Duke | MARGOLIS CENTER for Health Policy

Understanding Bias and Fairness in AI-enabled Healthcare Software

Virtual Public Meeting
December 17, 2021

Background and Glossary

This document provides a brief overview of the topic of bias in AI in health care in order to provide a short introduction to key themes that will be highlighted during the virtual public meeting: *Understanding Bias and Fairness in AI-enabled Healthcare Software*, convened by Duke-Margolis Center for Health Policy.

Artificial intelligence (AI) refers to the ability of a machine to perform a task that is normally done by humans, including problem-solving and learning. As with other domains, there has been a lot of excitement about AI's potential to improve health and medical treatments. Applications include clinical decision support, wearable remote monitoring and analyses, digital therapeutics, administrative software (such as scheduling software that can predict "no-shows", software used to determine home nursing aid hours, and supply chain management), and population health management. It also is being used in the development of other medical products, for example, identifying drug targets and within clinical trial data collection and management. Some of these tools and products are medical devices, under FDA authority, while others are not. These tools have the potential to significantly increase access and efficiency in health care but may also worsen health care disparities if careful attention isn't given to potential biases that may be programmed into the software or inherent in the framing of the problem. This results in AI perpetuating or even exacerbating current bias, inequity, and injustice in our society. And, because of the black box nature of many of these algorithms, identifying potential bias can be challenging. Ideally, there would be processes and tests throughout the development cycle that would allow products to "fail fast" or identify potential biases for mitigation as the software is being built, as well as overall tests for bias after a product is ready for testing and marketing.

Bias can be introduced into AI throughout the software development process. In this meeting, we conceptualize the stages into four overlapping categories: (1) bias introduced based on how the problem the algorithm is addressing is framed by the development team and actions that may be taken as a result of the algorithm that result in inequitable outcomes (2) bias due to unrepresentative or insufficient training data, (3) bias within training data, such as using biased proxies, and finally (4) bias introduced as the training data is prepared and used for learning, including decisions made by the development team when curating data and when deciding on optimization parameters.

Given the importance of addressing bias in AI overall, there are multiple toolkits that have been developed to address it including: IBM's open-source AI Fairness 360 toolkit and the University of Chicago's open-source Aequitas toolset for auditing for machine-learning bias. However, these frameworks represent only the first step of identifying the many ways that AI may lead to disparate and biased outcomes and then developing rigorous tests for these risks. Researchers have also proposed ways in which to label the data used for training the algorithms as well as called for companies to explicitly state the intended purpose of their algorithm so that it can be easily identified if it is used in an inappropriate context.

US government agencies have begun to address this issue as well. The FDA has discussed bias in artificial intelligence within the context of their medical device regulations purview. In June of this year, NIST published a proposal for reducing the risk of bias in AI and the FTC warned companies that biased AI

software could violate consumer protections. In this meeting, panelists will discuss where there are beginning to be consensus solutions for testing for bias and best practices for building unbiased and fair AI health products, and where there are still gaps.

However, there is also reason to hope that carefully built AI could be a tool to help reduce injustices and inequities that already exist in the healthcare system. For example, algorithms trained to expect biased operational data could make recommendations that nudge providers away from decisions partially based on implicit bias (similar to how digital financial services have reduced bias in lending). AI tools can be used by organizations as auditing tools to understand where their care may be having disparate outcomes and why. Another promising use case is for AI to reduce disparities in the quantification of pain. Such AI-based software systems could be critical tools to help health care systems reduce health disparities by better capturing and highlighting biases that currently exist in the healthcare setting.

Glossary Terms

Algorithm

A process or set of instructions, including data driven or human-curated, to be followed in calculation or other problem-solving operations. The technology of Artificial Intelligence uses a variety of algorithms as tools and applications. (ANSI/CES)

Algorithmic bias

When an algorithm demonstrates significantly different performance in a subgroup of the population of interest:

- Demographic (racial, ethnicity, age, sex, gender, etc.)
- Socioeconomic (income, insurance status, etc.)
- Geographic (rural vs urban)
- Health system (community hospital, academic health center)
- Comorbidities

Artificial intelligence

A general term addressing machine behavior and function that exhibits the intelligence and behavior of humans. (Duke-Margolis)

- **Machine Learning**
  algorithms that use data to create relationships without being explicitly programmed.
- **Rules-Based**
  algorithms programed to use (generally clinically accepted) rules to guide decision-making.

Black box

This term refers to users not being privy to how an algorithm works (although the developers may know and are concealing it as a trade secret). With machine learning, a "black box" algorithm may not be able to be explained, and even the developers don't know.

Health disparity

Avoidable differences in disease burden, injury, violence, or overall chances to attain the most desirable level of health. (CDC)

Health equity

In a given population, a lack of unjust and preventable health differences in different communities in a social, economic, demographic, and geographic context (WHO)

Health justice

Giving every community an equitable chance at being healthy by addressing persistent political, economic, and social inequities and injustices that affect all the sectors that serve our communities and have a disproportionate impact on the health of marginalized communities. (adapted from the AAMC Center for Health Justice)

Intersectionality

A framework and term coined by Kimberlé Crenshaw used to understand how social and political classifications such as gender, race, class, and sexuality overlap and intersect in people's identities, lives, in society, and in social systems to create different modes of discrimination and privilege. (Additional Information)

Learning data

Data used by the system to modify or refine treatment to achieve better outcomes with future patients (ANSI)

Operational data

the data used to make a prediction once the algorithm is in use (e.g., a patient's EHR and sensor data used to determine a risk score for a patient) (ANSI)

Structural racism

Racism is a complex system of beliefs and behaviors, grounded in a presumed superiority of whiteness. These beliefs and behaviors are conscious and unconscious; personal and institutional; and result in the oppression of minoritized communities and benefit the dominant group. Racism results in the combination of racial prejudice and institutional power.

Training data

Data used to initially train the algorithm. Learning algorithms use the data to find relationships between the data and the label/annotation (supervised machine learning) in order to make predictions.

References and other suggested reading

U.S. Government efforts

- **U.S. FDA |** Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan
- **NIST |** A Proposal for Identifying and Managing Bias in Artificial Intelligence
- **FTC |** Aiming for truth, fairness, and equity in your company's use of AI
- **OSTP |** Americans Need a Bill of Rights for an AI-Powered World

Proposed strategies for addressing bias

- Algorithmic Bias Playbook (sign-in required)
- **Toolkits:**
  - IBM's AI Fairness 360
  - University of Chicago's Aequitas

Other literature

- To stop algorithmic bias, we first have to define it
- FDA Review Can Limit Bias Risks in Medical Devices Using Artificial Intelligence
- Ethical Machine Learning in Healthcare
- Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril
- An algorithmic approach to reducing unexplained pain disparities in underserved populations
- ANSI/CTA Standard Definitions and Characteristics of Artificial Intelligence

Duke-Margolis

- Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care
- Trust, But Verify: Informational Challenges Surrounding AI-Enabled Clinical Decision Software
- Artificial Intelligence in Health Care Portfolio

Support for this project was provided by The Pew Charitable Trusts.