Stanford MEDICINE

Stanford | Center for Health Education

# EVALUATIONS OF AI APPLICATIONS IN HEALTHCARE STUDY GUIDE

## MODULE 4: DOWNSTREAM EVALUATIONS OF AI IN HEALTHCARE: BIAS AND FAIRNESS

### LEARNING OBJECTIVES

1. Overview of Bias and Fairness in AI solutions in healthcare
2. Understand the different types of bias in AI healthcare solutions
3. Describe algorithmic fairness
4. Identify solutions to address bias and fairness in AI solutions

### BIAS IN AI SOLUTIONS

Recently, reports have questioned whether AI solutions in healthcare might actually perpetuate discrimination if trained on historical data—which are often poorly representative of broader populations. Often, AI models are trained using historical or retrospective data which are often derived from affluent academic medical centers that likely do not contain all populations, particularly diverse populations for which the AI solutions will be applied. AI models exclusively trained on such data may further perpetuate disparities and fail to demonstrate external validity in broader patient communities. This is likely due to a lack of diversity represented in the training data, a lack of understanding how a disease may manifest and progress in different populations, and a lack of human understanding of the potential consequences and biases that may be inherent in AI solutions.

Fairness and bias in AI solutions may be a larger problem in countries where important health disparities exist based on patient demographics, such as the United States. Therefore, as tools proliferate across clinical settings, it is important to think about and understand potential demographic biases underlying model development and deployment.

Heart failure example:

> Not long ago we didn't know that symptoms of heart attack look different in women compared to men, which led to differences in cardiovascular mortality rates between women

> and men. The problem here is that the model was developed based on male symptoms so the model might be very accurate for identifying males with heart attack symptoms, but it might not work well in females, who present with different symptoms. The predictive accuracy, when analyzed against the true clinical outcomes, will decline for women, but not men.

Genomic database example:

> Genomic databases used across the research community where we find major racial bias in the genomic samples collected. These databases are the center of precision medicine, where our ability to identify whether a genetic variant is responsible for a given disease or phenotypic trait depends in part on the confidence in labelling a variant as pathogenic. Research suggests that these databases heavily reflect European ancestry, and in fact are missing major population-specific pathogenic information, particularly African-specific pathogenicity data. Therefore, genomic test results for persons of non-European ancestry could be less accurate, more challenging, or simply unattainable.

Dermatology example:

> In general, patients with darker skin present with more-advanced skin disease and have lower survival rates than fair-skinned patients. It is possible that the only fair-skinned populations may benefit because of the lack of inclusion of darker skinned patients in model training and development. If the algorithm is basing most of its knowledge on how skin lesions appear on fair skin, then theoretically, lesions on patients of color are less likely to be diagnosed and therefore benefit from the AI solution.

These examples provide you with an idea of how wide-spread the challenges are related to fairness and bias in AI solutions for healthcare.

## TYPES OF BIAS

Bias can occur during almost any stage of AI model building and implementation - from data collection to model deployment.

Types of bias:

- Historical bias
- Representation bias
- Measurement bias

- Aggregation bias
- Evaluation bias
- Deployment bias

**Historical bias** occurs if the present or past state of the world influences a model in a way such that its predictions are considered unfair given societal values and norms. Historical bias refers to judgement based on preconceived notions or prejudices. AI algorithms are entirely data dependent and historical bias encoded in real-world data cannot even be overcome by perfect sampling and feature selection. Therefore, it is important to remember that historical healthcare data, in general, is extremely male and extremely white, and this has real-world impacts.

**Representation bias** (also called sampling bias) arises when the sample collected to train an AI solution does not represent the actual distribution of the population it is intended to be applied to. It occurs when certain parts of the final use population are underrepresented in the training data.

**Measurement bias** arises in situations if the noise is not randomly distributed but differs across groups, which leads to differential performance. Often the only available and measurable features as well as labels are only noisy proxies of the actual variable of interest. Usually, one cannot change the data, some historical biases might be indistinguishably linked to the data – but the awareness about the problem is important and mitigation strategies can be identified by taking preventive measures such as pre- and post-processing actions.

**Aggregation bias** occurs while developing the model when we try to combine different populations whose underlying distribution of the outcome under study differs. This problem is known as infra-marginality and requires separate models for the different populations or including the demographic variable into the model to account for the systemic differences. In terms of model development, one size does not fit all. In order to identify aggregation bias, developers need to understand the meaningful distinct groups and reasons why they are different from each other.

**Evaluation bias** occurs during the model validation and tuning. Evaluation bias arises if the testing data, which often includes external benchmark datasets, is not representative of the final population to which the AI solutions will be applied. This is the difference between the data used for model evaluation and the data used for model's real-world predictions. Since developers mostly use a benchmark dataset or a synthetic dataset for training, their evaluation often does not fit the real-world. A solution of this problem is external validation of the AI model on a different unseen data selected from the targeted population. Evaluation bias can arise if inappropriate performance metrics

are used. Evaluation bias also refers to usage of evaluation metrics inefficiently and to avoid it, using granular and comprehensive evaluation metrics is suggested.

**Deployment bias** arises during the implementation of the model. It refers to using the model inappropriately or misinterpreting its results. In other words, if the model's intended use is different from the way it is used, deployment bias occurs. Deployment bias is the interaction of society with the AI solution - how society or the medical community uses the AI solution and its output.

## ALGORITHMIC FAIRNESS

Ethical analysis of AI Solutions in healthcare demand that we take a view of fairness, or more appropriately, justice that centers on the health and lives of people, not the outputs alone. A lot of historical medicine has been influenced by white normativity, which is the basis of many medical facts. This is evident because of a lack of inclusion of diverse patients in clinical research. have knowledge that the insiders don't. Bringing diverse perspectives actually enhances the quality and accuracy of your scientific endeavor.

A key difficulty in developing fair AI algorithms is the fact that no universal notion of fairness exists. Many different definitions have been proposed by researchers over the years and they can be broadly regrouped into three main classes: **anti-classification, classification parity** and **calibration**. These fairness definitions have been shown to all suffer from significant statistical shortcomings. Therefore, special caution and awareness about the notion's limitations and weaknesses is always necessary when applying these concepts in model evaluation settings.

### Anti-classification or "fairness through unawareness"

1. Requires the exclusion of any protected attributes in the outcome modeling
2. Requires the omission of any unprotected characteristics that are proxies of protected attributes

The main shortcoming of this fairness definition is that some clinical risk models need to explicitly include protected attributes for it to be equitable. In particular, this applies to situations where the true underlying risk distribution differs across subpopulations, known as **the problem of infra-marginality**. Therefore, an accurate model must include protected attributes, but must also learn to avoid bias based on these attributes.

AI solutions that use datasets which may be under-representative of certain groups, may need additional training data to improve accuracy in the decision-making and reduce unfair results. Anti-classification requires the exclusion of any protected attributes in the outcome modeling.

**Classification parity of fairness** asks for equal predictive performance across any protected group. When selecting the metrics to examine, it is important to keep in mind the actionable insights resulting from a model output.

Two measures have received particularly high attention by the machine learning community:

1. False positive rate: The probability of predicting a positive outcome when the true outcome is negative should be the same regardless of protected attributes of the patients
2. Proportion of positive decisions: The probability of predicting a positive outcome should not vary across different demographic groups given all else equal. It is also known as demographic parity as it requires the classifier's predictions to be independent of protected traits.

These definitions are problematic when the risk distributions are different for different groups, a problem known as infra-marginality. Classification parity asks for equal predictive performance across groups.

**Calibration** is when a model reaches a good agreement between model predictions and observed outcomes. Calibration means that when conditioning on risk estimates, outcomes should be independent of the protected attributes. Think of calibration as a comparison of the actual output and the expected output given by a system. Calibration is open to manipulation of risk distributions for different groups. Model calibration is an important aspect of development and must be evaluated before model deployment.

Applying fairness measure:

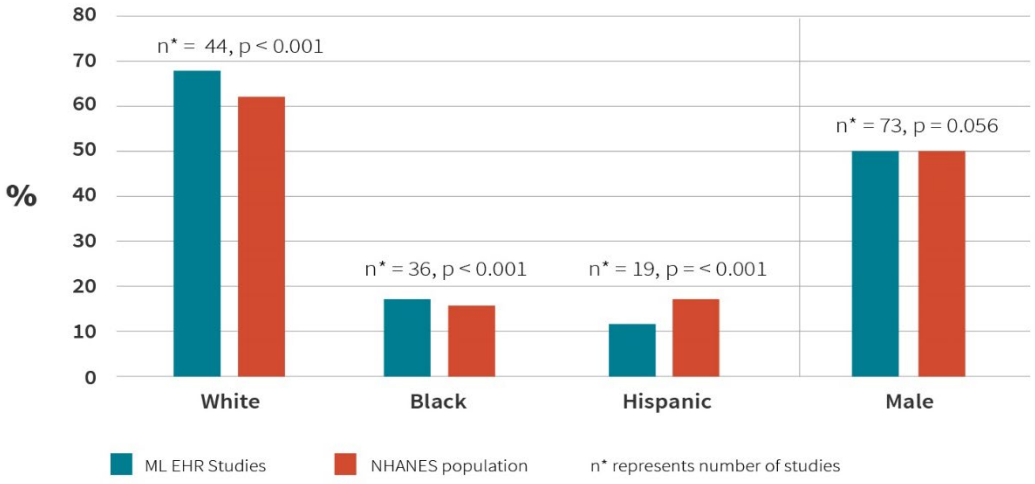
- Anti-classification: Checking the requirement of not using any protected attributes in the decision rule. It gets more complicated when trying to impose the stricter condition to also not use any proxies of protected attributes.
- Classification parity: Once the 2-3 most relevant performance metrics are selected, we evaluate performance differences on the test-set across different demographic groups. It is

always better to calculate empirical confidence intervals of this statistic to decrease the dependence on the test set.

- Calibration: Only look at the overall rate of predicted and observed outcomes across demographic groups.

## TRANSPARENCY

It is important to think about whether the population captured in the EHRs system is representative of the broader community, particularly if these are coming for an academic medical center. If AI solutions are being developed on non-representative populations, the utility and applicability of these advances for the broader patient community falls into question. This relates to an important issue around **transparency and reporting** and many AI studies, particularly machine learning, do not report the demographic breakdown of the training data used to develop and train the models. Lack of reporting makes it very difficult to evaluate the bias and fairness of the AI solution and its applicability across populations.




Studies that mentioned variables often did not report if they were included as model inputs. In the studies that reported demographics, the average populations included higher proportions of whites and Blacks yet fewer Hispanics compared to the general population. While each study might not be applicable to the general population, these findings emphasize the present lack of transparency in reporting details of training data used for development and evaluation by machine learning models in healthcare. To ensure the unbiased deployment and application of any AI model in healthcare, detailed information on the data used to develop and train the model are necessary.

As a solution for transparent reporting and to identify best-practices for designing machine learning models to account for biases and fairness, MINIMAR template is suggested. (MINIMAR = The **MIN**imum **I**nformation for **M**edical **A**I **R**eporting)

MINIMAR Requirements:

1. Include information on the population providing the training data, in terms of data sources and cohort selection
2. Include information on the training data demographics in a way that enables a comparison with the population the model is to be applied to
3. Provide detailed information about the model architecture and development so as to interpret the intent of the model and compare it to similar models
4. Model evaluation, optimization, and validation should be transparently reported to clarify how local model optimization can be achieved and to enable replication and resource sharing

You can understand that by providing these details of the training data and population, model design and intent, an end-user will have a great understanding of how to best deploy the model and in which populations.

## DOWNSTREAM EVALUATIONS

While we have covered many topics in this lecture regarding bias and algorithmic fairness, there are still many more challenges and opportunities for Fair AI research:

- Defining fairness: There are several definitions of AI fairness that have been proposed in the literature. It is nearly impossible to understand how one fairness solution could address all challenges. Identifying the correct or best definition is an ongoing debate.
- From equality to equity: Equity suggests that each group is given the amount of resources needed to have similar outcomes. Understanding how to develop and implement a model that provides both equality and equity presents a paradigm shift in the way to think about healthcare delivery and is an active area of research.
- Identifying biases in models, and particularly in datasets: Many biases are systematic and we are often unaware they exist. We still have a long way to go before we can systematically mitigate these biases and provide our professionals with the appropriate tools they need to address these issues at point of care.

The perspective collection and reporting of AI outputs, clinical recommendations and patients decisions coupled with eventual outcomes is essential in being accountable as healthcare institutions

and as clinicians. This work and transparency in reporting AI solutions is absolutely critical for populations who have difficulty trusting the medical establishment. The key is to use AI in a way that actually does benefit all groups, which requires thoughtful evaluations and human interpretations.

## CITATIONS AND ADDITIONAL READINGS

Adamson, A. S. and A. Smith. 2018. "Machine Learning and Health Care Disparities in Dermatology." *JAMA Dermatol* 154(11): 1247-48.

Beery, T. A. 1995. "Gender bias in the diagnosis and treatment of coronary artery disease." *Heart Lung* 24(6): 427-35.

Bozkurt, S., E. Cahan, M. G. Seneviratne, R. Sun, J. A. Lossio-Ventura, J. P. A. Ioannidis, and T. Hernandez-Boussard. 2020. "Reporting of demographic data, representativeness and transparency in machine learning models using electronic health records."

Corbett-Davies, S. and Goel, S., 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Hernandez-Boussard, T., S. Bozkurt, J. P. A. Ioannidis, and N. H. Shah. 2020. "MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care." *J Am Med Inform Assoc*.

Kessler, M. D., L. Yerges-Armstrong, M. A. Taub, A. C. Shetty, K. Maloney, L. J. B. Jeng, I. Ruczinski, A. M. Levin, L. K. Williams, T. H. Beaty, R. A. Mathias, K. C. Barnes, T. D. O'Connor, and C. o. A. a. A.-a. P. i. t. A. (CAAPA). 2016. "Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry." *Nat Commun* 7: 12521.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366(6464): 447-53.

Parthipan, A., I. Banerjee, K. Humphreys, S. M. Asch, C. Curtin, I. Carroll, and T. Hernandez-Boussard. 2019. "Predicting inadequate postoperative pain management in depressed patients: A machine learning approach." *PLoS One* 14(2): e0210575.

Spanakis, E. K. and S. H. Golden. 2013. "Race/ethnic difference in diabetes and diabetic complications." *Curr Diab Rep* 13(6): 814-23.

Suresh, H. and J. V. Guttag. 2019. "A framework for understanding unintended consequences of machine learning." *arXiv preprint arXiv:1901.10002*.