PROVIDING GUIDELINES FOR THE RESPONSIBLE USE OF AI IN HEALTHCARE

COALITION FOR HEALTH AI
VIRTUAL WORKGROUP SESSION: TESTABILITY, USABILITY, AND SAFETY

JULY 13, 2022, 2-3:30PM ET

SUMMARY

This Virtual Workgroup Session was convened by the Coalition for Health AI to develop a collective understanding of the definitions, important considerations, and open questions for the concepts of testability, usability, and safety in health. With input and participation from a group of subject matter experts from healthcare and other industries, this session included a series of three lightning talk presentations and group discussions centered on preselected use cases. It also featured a set of breakout sessions that addressed the themes of testability, usability, and safety. The aim of this and other planned meetings is to develop a practical guide for implementing AI and ML tools in healthcare, one that establishes clear and appropriate guidelines and guardrails for the fair, ethical, and useful application of machine learning in healthcare settings.

INPUT AND FEEDBACK

We welcome feedback and input on the ideas presented here, on additional ideas and concepts, and on the future direction of work pertaining to testability, usability, and safety in health AI.

Input and feedback are requested via submission form on our website during a 30-day comment period, ending October 14, 2022.

OBJECTIVE

The objective for this Health AI Virtual Workgroup Session was to develop our collective understanding of definitions, important considerations, and open questions for the concepts of testability, usability, and safety in health AI.

LIGHTNING TALKS & USE CASES

To articulate key themes and ground discussion in real-world issues affecting healthcare and healthcare delivery, invited experts selected use cases from published reports that examined the development and deployment of algorithmic analytical tools in healthcare and other settings, and examined them in a series of brief lightning talks that were followed by focused discussions. The three use cases, which are being used throughout this series of talks, were selected to inform these discussions with real-world examples. They include:

1. Hospitals, providers, and insurance companies implementing patient-level prediction of all-cause 30-day hospital readmission using claims data or electronic health record (EHR) data[1];
2. A large health system implementing 12-month mortality estimates to support advanced care planning[2]; and
3. A machine learning algorithm being developed to triage, diagnose, and/or monitor for skin cancer using clinical or dermoscopic images of skin disease.[3]

LIGHTNING TALK 1: HEALTHY AI BETTER DATA TO BE ROBUST, PRIVATE & FAIR

Presented by Marzyeh Ghassemi, PhD (Institute for Medical Engineering and Science, Massachusetts Institute of Technology)

Despite many of the advances that have been made in the field of AI, if we want to achieve robust, private and fair machine learning, we need better data. Specifically, when working with embodied data – data derived from human bodies – we need to ensure more diverse datasets for research use in order to improve science and prevent medical harms.

A recent paper[4] examines a number of ML models that report performance at or above that of humans on a range of tasks over the human lifespan. However, upon examining the data contained in Table 3 of the paper, it's clear that there is a great deal of variation in sample size, ranging from a few hundred to hundreds of thousands. Furthermore, the metric of predictive performance, area under the curve (AUC), is a single number. How should this number be interpreted?

AI learns from human practice, which means that medical AI models are trained with existing data from historical practice. These data inevitably reflect inequities because doctors, like all humans, have biases. So, unfortunately, when we train AIs on all the data we have in large hospital systems, we are in effect training them to do as we do, not as we believe we should do or aspire to do. We are training them on examples of medical practice in which clinicians were upset, tired, or made mistakes. Nor are these data coded in a way that allows us to distinguish between what we do versus what we should do. Therefore both eventually show up in the representation. Furthermore, if that representation remains, the bias will persist even if the model is retrained on different data.[5] It's unclear how well known or widely understood this issue is in the larger AI community.

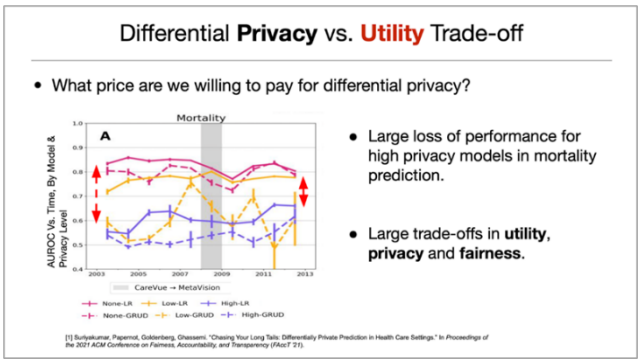
Making models that are healthy requires auditing healthcare systems to ensure that the practices we are using are healthier. We also must understand how we can deliver that information to clinicians in such a way that they can use it well. One frequently used method for this is to develop a decision-support checklist, which often depend on scores developed by domain experts. However, research has shown that for several clinical risk scores, risk was significantly over- or under-corrected for African American patients.[6]

Importantly, we should be aware of the potential for hidden issues in the data, given the layered proxies present in medical data. This can be countered at least in part by creating a fairness constraint for the training data, yielding a checklist that works well and is not biased.

Differential Privacy

Another issue to address is the conundrum of balancing utility, privacy, and fairness. Differential privacy is an approach that protects patients who have a combination of attributes that are uniquely identifiable. Although differential privacy is widely used, it often results in a significant loss of utility

– so much so that the model cannot be deployed.[7]



AI applications are built on finding and enforcing similarities. Whe privacy is added to the model, it removes the patients who are most different with respect to the larger group. In this example we saw that it changed the most helpful group training data for Black patients from Black patients to White patients. This is a problem that cannot be solved without more diverse datasets.

Does Biased AI Affect High-Stakes Decisions?

What happens when we give ML output to a doctor and the doctor uses it to make a decision? In a survey study that put volunteers in the position of having to make a crisis help-line assessment about whether to call for healthcare or police intervention for a particular person, they were initially given only minimal instructions to call the police if there was a threat of violence.[8] We then created an "evil" AI by training it with biased language. When the advice generated by this AI was provided in a prescriptive fashion, subjects (clinician or non-clinician) were more likely to call the police on Black or Muslim people. But when the same biased advice was given in a descriptive (not prescriptive) fashion, the study subjects were not more likely to summon the police on Black or Muslim persons. What this example shows is that we need to have this higher level ethical discussion about the entire pipeline from data collection, to defining outcomes, to developing and deploying the model. This is an ongoing process, and progress will require diverse data and diverse teams.

Key Discussion Points

- Evidence suggests that adherence to care and treatment guidelines improve patient outcomes while reducing variability of care across providers. Adhering to these guidelines tends to "bake in" a set of descriptors that in turn imply a prescription for action, but this comes as the result of significant amounts of research and training. Outside the world of AI, there is this constant tension between description and prescription.
- Within the world of AI, no one thinks about these tensions when building models – we don't think to tell users that when the model is deployed, the recommendation it generates should be phrased only in this one specific way, so that when it gives the wrong advice (as all models inevitably will at some point), the users don't over-anchor to that recommendation. The tendency toward automation bias – of uncritically accepting the model output – is potentially concerning, because in the case of healthcare decision-support models no one is validating all the different ways the information can be presented to the clinician or considering that that may bias toward certain actions.

LIGHTNING TALK 2: PRINCIPLES FOR EVALUATION OF CLINICAL DECISION-SUPPORT TOOLS

Presented by Michael Pencina, PhD (Duke AI Health/Department of Biostatistics & Bioinformatics, Duke University School of Medicine)

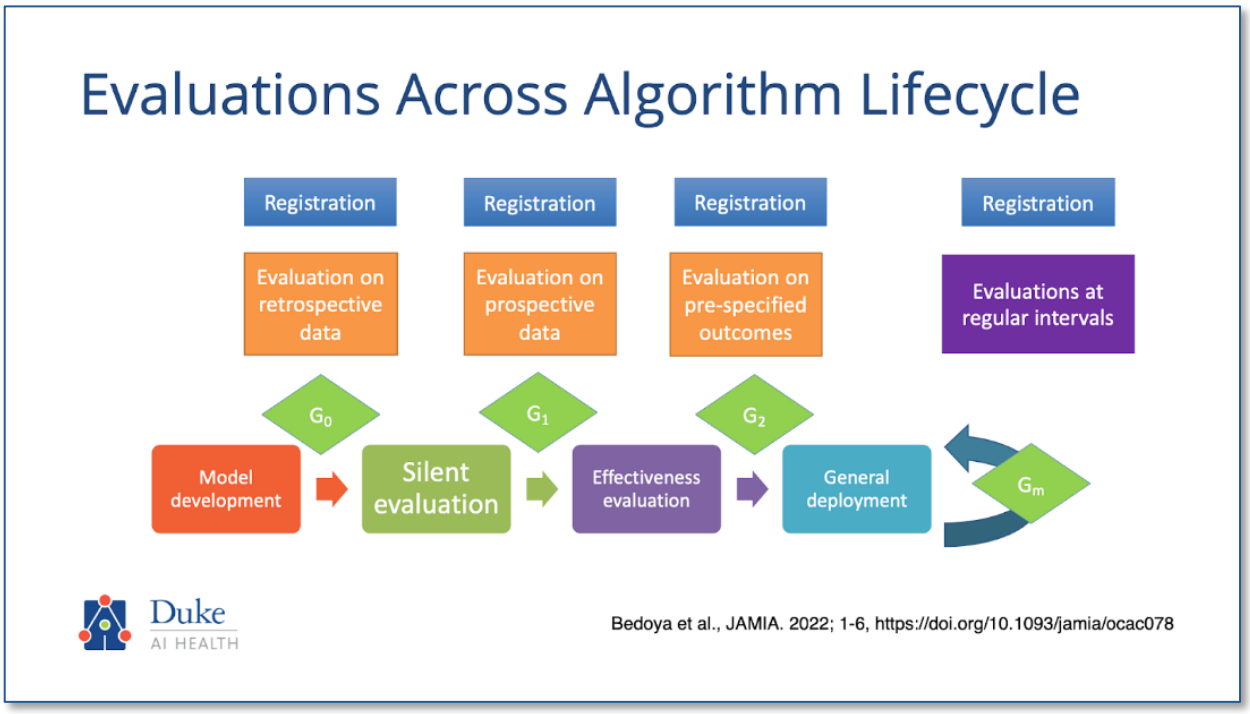
It is important to consider the evaluation of a clinical decision support (CDS) tool in the context of its application – what is the clinical story? Do you need this model, and if so, why? How will you use it? What decisions will you make based on its output? Without this rich context, some statistical metrics, such as the *C* statistic, may offer limited insight but will not portray the full picture.

For example, a now-famous paper published in *Science*[9] documented racial bias in an algorithm used by many health systems. The model evaluated the percentage of patients who needed additional preventive healthcare. But because the model used the proxy, which was healthcare costs (related to the amount of interaction with the health system), rather than the actual health status of individuals, it led to a massive racial bias. The algorithm reported that about 18% of Black patients needed the additional measures; however, in reality, for the healthcare to be equitable the true proportion would be over 46%. This approach of using a proxy for an outcome that does not match what's clinically relevant is a failure of design. This example prompted us to propose eight criteria for developing and evaluating clinical decision support tools in the context of heath systems, but also for clinical guidelines and beyond.[10]

Principles for Evaluation

- **Population at Risk:** The population in which the model is developed and the population(s) for which the model is deployed should match closely if the model is to perform adequately.
- **Outcome of Interest:** The outcome used for modeling must closely match the outcome of interest in the clinical setting. Poor proxies will create more harm and less value.
- **Time Horizon:** The timeframes used in the model must be relevant and applicable to the populations for which the model is being applied.
- **Predictors:** Are clinical predictors measurable within the model's clinical use context, and can they be measured without bias? If you are using multiple predictors, do they add value, or just complexity?
- **Mathematical Model:** Despite the current vogue for complex (and sometimes inscrutable) machine learning models, do the job with the simplest model that will achieve your goal.
- **Model Evaluation:** What can you truly do?
- **Translation to Clinical Decision Support:** How is the model going to be used in the clinical setting? What is the value of the tool, beyond the assessment of the model?
- **Clinical Implementation:** Monitoring and maintenance bring the model back full circle to the beginning.

The figure below shows the algorithmic governance process that we employ at Duke Health,[11] which views a new algorithm in the context of its lifecycle. Because a single, one-time assessment is not enough, we have multiple checkpoints for different stages of evaluation.



After testing the model on retrospective data from our health system, we run it in the context of the health system application in silent mode (results are not given to practicing clinicians at this stage) and prospectively evaluate its performance along with any issues that arise. The third stage of evaluation is evaluation on pre-specified outcomes. Ideally, this would be a randomized experiment comparing the model against current practice. This is followed by multiple evaluations after implementation at regular prespecified intervals.

Performance Metrics

Algorithm

- **Discrimination:** The ability to separate those who have events and those who don't
    - Area under the curve (AUC), C-index, Brier score, etc.
- **Calibration**: How close the predicted and observed risks are
    - Graphical displays
    - Calibration in key subgroups
    - Useful for detection of bias

CDS Tool

- Sensitivity, specificity, PPV, NPV, net benefit, relative utility
- Net benefit curves
- Weighted metrics[12]

In a paper with colleagues in Europe,[13] we pointed out that perfect calibration across all possible subgroups is not achievable: the model will always be biased in some way.

We therefore need to decide what types of miscalibration are not acceptable. Ensuring that the metric is meaningful and aligned with the clinical use case is critically important. Equally important is the ability to show that a strategy based on the new algorithm improves current practice. Rigorous study designs are needed. Randomized experiments are ideal, and we don't do enough of them in the context of health systems research. However, they can be done, and they are not as difficult or expensive as randomized clinical trials in pharmaceutical research.

Key Discussion Points

To demonstrate the value of a strategy based on a new algorithm, we should identify outcomes that matter (clinical, operational, financial), appropriate study designs and relevant comparators. We often forget that new algorithms and related strategies are rarely introduced in a vacuum. When we build systems, we need to understand how these likelihoods are linked to diagnoses and predictions, and how they affect larger decision chains about costs, benefits, thresholds, etc. as we design and implement evaluations. Unfortunately, we currently know very little about the status quo for any given care scenario for which we're trying to develop a model. Do we understand how introducing any models change the status quo?

In fact, the status quo is rarely explicitly defined and quantified. Having a new model forces the conversation and standardizes the process. Having a model forces the conversation and standardizes the process. Ideally, you should have a randomized experiment to run against the status quo.

LIGHTNING TALK 3: TESTABILITY, USABILITY & SAFETY: LANDSCAPE ANALYSIS AND EVALUATION CRITERIA

Presented by Shauna M. Overgaard, PhD (Mayo Clinic AI Translation Assessment Group)

If approached responsibly and meticulously, the application of ML to available medical data could revolutionize healthcare. However, prioritizing testability, usability, and safety of solutions based on ML is critical to the successful translation and evolution of healthcare AI.

Mayo Clinic has formed an enterprise-wide, cross-functional team of experts to streamline organizational processes and assess the implementation and integration of AI models into the clinical workflow. This multidisciplinary advisory group:

- Assesses the feasibility of AI to solve a given problem (i.e., right fit);
- Assesses the risk of patient harm, clinical benefit, and cost/return on investment;
- Assesses technical and operational feasibility of proposed AI solutions; and
- Guides the testing and implementation of AI within Mayo Clinic.

Concurrently, Mayo is also systematizing scientific rigor in this work through the development of streamlined processes, frameworks, and tools to develop trustworthy explainable and responsible AI and accelerate translation of AI into clinical

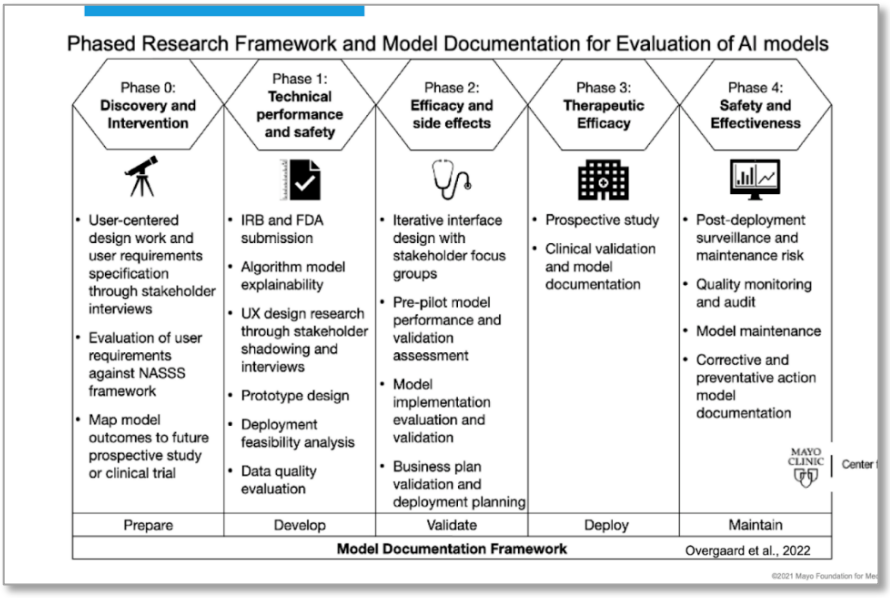
practice. We approach testability, usability, and safety in multiple phases aligned with those of clinical research studies.[14] Work is performed with the ability to reproduce findings by providing replicable methods and algorithm code.

Research Framework and Model Documentation

We know that AI models run the risk of overfitting or working only with the specific data set being tested, so we begin to address this by understanding and testing the methods outside of the original study. For this reason, making models, software, code, and data available for independent validation remains a priority in our data science work products. As the development of a tailored, risk-based framework for testing and safety which is done through strong partnership with our Software as Medical Device (SaMD) governing body.

The extent to which machine learning system can be evaluated for usability may be characterized by the achievement of specified goals with effectiveness, efficiency, and patient satisfaction, sometimes in multiple healthcare environments. These applications often must be scalable across multiple settings and offer an improvement on usual care, while also avoiding additional burden on providers and patients. The current clinical workflow and its constraints must be understood so that interventions guided by machine learning systems can be compared with usual care. Then, a documented evaluation of sustainability and sustained scalability of requirements can generate value, and the tool's adoption can be tracked. Methods and instruments to support these priorities included a tailored risk-based framework for testing and safety. This is approached carefully through quantification of risk and by defining control mitigation strategies.

The widening of responsibility gaps and the additional risk of negative side effects are inherent in healthcare interventions and increasing the scope and authority of digital health systems is challenging. By prioritizing consistent and sustainable risk reduction, reducing the occurrence of avoidable harm, and making errors less likely, we can systematically protect patient safety. One aspect of this work is understanding how clinicians, patients, and AI systems adapt their behaviors throughout the course of their interactions.



Transparent and Explainable Documentation

Communication among groups, disciplines, and sectors is essential to ensure development of useful and beneficial systems. To address gaps in explainability, transparency, accountability and trustworthiness, Mayo has created scalable documentation that addresses and communicates a solution's purpose, development, implementation strategy, and limitations.[15,16]

Mayo is currently developing another checklist for developing, validating, deploying, and maintaining solutions, one that aligns with clinical research phases and associated subphases complementing production lifecycles of documentation. Together, it comprises a framework for AI translation.

Specifically, it encourages knowledge continuity and ensures that relevant components are considered, reported, maintained, and communicated across stakeholder groups, product teams, and end users through evidence-based reporting driven by subject matter experts (SMEs). The individual phases are listed below:

- **Prepare:** patient impact, purpose and indications, model planning and architecture, and data bias evaluation.
- **Develop:** risk assessment, usability, formative, and model bias evaluation.
- **Validate:** deployment and validation planning. This includes how and when to use a model, engaging appropriate teams for translation to coordinate use, and adoption strategies, detailing pathways for application, and emulating and testing end-to-end workflow.
- **Deploy:** clinical validation, user education and training, monitoring, and reporting, corrective and preventive action. These include including verifying and validating model performance, quality testing the full pipeline, and discussing limitations.
- **Maintain:** post-deployment maintenance risk, quality monitoring and audit, maintenance.

This framework affords an efficient way to engage stakeholders, store and reuse content prepared by SMEs, and effectively facilitate knowledge continuity. This tool constitutes a central documentation platform that serves a model's health and life cycle, like how an EHR serves as a patient record and repository. Leveraging mated metadata will allow studies to be scoped and classified. Downstream impact of decisions can also be identified based on provided information and stakeholder associations, serving to compile and submit relevant extracts developed by SMEs for reuse.

Key Discussion Points

Exploring questions about safety and performance of models evaluated in this framework may present opportunities to drill down on detailed data. In addition, how tools are used in real-world clinical settings,[17] and how data are presented to users to guide decisions, present critically important issues for further discussion.

BREAKOUT SESSIONS

Following the conclusion of the lightning talks, conference attendees were divided into groups to participate in breakout sessions that addresed the topics of testability, usability, and safety in healthcare AI applications. Each breakout session included a series of key topical questions intended to focus the resulting discussions.

Testability

Prompt for Discussion

Let us consider testability to be the extent to which an ML algorithm's performance can be verified as satisfactory. Specifically, we assume the algorithm developers have tested the algorithm's performance, so we are focusing on algorithm testability by groups other than the developers.

Key Questions

- What would you consider a testable health AI tool?
- How is testability measured, and who bears the responsibility to test an algorithmic tool?
- If an algorithm contributes to medical decision making, should the clinician and/or patient be able to request an in-depth summary of the algorithm's performance?

For comparison, consider areas in care delivery that end-user testability is currently standard practice, such as usage metrics for rules based CDS alerts and calibration of laboratory testing equipment. Also, consider areas where end-user testability is not standard of care, including pharmaceuticals and vaccines.

Discussion Points

- Testability may be a sliding scale from one device to another, depending on who (Clinician? Developer? End user?) is assessing it. Ensuring that the application behaves ethically throughout its lifecycle will require a shared partnership.
- Ensuring strong sensitivity and specificity of models is important. Many evidence-based models do not generalize well, because the environment in which the model is trained is different from the one in which it is deployed. Overpopulating the training data may contribute to this issue.
- Most AI applications are optimizing tools - they are constantly evolving and changing. Testability is not a single evet, but a continuum, and as such requires continuous feedback loops, especially in healthcare.
- Developers have the responsibility for testing systems on the conditions they will encounter when the tools are deployed.
- Whenever any new technology is introduced into something as complicated as a hospital system, it raises issues about how to go about

evaluating that technology in its new environment. What standards of practice and policy are applied? Tracking and rigorous evaluation on dynamics of care both before and after the application of the new system is needed; otherwise, ripple effects (both good and bad) will be hidden. Meta-models can permit holistic evaluation involving multiple scenarios.

- Personal experiences shape our professional lenses. Patients feel strongly about 1) co-designing technology (you can't talk about patient experience without patients); 2) transparency of the data (how are data used and stored; how are algorithms applied to it); and 3) education (the latest and most sophisticated tools won't help if patients can't use them).
- Patients should own their data. Incentives/compensation could be offered in return for allowing use of that data; this could facilitate a more global database that can be trained (e.g., genomics for risk stratification).

Usability

Prompt for Discussion

Let us consider usability to be the quality of the user's experience, including effectiveness, efficiency, and satisfaction, when using an algorithm's output. Specifically, consider the point at which the output is integrated into a clinical or operational workflow.

Key Questions

- What would you consider a usable or non-usable AI tool?
- Are there components of good usability that are unique to ML applications, compared to traditional software applications?
- How is usability measured, and by whom?
    - Who is responsible for determining what is an acceptable level of usability? Consider algorithm developers, EHR vendors, hospital IT departments, clinicians or patients.

Discussion Points

- AI tools that create alerts or reminders every time they are used are not optimally useful. AI tools that will create a closed-loop intervention with patients that is not intrusive are needed.
- Patients' perspectives must be incorporated; otherwise, we don't know how it affects their lives.
- Tools that are too complex or too difficult to explain impose a significant cognitive workload on patients, especially when not even physicians understand the inputs/accuracy of predictive algorithms. We need to listen to patient communities to better understand how we can help and formalize that role.
- A good first step is to let people know decisions are made from an algorithm. Do the patients have the educational and literacy capacity to understand?[18] Does

being open and authentic with patients, and being transparent about the fact that the models are not perfect, erode patient confidence in the assessment?

- Explainability makes models more biased and worse in terms of performance and adds disparities for marginalized groups. We should instead move toward transparency and maintaining rigor in the data we select and the outcomes we define.
- As models become increasingly complex, the language and concepts we use can contribute to disparities. We need to be sensitive the level of healthcare knowledge a given person has. How to customize information for patients, including those with disabilities, needs to be incorporated into the design of these tools.
- The time horizon over which you generate predictions has a large impact on usability. We must be careful about the point at which we
- expect the human to affirm or dismiss the notification and in the interest of fairness ensure that the person is capable doing what they are being asked to do.
- A missing piece of the healthcare continuum is patient communities with the capacity to teach each other among peers, guided by experts. We should consider learning health network models and communities of peers that can meet people where they are and help guide technical discussions.
- It could be argued many Americans have little insight into the analytics that shape a significant portion of their financial lives. We have yet to really address financial literacy with the public today. We appear to be following suit with the powerful, life-changing analytics used in healthcare unless we find ways to address this knowledge gap across the entire patient continuum, from the education system on through to employment and everyday life.
- Biases are not unique to the ML domain – they exist in medicine with or without ML in the picture. The problem ultimately resides with care providers and lack of oversight.

Safety

Prompt for Discussion

Let us consider safety in the context of the potential for worse outcomes for the patient, provider or health system to accrue as a result of use of an ML algorithm.

Consider the Risk Categorization framework offered by the FDA (see table), which combines both characteristics of the patient's health (critical, serious, non-serious) and the significance of the information provided by the ML algorithm (diagnosis/treat, drive clinical management,

	Significance of information provided by SaMD to healthcare decision		
State of healthcare situation or condition	Treat or diagnose	Drive clinical management	Inform clinical management
Critical	IV	III	II
Serious	III	II	I
Non-serious	II	I	I

Source: https://www.fda.gov/medical-devices/software-medical-device-samd/global-approach-software-medical-device
SaMD, Software as Medical Device

inform clinical management). While this framework focuses on risk to the patient, it shows the spectrum of risk profiles that can be associated with ML algorithms.

Key Questions

- What would you consider a safe AI tool? What characteristics make an ML model 'unsafe'?
- How is safety measured and by whom?
- Who is responsible for the safe use of ML algorithms? Consider federal regulators, ML developers, health systems, clinicians, or patients.

Discussion Points

- Safe ML model is something that does not create an outcome worse than the status quo. A safe AI tool will not inflict any harm and is transparent if it is not sure (in other words, the level of confidence of the prediction will be clear).
- AI can be considered safe when it enhances the ability of the doctor to make better decisions or enhances the patient's understanding. Unsafe AIs are models that, while accurate at the time of deployment, are set in motion without any observation or oversight in patient-care environments.
- The safety of a system may be numerator-driven, in that one harmful event is one too many, whereas the quality of a system is a numerator/denominator consideration. We need a nuanced understanding about what the standard of care is, what constitutes an improvement in safety or performance regarding the status quo, and how to develop a quantified sense of how things might go wrong (and how to anticipate that).
- All data that is important, but from a patient's point of view, considerations such as where the data is going, how it is collected, who is receiving it, are particularly important. Technologies, including AI applications, can help build relationships between physicians and patients as we move toward decentralized trials. But we must also ensure that patients are better prepared as partners. This requires taking a deeper look at how AI is used and getting inputs from different communities. Safe AI requires involving organizations at the patient level who are conversant with the impact of technology on personal and social dimensions.
- When we think about safety and effectiveness of other types of medical products like drugs and other types of medical devices, we generally think about the intended, targeted response in terms of effectiveness, while safety is thought of in terms of an unintended side effect that the drug may have. AI, especially when it is not part of medical device is different: AI is working outside of your body. Thinking about side effects requires thinking more broadly.
- What is the risk for the patient? For drugs or medical devices, it is serious adverse events vs non-serious adverse events. Why should there be a difference between device and non-device when deploying these tools? When the model is making decisions and we are applying those decisions in care settings, how do

we report serious adverse events vs non-serious adverse events?

- The field of AI lacks a clear understanding of who is responsible for user safety, for defining what is appropriate and what safe use is. We also don't know that implications are for insurers. We have some idea of the status quo performance; what does the AI tool or ML algorithm do? We must think about this in the context of systems, not just assessing the tool by itself.
- As we move toward decentralized clinical trials, we will rely increasingly on patient-generated data as parsed by an AI algorithm. We will need to examine whether people are collecting the right types of data, in the right manner, in the right settings. Safety should be measured by a physician with input from the patient – relying solely on the algorithm may lose the human dimension.
- The framework referenced in the table above is useful but could be improved. In the context of algorithms, we can ask questions such as: Does it improve the C-statistics? Will it rank-order people better? Will it place them over different threshold? But we must also ask what consequences it has. Does it introduce bias? Those responsible for model governance also have the responsibility to demonstrate a given tool's safety regarding pre-specified measures
- Even though they may be the ones blamed if something goes wrong, it's not clear that clinicians – some of whom may lack data science expertise - are in a reasonable position to deconstruct and algorithm and understand all the nuances of its pros and cons. As with drugs and devices, developers of ML tools and the health systems that deploy them should share responsibility for the safety of those tools, and regulatory frameworks will need to adapt accordingly.

REFERENCES

1. Wang HE, Landers M, Adams R, Subbaswamy A, Kharrazi H, Gaskin DJ, Saria S. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. J Am Med Inform Assoc. 2022 Jul 12;29(8):1323-1333. doi: 10.1093/jamia/ocac065. Erratum in: J Am Med Inform Assoc. 2022 Jun 17; PMID: 35579328; PMCID: PMC9277650.
2. Li RC, Smith M, Jonathan Lu, et al. Using AI to empower collaborative team workflows: Two implementations for advance care planning and care escalation. NEJM Catalyst Innovations in Care Delivery. 2022; DOI:https://doi.org/10.1056/CAT.21.0457
3. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. JAMA Dermatol. 2021 Nov 1;157(11):1362-1369. doi: 10.1001/jamadermatol.2021.3129. PMID: 34550305.
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44-56. doi: 10.1038/s41591-018-0300-7. Epub 2019 Jan 7. PMID: 30617339.
5. Dullerud N, Roth K, Hamidieh K, Papernot N, Ghassemi M. Is fairness only metric deep? Evaluating and addressing subgroup gaps in deep metric learning. March 23, 2022. Preprint available from arXiv.org at: https://doi.org/10.48550/arXiv.2203.12748
6. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - Reconsidering the use of race correction in clinical algorithms. N Engl J Med. 2020 Aug 27;383(9):874-882. doi: 10.1056/NEJMms2004740. Epub 2020 Jun 17. PMID: 32853499.
7. Suriyakumar VM, Papernot N, Goldenberg A, Ghassemi M. Chasing your long tails: Differentially private prediction in health care settings. October 13, 2020. Preprint available from arXiv.org at: https://arxiv.org/abs/2010.06667
8. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021 Jul;4:123-144. doi: 10.1146/annurev-biodatasci-092820-114757. Epub 2021 May 6. PMID: 34396058; PMCID: PMC8362902.
9. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342. PMID: 31649194.
10. Pencina MJ, Goldstein BA, D'Agostino RB. Prediction models - development, evaluation, and clinical application. N Engl J Med. 2020 Apr 23;382(17):1583-1586. doi: 10.1056/NEJMp2000589. PMID: 32320568.
11. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc. 2022 May 31:ocac078. doi: 10.1093/jamia/ocac078. Epub ahead of print. PMID: 35641123.
12. Reyna MA, Nsoesie EO, Clifford GD. Rethinking algorithm performance metrics for artificial intelligence in diagnostic medicine. JAMA. 2022 Jul 8. doi: 10.1001/jama.2022.10561. Epub ahead of print. PMID: 35802382.

REFERENCES

13. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016 Jun;74:167-76. doi: 10.1016/j.jclinepi.2015.12.005. Epub 2016 Jan 6. PMID: 26772608.
14. Overgaard SM, Peterson KJ, Wi CI, et al. A technical performance study and proposed systematic and comprehensive evaluation of an ML-based CDS solution for pediatric asthma. AMIA Annu Symp Proc. 2022 May 23;2022:25-35. PMID: 35854754; PMCID: PMC9285150.
15. IEEE Ethics in Action. Addressing ethical dilemmas in AI: Listening to engineers report. 2021. https://standards.ieee.org/initiatives/artificial-intelligence-systems/ethical-dilemmas-ai-report/
16. Consumer Technology Association. The use of artificial intelligence in healthcare: Trustworthiness. ANSI/CTA 2090. February 2021. https://shop.cta.tech/products/the-use-of-artificial-intelligence-in-healthcare-trustworthiness-cta-2090
17. Fogliato R, Chappidi S, Lungren M, et al. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. Accepted at ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2022. Preprint available from arXiv at: https://arxiv.org/abs/2205.09696
18. Light Collective. No aggregation without representation. Available at: https://vimeo.com/564446961.