch/backend-services.html
30. CDS Hooks. HL7 International. Version 2.0 STU 2 Release, August 23, 2022 Available at: https://cds-hooks.hl7.org/
31. Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) Program, National Institutes of Health Office of Data Science Strategy. Available at: https://datascience.nih.gov/artificial-intelligence/aim-ahead
32. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food and Drug Administration. January 2021. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device
33. Bertoletti, A., Eeles, P. "Use an IT Maturity Model." *IBM Garage Methodology*. Available at: https://www.ibm.com/garage/method/practices/think/it-maturity-model/
34. "How We Measure AI Readiness," Artificial Intelligence Center of Excellence, IT Modernization Centers of Excellence, U.S. General Services Administration. October 28, 2020. Available at: https://coe.gsa.gov/2020/10/28/ai-update-2.html
35. The Software Precertification (Pre-Cert) Pilot Program: Tailored Total Product Lifecycle Approaches and Key Findings. U.S. Food and Drug Administration. September 26, 2022. Available at: https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-pilot-program
36. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food and Drug Administration. January 2021. Available at:

https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

37. Executive Order 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. Office of the President. December 3, 2020. Available at: https://www.federalregister.gov/executive-order/13960
38. Blueprint for an AI Bill of Rights. White House Office of Science and Technology Policy, October 4, 2022. Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/
39. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. License: CC BY-NC-SA 3.0 IGO. Available at: https://www.who.int/publications/i/item/9789240029200
40. OECD (2021), "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems", *OECD Digital Economy Papers*, No. 312, OECD Publishing, Paris, https://doi.org/10.1787/008232ec-en.
41. Bonta, R. "Attorney General Bonta Launches Inquiry into Racial and Ethnic Bias in Healthcare Algorithms." State of California Department of Justice State Attorney General Office, August 31, 2022. Available at: https://oag.ca.gov/news/press-releases/attorney-general-bonta-launches-inquiry-racial-and-ethnic-bias-healthcare
42. "The Road to Full Automation: Levels of Automation." U.S. Department of Transportation, National Highway Traffic Safety Administration. Available at: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety#the-topic-road-to-full-automation
43. Wang, R., Williams, P. "Five Levels That Will Define the Future of Autonomous Enterprises." *IBM Cloud*, July 1, 2021. Available at: https://www.ibm.com/cloud/blog/five-levels-that-will-define-the-future-of-autonomous-enterprises
44. Executive Order 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. Office of the President. December 3, 2020. Available at: https://www.federalregister.gov/executive-order/13960