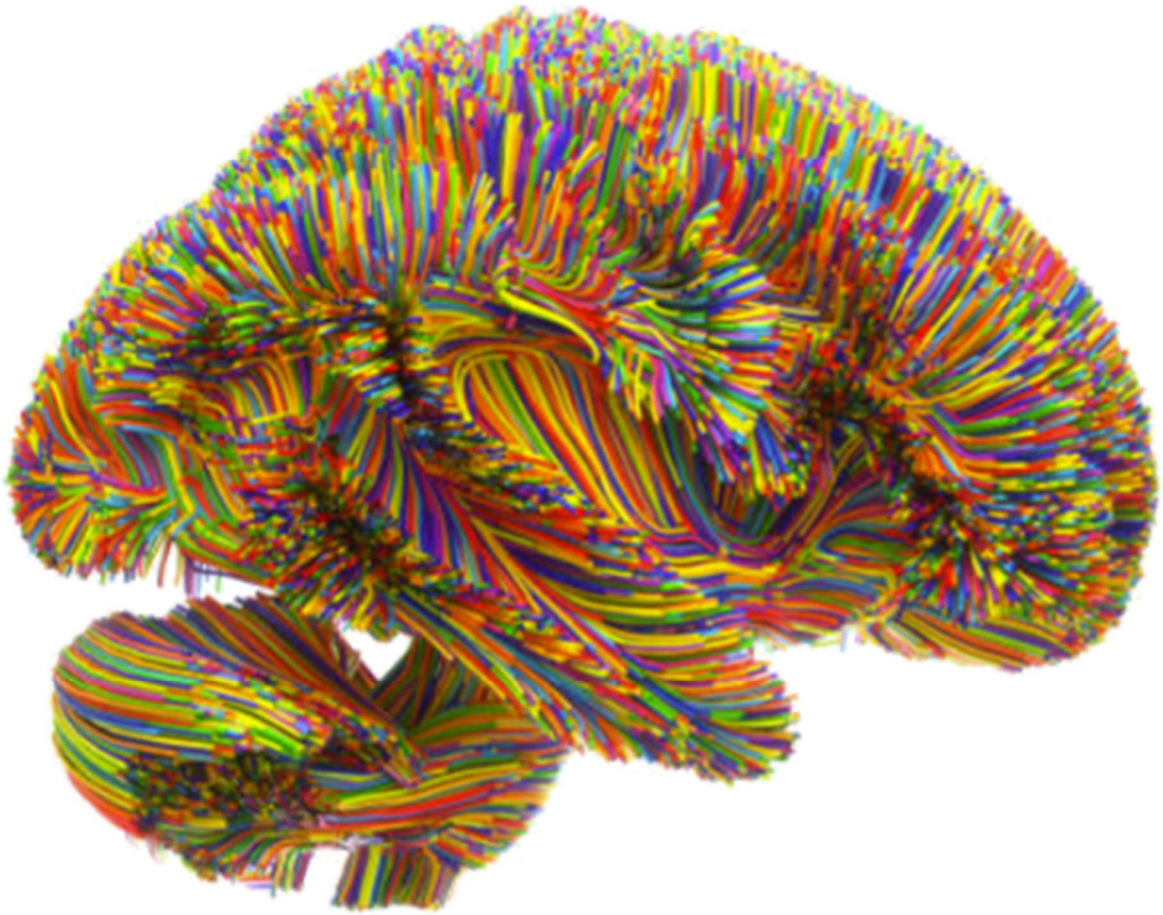

PROVIDING GUIDELINES FOR THE RESPONSIBLE USE OF AI IN HEALTHCARE

COALITION FOR HEALTH AI VIRTUAL WORKGROUP SESSION: RELIABILITY AND MONITORING

SEPTEMBER 22, 2022, 2:30-4 PM 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 reliability and monitoring in the development and use of artificial intelligence and machine learning applications for healthcare. 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 that explored issues related to reproducibility, reliability, and monitoring in health AI, followed by brief group discussions. It also featured a set of focused breakout sessions that addressed these themes in the context of the selected use cases. 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.

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 reproducibility, reliability, and monitoring 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: AI REPRODUCIBILITY

Presented by Christine Kirkpatrick (San Diego Supercomputer Center) and Kevin Coakly (San Diego Supercomputer Center/Norwegian University of Science and Technology)

Overview

Although a "reproducibility crisis" in science at large has received much attention, reproducibility specifically for AI is an exponentially more difficult problem, because so many AI algorithms rely on "black-box" processes that are not transparent to users. There are also questions about how data are collected and prepared for use in training datasets. Adding to these challenges are issues such as the significant computational resources and large amounts of data that would be needed to reproduce someone else's work.

The term "reproducibility" is sometimes used interchangably with terms such as repeatability and replicability; however, there are some important differences. We can think of repeatability – of a researcher trying to repeat their own experiments – as one end of a continuum - being able to document a process and provide resources such that someone else could independently reproduce your results. If we think about, for example, a cancer drug in development, clearly we want to know that the results seen by one person in the lab can be reproduced by others. However, standards for reproducibility will vary across different disciplines.

If we think about reproducibility in context of the scientific method, then we can think about reproducibility at different levels. For example, you may be able to repeat an experiment's results exactly, or you might also be able to apply a different analysis but still come to the same conclusion. Implementation is interwined with reproducibility, which in turn is sensitive to the environment in which the research is done. In the context of machine learning and deep learning, some of these factors would include aspects such as initialization software, parallel execution, compiler settings, auto-selection of primitive operations, processing units, and rounding errors, among others.

Case Study: Reproducibility Exploration Using Open Science Grid

Not all of these reproducibility factors can be controlled for. If you only have one machine or rely on cloud computing, you may not be able to be able attempt to reproduce findings on different machines. However, it is important to understand that factors affecting reproducibility can be examined by running examples multiple times in heterogeneous computing environments with different hardware and software.

One of the biggest challenges for researchers exploring reproducibility is getting access to multiple computing environments. For this reason, we used the Open Science Grid to access 10 different central processing units (CPUs) of four different graphical processing units (GPUs), running different versions of the machine learning platform TensorFlow. Three simple use cases were chosen from the Keras website, with the examples changed to run deterministically.

One of the examples was in computer vision; another used natural language processing (NLP); and the third involved structured data. Each example was run multiple times on different hardware.

A heat map provides an overview of the results. These results were achieved using the exact same code and computing evironments – the only thing that changed was the version of TensorFlow and the CPU. Even with these simple cases, a 6% difference emerged with the NLP example. There was less variance with the other examples: 3 percent for the structured data example, and almost none for the computer vision example. However, when these examples were run on GPUs, it became evident that TensorFlow's Docker containers included a software bug that caused the NLP example to produce non-deterministic results and a variance of over 8%. However, after changing to an updated version of TensorFlow that was not affected by that bug, deterministic results were again seen.

One salient point here is that due to limitations on time and resources, researchers may be unaware of issues such as software bugs, which are only revealed by tests like these. A variance of 3%-5% may be acceptable for a given task. But if the application is something like medical imaging, it may be more important to have accurate results.

Key Takeaways

- AI and ML are under constant development, with new implementations of algorithms and new versions (e.g., operating systems, frameworks, processors).
- Commitment to repeatability is needed when working with AI.
- Reproducibility requires documenting all applicable implementation factors:
    - Publishers should consider guidelines for AI/ML-driven work
    - Nanopublications are needed for replicable documentation
- Awareness building is needed, especially when data with embedded bias is used in ML.
- Research and tools are needed to assist researchers in prioritizing reproducibility and documentation.

Performing experiments in multiple environments can expose how sensitive your experiments are to these factors, and whether and how this variation could affect your analysis or conclusions.

Machine learning tools are constantly changing. For this reason, it's important to partner with technologists who can track these changes and help you to factor this into your replicability and reproducibility documentation. In applications of life-or-death importance, it's imperative to put in the work to understand these variations. AI is computationally intensive and expensive. There is only so much money to spend on the cloud, or on so many computing cycles, and so people might just be glad to run something, get an answer, and be finished. But it is essential to run experiments multiple times and examine the variation across experiments.

We focused here on implementation factors, but there are other categories of reproducibility, including design and evaluation factors.[4] Some computer science conferences, for example, require a supplement addressing reproducibility when submitting a paper. However, there is room for publishers and researchers to do more, particularly with regard to "nanopublications" (brief reports of 1,000 words or fewer that could be used to document work and be reused by others. There are also efforts underway to build communities aound promoting of better practices for AI and to improve efficiency and reproducibility, such as the recently launched Fair AI Readiness & Reproducibility (FARR) initiative.

Key Discussion Points

- There are multiple definitions of reproducibility.
- Professional societies that interact with the FDA are trying to provide some additional clarity on concepts we defined reproducibility with respect to different teams, repeating the same experiment and depending on your perspective.
- The same experiment can mean different things in the context of health care. For example, as the perspective is widened, one can find more variability of sources of irreproducibility.[4] It might be necessary to preprocess the data before it's used as an input for machine learning model. Although different methods may aim to accomplish the same pre-processing, there may be differences in their implementation, which could itself be a source of variability.

LIGHTNING TALK 2: RELIABILITY & MONITORING IN AI

Presented by Irene Dankwa-Mullan, MD, MPH (Merative/formerly IBM Watson Health)

Health equity efforts at Merative include programs focused on building inclusive technologies and promoting inclusive language, figuring out how to build on data and diversity, and how to build ethical AI and machine learning that incorporate the central tenets of fairness, trust, and transparency. Merative works with the entire health ecosystem and its stakeholders to optimize solutions for improved health, and a significant part of this involves addressing bias, evaluation, monitoring, and building on scientific evidence and social impact.

In previous meetings we've talked about designing for equity–designing better data to advance our efforts to achieve health equity for all, and sharing expanded or general concepts of health AI bias model assessments, which will lead to important considerations about reliability as well as monitoring.

The Five Es – Broad Aspects of Bias Across the Data Generation and Model Development Continuum

Evidence

In thinking about monitoring and mitigating AI bias, one of the first things we must consider is the fundamental bias that affects clinical research and scientific evidence. This bias stems from how we translate

science and evidence into care for all populations, especially communities that experience barriers to optimal health. Clinical decisions are informed by a synthesis of evidence tied to rigorous randomized clinical trials and studies using real-world evidence. It's also tied decision about what gets funded by the National Institutes of Health and other agencies, and what research is published in peer-reviewed journals.

Experience/Expertise

Experience/expertise is an integral part of translating patient data into improved health outcomes. When a patient arrives at a medical facility, the actions of the provider as expressed through examining the patient, listening to their story, taking into acccount their understanding, preferences, culture, beliefs, life experience – all of this is translated into EHR data and into care. We also know that a patient's health plan may or may not pay for certain treatments, or make available the full range of treatment options for the patient. Unconscious bias in the health care system is a part of what we need to think about when we're considering data bias.

Exclusion

The third aspect of bias concerns the exclusion of key information, such as patients' life course, history, and other determinants that shape health and outcomes. Bias can be introduced when there is a lack of standards. For example, for categories such as race, ethnicity, gender, disability, socioeconomic status, or occupation, we know that these are linked to understanding disparities and promoting equity. We need to build on that data information architecture, because it does not fully capture the scientific evidence that has accumulated around health equity. Most of our current research practices around data reinforce norms of homogeneity against Black or Hispanic communities. We often apply similar standards, comparing Black versus white, or white versus Hispanic, despite the fact that there are known within-group differences or different risk attributes.

The Five Es

- **Evidence.** Researcher bias – lack of equitable standards around how science is funded, conducted, reviewed, published and disseminated; lack of inclusion in clinical trials; lack of diversity in researchers, evidence base, and real-world data.
- **Experience/Expertise.** Provider bias: provider expertise and experience; cognitive biases and in-group biases; preexisting stereotypes or discriminatory practices from providers/health professionals.
- **Exclusion.** Embedded data bias: incomplete/missing health data; data bias in sample selection, modeling structure and selection of metrics for prediction; lack of cohort diversity; unrepresentative training data.
- **Environment.** Data invisibility: lack of data on important factors (social determinants of health and environmental triggers) that can trigger discriminatory outcomes.
- **Empathy.** Data empathy: lack of knowledge, understanding, and/or experience about the people, places, and factors that make up the data; inability to recognize bias and optimze analysis; lack of knowledge about data sources and real-world evidence or social implications.

Environment

Designing robust data representations that capture life experience, exposures, health determinants, and other relevant data points would allow us to integrate these data into our health AI. This is not as simple as recording ethnicity or socioeconomic status, but requries thinking carefully about all of these elements and how they relate to different aspects of environmental and life-course exposure.

Empathy

The fifth "E" is data empathy, which refers to the degree that empathy, patient values, preferences, or reported experiences or outcomes, are integrated into care and decision-making, or into our own reporting measures and benchmarks.

Monitoring Our Efforts

We know that our goal is to understand or optimize AI based on machine learning tools, increasing beneficial impact and reducing risk and adverse outcomes for all populations, while prioritizing human agency and well-being. We developed a framework of rubrics that provide a high-level structure for thinking about accountability.[5]

This framework can guide us in ensuring equity and acknowledged values for those groups to which the AI Tool will be responsible. It can inform how we promote algorithmic impact assessment to understand the health and social implications of AI, particularly as they are being integrated into our health systems and hospitals. We are all thinking about how we can use AI for the right purposes to bring out the best in humanity. A key part of this is incorporating the entire continuum in reliability and monitoring efforts.

Integrating Equity and Design Justice Across AI and Model Development

- Accountability
- Impact of algorithms
- Data responsibility
- Design equity
- Discrimination and bias
- Empathy
- Explainability
- Fairness
- Human oversight
- Human autonomy
- Inclusion
- Social cohesion
- Inclusive technology
- Moral agency
- Privacy protection
- Robustness and safety
- Transparency and trust
- Value alignment

Key Discussion Points

- You can only get out what you put in, and if you're not giving all the proper context, you're going to get a very uninformed response from your AI process, and thinking about sources of bias through omission is really important.
- "Accountability" is a word that's not used enough in the context of AI. We need to remember that AI is a tool to serve humanity.

LIGHTNING TALK 3: RELIABILITY IN MACHINE LEARNING: DEFINITIONS AND CONSIDERATIONS

Presented by Adarsh Subbaswamy, PhD Candidate (Department of Computer Science, Johns Hopkins University) and ORISE Fellow (US Food and Drug Administration)

As we examine the topic of reliability in depth, we encounter the subdiscipline of systems engineering called Reliability Engineering, which has a long history in safety-critical systems in more traditional engineering fields outside of healthcare (such as aviation, civil engineering, or nuclear power). This concept of reliability is probably familiar to many. One definition offered by the Institute of Electrical and Electronics Engineers (IEEE) is:

> *The ability of an item to perform a required function under stated conditions, quantified in terms of the probability of success (i.e., avoiding failure).*

The two key pieces here are being able to define the required function and the intended use of the tool, as well as the stated conditions (what is the working environment? What is its integration?). There are aspects of machine learning and modeling in healthcare that make defining the required function and stating these conditions a challenge.

Principles of Reliability

In the sphere of reliability engineering, there are three core principles[6] of reliability for ensuring the reliability of a tool:

1. **Failure prevention:** prevent or reduce the likelihood of failures. In other words, taking a proactive approach to prevent and reduce the likelihood of failures.
2. **Failure identification and reliability monitoring:** identify failures and their causes when they occur (in other words, regular stress testing after a tool is deployed).
3. **Maintenance:** fix or address failures when they occur. In other words, as risks are identified, have protocols in place to fix and address them.

One key challenge related to reliability when talking about ML or AI tools is dataset shift, in which the environment in which an AI model or tool operates differs from the one in which it was originally developed. These differences can arise across sites and over time. There can be changes in technology and equipment used to acquire data (for example, different types of X-ray machines). There can be differences in the population, in demographics, in disease severity. There can be changes in the prevalence of a disease over time (including seasonality) and, interestingly, changes in the behavior of agents who interact with the model or tools (i.e., patients and clinicians). People who have worked within the healthcare setting are familiar with how clinical practice patterns can vary widely across sites and over time. This variation influences the data going into the model as well as how the model is used.[7]

Dataset shifts can affect the working of the different reliability principles. In the case of failure prevention, that may include thinking about how models are being trained and the need for robust training that allows for specific types of dataset shift. In fact, this consideration should start before we start training the models in the design specifications.[8]

Regarding monitoring and governance: a quality improvement/quality assurance approach to these tasks would consist of "improv(ing) care through the use of standardized processes and structures to reduce variation, achieve predictable results, and improve outcomes."[9] An important aspect of a quality approach is the need to understand the root causes of changes. As data changes, the performance of the models will begin to decay. Understanding the role of dataset shifts is important to understanding the root causes for those changes in performance.

A key question related to maintenance concerns how to update models as more data become available – as models are deployed at a larger number of sites and we observe the feedback. how do we update model as

more data becomes available, as the models are as models deployed at a larger number of sites as we observe feedback. What are the protocols for when to perform maintenance? How do we perform these updates? A key consideration here involves the idea of backwards compatibility—ensuring that updates are not disruptive to the way the human-AI team performs.[10]

Key Takeaways

- Reliability is the ability of an item to perform a required function under stated conditions.
- Specification of an AI tool's intended use or required function is heavily affected by dataset shift.
- Principles of reliability include failure prevention, failure identification, reliability monitoring, and maintenance.

Discussion Points

- Prevention, identification, monitoring, maintenance would constitute a potentially useful ontology, possibly adding notions from engineering practices about failsafe technologies and thinking about what that would mean for a medical system: for example, taking it offline and transferring control over to humans if errors are identified.
- There is a question of what to look for when detecting failures, whether it be monitoring or being proactive. Average statistics will not suffice, because a model can perform within an acceptable range on average, but with substantial variation across different demographic cohorts. Monitoring needs to include subsets and cohorts and to address shifts in quality of response, especially along different demographic lines.
- There is a family of algorithms being developed now under the heading of robust machine learning that has mostly been applied on the deep learning and computer vision front, but which will likely have applications in healthcare and lead to fundamentally more robust procedures.

BREAKOUT SESSIONS

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

Session 1: Definitions

Key Questions

- What is the definition of reliability?
- How is reliability assessed?
- What is the definition of monitoring in the context of health AI tools?
- What measures should be monitored?

Discussion Points

- Definitions of the same features will change depending on the hospital without the EHR having a different identifier for those features.
- When dealing with training data, engineering and data science teams don't always know the embedded healthcare context but may not be able to tell if changes in the model output are due to

the model's inner workings or because the data being used was missing something salient.

- "Safe" and "effective" are qualities that are linked to reliability; however, you could also have a reliable algorithm that isn't safe or effective.
- What is the scope of reliability? Is the definition embedded in the reliability of the care being delivered, with all the constituent elements that go into that? Where do we draw the line? How narrow or broad is that definition? Is it everyone who is involved and impacted? The patients, the doctors, the healthcare system?
- What features should we standardize to deploy AI in that way that gets us to the long-term goals of democratization and AI benefiting healthcare? How do we monitor AI safety once these tools are deployed in the real world?
- In the case of larger institutions and corporations, central monitoring may make more sense because of the need for some degree of triaging for safety reporting (and for doing it within a reasonable timeframe).
- A cockpit is a relatively calm and standard environment, but if you take that to an emergency department setting, there is no expectation on how you establish that type of environment. Is this an expectation for machine learning in healthcare?
- There is a need to be explicit about the scope of the AI application in question. What is the role of a given tool within a potentially chaotic setting (such as the emergency department or the intensive care unit), and how is safety understood in that context. It is not possible to separate the reliability of an application from a clear specification of its intended use, or from how it is integrated into the clinical workflow. Deviations of use and the degree of adherence to guidelines about use are separate considerations, but this might serve as a key starting point.
- Military systems are often used in chaotic settings, and reliability for such systems is weighed according to achieving a defined success and whether the cost to fix failures are appropriate for that setting. The language used in specifications for military applications might be helpful.
- One current approach to monitoring is to perform small-scale testing and ask questions about what outcomes and measures will be monitored moving forward. The next step is to perform small-scale testing in real-world situations, and then move into the real monitoring phase following review.
- Building bespoke reliability measures for each algorithm deployed would create a long tail of active monitoring and require substantial resources to efficiently monitor. Can such monitoring be scaled effectively?
- Much of the benefit and risk we have been discussing falls on the project teams.
- There are two basic types of monitoring: 1) the technical/algorithmic and 2) the socio-technical type concerned with how

the medical professional or patient uses a particular technology. (Jennifer)

- Checklists may help organization be more proactive about monitoring for reliability.

Session 2: Reliability

Prompt for Discussion

Reliability is defined as the degree to which a machine learning model and/or health AI tool performs according to its specification. We will assume that the specification of intended use is designated by the model developer and approved by regulators as necessary.

Key Questions

- What information is relevant to include in the specification for a machine learning model and encompassing tool?
- What types of protocols should be in place for when unintended model behaviors arise? For example, should service be discontinued or stepped back?
- How can reliability be ensured as new versions of a model become available? Can rollout be fast-tracked if the model meets performance & usability standards of previous versions?

Discussion Points

- We think in terms of efficacy (how a technology performs in the ideal conditions of a laboratory setting) and effectiveness (how it performs in the real world). These are not specific measurements for reliability, but rather an aspirational view of what we would like to see in the real world. But "reliability" is not currently a consistent measurement across healthcare.
- When we look at reliability and the information that we include in evaluating it, we also need to include operational components - reliability depends on the next action we will take. The urgency of the action also plays a role, creating different levels of reliability and risk.
- The difference in terms of reliability between machine learning models and other clinical devices is what the tool works on (data for AI, human body for most tools) – and data changes much faster.
- Some devices are rules-based (for instance, some of those that provide alerts) and not a true predictive machine learning model. However, risk assessment impact might be the same when looking at a rules-based vs ML model. What is the dependence on the decision-making process on any data-driven output/input? Is it just one of several key metrics or the main one that is being examined? Can we put controls in place when model drift occurs, or the required performance isn't being achieved?
- "Reliable" compared with who or what or over what timeframe are great questions. If the comparison is to a human in a flawed system, versus a perfect ten, other models…that's an important consideration.
- It is critical to know what the expected inputs are and to be able to see how the inputs deviate from the intended use

case. On the other hand, what about the outputs? Currently, we deploy models that don't involve automatic decision-making, but rather assistance with decision-making. How is the end user being trained to interpret the output of the ML model? There should be specific training for interpreting outputs and detecting when the model may not be performing correctly.

- Models need to be seamless as possible, but the need for clinicians to understand the assumptions behind the model might add a layer of complexity. Is there a better way to engineer that behind the scenes? There definitely must be an inventory of the assumptions and caveats that were used to build a given model.
- How can we really understand that an unintended model behavior has occurred? The labeling currently being used might have changed since the original specifications. Further, the outcome itself may change because you intervene after the model triggers.
- Having good data governance is key—making data made available for consumption and being able to see how this data is compared to other data histories. Does the new model have better performance, or is it expanding the limitations of the other models?

Session 3: Monitoring

Prompt for Discussion

Monitoring refers to the ongoing surveillance of a health AI tools with the goal of raising an alarm when shifts in the input data, model predictions, or use are detected.

Key Questions

- What are you currently doing to monitor for shifts in the data, predictions, or way the model/tool is used? What metrics do you consider or should be considered?
- How can we monitor for overreliance on or changes in how a model or tool is used? What metrics and safeguards could be used?
- When does a model or tool need to be updated? What practice(s) are you currently using to determine that an update is required? How do you ensure that updates are backward compatible?

Discussion Points

- One consideration is whether models are being monitored from a numerical standpoint or from the standpoint of clinical relevance. Monitoring performance only numerically may take a shortsighted view of tracking, as opposed to measuring things that clinicians and patients really care about. And there is the further distinction of whether you are measuring the clinical outcome you want to measure, or merely the one you are able (as a proxy) to measure.
- As we examine outcomes or the performance of AI tools, we need to broaden our definition of monitoring, but understanding what we want to measure can be a challenge. We may need to map out our ideas further.
- The use case of algorithms affects the timing of the ground truth. For outcomes

for which ground truth is available relatively quickly and easily (for example, deterioration in an ICU or 3-day readmission), a tightly schedule monitor is feasible. However, for anything in the outpatient realm (where measures may be gathered at intervals of years, such as in the case of the development of chronic kidney disease), ground truths can't be produced any faster.

- For a further example: we added a new model for detecting colon polyps and compared 6 months of using the AI tool with the previous 6 months without using the AI tool and found no added value from using the tool. The lesson is that once you implement an AI tool, you must compare it with something and test it along the way.
- Unit testing offers one approach for monitoring specific algorithms to ascertain whether they are operating as you expect them to. Is the accuracy the same?
- Monitoring needs to remain flexible regarding model performance in a particular setting. If the performance of a model degrades, its use cannot be justified, but if a model's performance degrades but is still superior to not having the model at all, it may still be justifiable.
- However, the threshold for action will depends on the particular project or application – if a false prediction can create substantial risk, then that raises the stakes of the decision. These are nuanced issues, and they need to be discussed before systems are deployed.
- There is a dilemma in terms of who defines how monitoring is done—is it developers? Clinicians? A combination of both? One key aspect of training for a tool should include how to monitor. Someone needs to be identified as responsible and accountable.
- Engineers and practitioners, share a need to be able to monitor tool performance. End users can provide feedback back to engineers as needed.

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, Lu J, Avati A, Wang S, Teuteberg WG, Kenny Shum, Hong G, Seevaratnam B, Westphal J, 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. Gundersen OE, Coakley K, Kirkpatrick C. Sources of irreproducibility in Machine Learning: A review. Preprint available from arXiv at https://arxiv.org/abs/2204.07610. Accessed September 29, 2022.
5. Dankwa-Mullan I, Scheufele EL, Matheny ME, Quintana Y, Chapman WW, Jackson G, South BR. Journal of Health Care for the Poor and Underserved. Johns Hopkins University Press; 32(2) Supplement: 300-317. https://muse.jhu.edu/article/789672
6. Subbaswamy A, Saria S. Tutorial: Safe and reliable machine learning. ACM Conference on Fairness, Accountability, and Transparency. 2019: Atlanta, GA. Available at: https://arxiv.org/pdf/1904.07204.pdf. Accessed October 3, 2020.
7. 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. PMID: 34260843; PMCID: PMC8665481.
8. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020 Apr 1;21(2):345-352. doi: 10.1093/biostatistics/kxz041. PMID: 31742354.
9. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, Pirracchio R. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022 May 31;5(1):66. doi: 10.1038/s41746-022-00611-y. PMID: 35641814; PMCID: PMC9156743.
10. Srivastava M, Nushi B, Kamar E, Shah S, Horvitz E. An empirical analysis of backward compatibility in machine learning systems. KDD '20: Proceedings of the 26th ACM

SIGKDD International Conference on Knowledge Discovery & Data Mining. August 2020. 3272–3280. DOI: https://doi.org/10.1145/3394486.3403379.