Bias and Fairness

Michael Yao, Allison Chae

Learning Objectives

1. **Define** algorithmic bias and recognize that bias is often a subjective property of an algorithm.
2. **Reflect** on important case studies demonstrating the real-world impact of bias.
3. **Describe** potential bias mitigation strategies and how we can incorporate them into clinical decision making.

What is Bias?

Bias is a term that is often used broadly but has a very precise definition. Bias is always defined with respect to two related concepts: (1) protected attribute(s), and (2) a definition of harm.

1. A ***protected attribute*** is an attribute about a patient that we want to ensure there is no bias against. Examples of protected attributes might be patient age, gender, or ethnicity.
2. A ***definition of harm*** is how we choose to define when bias is present. Common definitions of harm include an algorithm's overall error rate, or its false positive or false negative rate.

When we choose the protected attribute and a definition of harm, we can then define when an algorithm is biased. Namely, **an algorithm is biased if it causes an increase in harm**

for a subpopulation of patients with respect to the protected attribute(s). For example, if we define the protected attribute as a patient's race and the definition of harm as the algorithm's error rate, then the algorithm is biased if its error rate is higher for Black Americans than White Americans.

Key Terminology: Binary Classification

In this section, we primarily focus on understanding algorithmic bias as they pertain to *binary classifiers*, which are algorithms that seek to classify patients into one of two categories. Binary classification is common in clinical medicine: we often want to classify patients by…

- diagnosing them with a disease or not;
- identifying if they are sick or not; or
- deciding whether a patient can be discharged or not.

In each of these situations, our goal is to label an input patient using a positive label or a negative label.

As with any clinical task, algorithms rarely get 100% accuracy. We may mistakenly diagnose a patient as having a disease when they in fact do not have it, or decide that a patient can be discharged when they need to stay in the hospital for additional treatment. These are situations where an algorithm might cause harm to a patient.

1. An algorithm's false positive rate (FPR) is the rate at which an algorithm falsely assigns the positive label to a patient (i.e., telling a patient AMAB[1] that they are pregnant).
2. An algorithm's false negative rate (FNR) is the rate at which an algorithm falsely assigns the negative label to a patient (i.e., telling a patient that is 8 months pregnant that they are not pregnant).
3. An algorithm's overall error rate is the rate at which an algorithm assigns the wrong label to a patient.

[1] Assigned male at birth, meaning the patient has XY sex chromosomes.

In general, we want to minimize an algorithm's FPR, FNR, and overall error rate. FPR, FNR, and overall error are commonly used definitions of harm in assessing for algorithmic bias.

It is important to recognize that **the consequences of false positives and false negatives can be different** depending on the task. For example, in colon cancer screening, a false negative (e.g., missing a precancerous polyp on a colonoscopy) is much worse than a false positive (e.g., taking out a potential polyp that turns out not to be precancerous). This means that the consequence of a false negative is much more significant than that of a false positive for colon cancer screening.

Sources of Bias

What *causes* an algorithm to be potentially biased? Bias can be due to a wide variety of reasons, including population-dependent discrepancies in...

Other than the sources listed below, what are some other potential causes of bias that algorithms might suffer from?

Availability of Data

Especially for machine learning algorithms, it is important for models to be trained on diverse datasets from many different patient populations. If the dataset used to train a model is composed of 90% White patients and only 10% Black patients, then the resulting algorithm will likely perform inaccurately on Black patients.

This is a common problem not just for machine learning algorithms, but also in insights from randomized control trials! For example, take a look at the 2023 ISCHEMIA Trial[2] from the American Heart Association. According to Supplementary Table 1, approximately 77% of the patients in the study were male. Would you trust the insights from the trial for your female patients?

[2] Hochman JS, Anthopolos R, Reynolds HR, et al. Survival after invasive or conservative management of stable coronary disease. Circulation 147(1): 8-19. (2022). doi: 10.1161/CIRCULATIONA HA.122.062714. PMID: 36335918

Pathophysiology

Different patient populations may have different underlying mechanisms of disease, and so lumping patients together using a single predictive algorithm may limit that algorithm's ability to represent all the different mechanisms of disease.

Quality of Data

Suppose we have two CT scanners in the hospital: Scanner 1 and Scanner 2. Scanner 1 was made in 1970 and Scanner 2 was made in 2020; as a result, Scanner 1 produces very low-quality, low-resolution images compared to Scanner 2. If we learn an algorithm to diagnose a disease from CT scans, then the algorithm will likely perform worse on input scans from Scanner 1. This is because lower quality scans contain less information about the patient, and so patients imaged with Scanner 1 will be inherently less predictable.

How Data is Acquired

The data that we choose to collect to learn an algorithm can also introduce biases. For example, suppose you are conducting a study investigating the relationship between number of leadership positions and success in medical school for medical students. Focusing only on leadership positions might result in algorithms that are biased against students from lower socioeconomic backgrounds who can succeed in medical school but may have to focus on things such as taking care of loved ones or part-time employment that was not factored into the initial algorithm design. In summary, it is important to be thoughtful about not only algorithms, but also datasets as potential sources of bias!

How might collecting too *many* data features bias algorithms?

Case Studies

To better understand the sources of algorithmic bias and why they are important, let's look at some commonly cited case studies (both clinical and non-clinical):

Pulmonary Function Testing (PFT)

PFTs are used in clinical practice to evaluate lung health. The patient's measured lung values are compared with the expected lung values given the patient's age, height, sex assigned at birth, and ethnicity among other factors.

Researchers have found that using a patient's *race* as input into the expected lung value calculation can result in different PFT results, with implications for access to certain disease treatments and disability benefits. However, they also report that using race has also allowed patients to **benefit** from treatment options that they would have otherwise not had access to based on societal guidelines.

The 2019 American Thoracic Society (ATS) guidelines currently offer both race-specific and race-neutral algorithms, leaving it up to the discretion of the provider to determine how PFTs are used in clinical practice. Other historical examples of bias in clinical medicine include eGFR calculations, opioid risk mitigation, and care assessment evaluation as functions of race and other patient demographic information. To learn more, check out the Health Equity article from List et al.[3]

COMPAS/ProPublica

In 1998, a private company built the **COMPAS algorithm**, which is an algorithm that takes in information about arrested/incarcerated individuals and returns a prediction of how likely the individual will commit future crimes. Inputs into the COMPAS algorithm include individual demographics, criminal history, personal and family history, and the nature of the charged crime among others.

[3] List JM, Palevsky P, Tamang S, et al. Eliminating algorithmic racial bias in clinical decision support algorithms: Use cases from the Veterans Health Administration. Health Equity 7(1): 809-16. (2023). doi: 10.1089/heq.2023.0037. PMID: 38076213

The COMPAS algorithm is used in the real world to help the judiciary system set bonds and evaluate arrested individuals. High (low) COMPAS scores mean more (less) likely to commit future crimes.

ProPublica is a nonprofit journalism organization that conducted an independent evaluation of the COMPAS tool in 2016. Their main finding was that COMPAS is biased.

- The distribution of COMPAS risk scores skews low for white individuals but more uniform for black individuals. In other words, white individuals were more often "let off the hook" than black individuals for the same crime.
- Looking at historical data, the false positive rate (FPR) was significantly higher for black individuals than white individuals. In other words, COMPAS was more likely to incorrectly predict that a black individual would commit a future crime.

> Does this mean that the COMPAS algorithm is racially biased (using the risk score as the definition of harm)? What are some potentially other reasons why white scores may be lower?
>
> Other reasons might include...
>
> 1. the nature of the subjective questions asked about the individual's personal and family history;
> 2. other causal variables like upbringing and socioeconomic status that might be racially correlated; and
> 3. the number of black individuals used in the development of the COMPAS algorithm.
>
> What other potential reasons did you think of?

COMPAS Developer Response to the ProPublica report was that COMPAS is not biased.

- Area under the receiver operating curve (AUROC), a metric of classifier "goodness," for black and white subpopulations are equal. Therefore, COMPAS is not biased and we can make sure the FPR of both populations are equal by setting different classifier thresholds for the two subpopulations.

Image Generation Using Google Gemini

In late 2023, Google introduced a new generative AI model called Gemini, which is able to complete a variety of tasks such as generating images from input text descriptions. Gemini was released to the public, and users quickly found that Gemini stood out from prior generative models because it was able to generate more diverse sets of images, such as producing images of people of color when prompted for an "American woman" or producing images images of women when prompted for historically male-dominated roles, such as a lawyer or an engineer. This was seen as a major step forward in tackling the bias associated with other generative models.[4]

However, an unintended side effect was that the model also generated historically inaccurate images when prompted for images of "1943 German soldiers" or "US senators from the 1800s." In these settings, it would be inaccurate to generate images of people of color given these input prompts.

How can we reduce bias?

The most common reason why bias occurs in machine learning is that we train models to be accurate on the "average" patient. In other words, if there are more patients from one subpopulation than another in the dataset used to learn an algorithm, then the algorithm will almost certainly be more accurate on the majority population. Naively learning algorithms in this fashion will result in accurate but biased models.

On the other hand, we could instead implement a *completely random algorithm* - for example, deciding whether a patient should be admitted or not solely based on the flip of a coin. Such an algorithm would be completely unbiased, but not very accurate.

[4]Nicoletti L and Bass D. Humans are biased. Generative AI is even worse. Bloomberg. (2023). Link to article

These two examples demonstrate that in general, there is a tradeoff between **accuracy and bias**. As we train models to be more accurate, they often become more biased at the same time. Researchers are currently working on ways to overcome these limitations[5], but this is an incredibly common empirical finding that we see in practice.

Fairness is not Compositional

Why is training both fair and accurate models hard? There are a lot of complex answers to this question, but one important reason is that **fairness is not compositional**. In simpler terms, imagine that we have a screening tool that seeks to predict whether a patient has a disease with 50% prevalence in the population. The algorithm has been shown to be unbiased against patient gender (male versus female) with regards to sensitivity, and also unbiased against patient color ("blue" versus "green" people) using the same metric. Does this mean that the algorithm is also unbiased against blue female people?

No! The screening tool can still be biased against blue female people in this toy case. Here's an illustrative diagram of one possibility:

In Figure 1, we can see that our screening tool is "fair" with respect to gender and color by only accepting individuals that are either both blue and male or both green and female. Ensuring that models are fair for certain subgroups doesn't mean that those models are also fair for members of the *intersections* of those groups, or other entirely unrelated subgroups. In other words, fairness conditions do not compose. **This is why fairness is such a hard problem to tackle!**

What can we do about this as clinicians?

The most important thing to help reduce bias is recognize that ***all real-world algorithms are biased***! How algorithms are biased and to what extent depend on your definition of harm

[5] Chouldechova A, Roth A. The frontiers of fairness in machine learning. arXiv Preprint. (2018). doi: 10.48550/arXiv.1810.08810

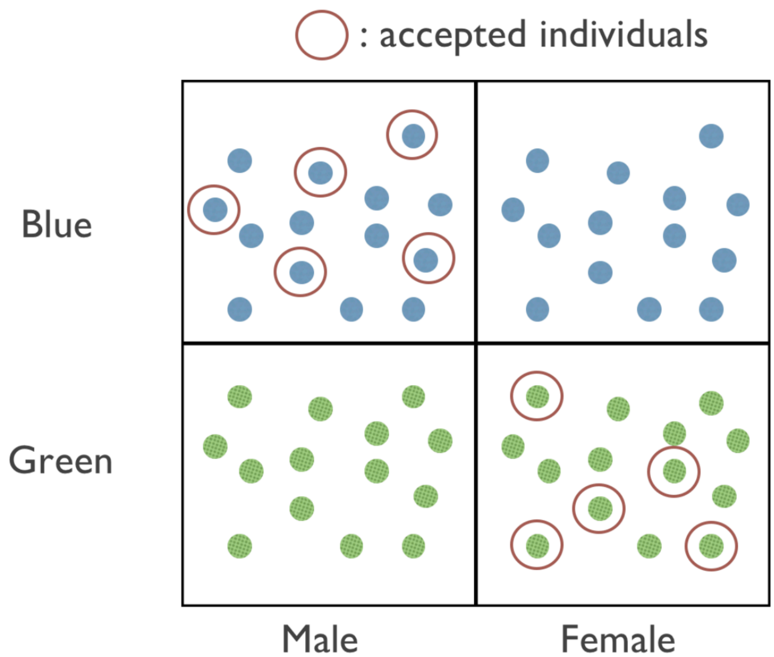

Figure 1: Fairness Gerrymandering. Adapted from Kearns M, Neel S, Roth A, Wu ZS. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. Proc Int Conf Mach Learn 80: 2564-72. (2018). Link to Paper

and the patient attribute(s) that you're focusing on. These definitions inherently differ between persons and scenarios. Recognizing our own biases in the algorithms used by both computers and humans is critical so that we make the best decisions for each individual patient.

Hands-On Tutorial

To better understand how bias can impact algorithms, let's take a look at a simple example of a binary classifier algorithm that seeks to predict whether a loan applicant will either (1) pay back the loan or (2) default on the loan.

> Will a loan applicant pay back the loan?
>
> 1. Go to this study from Wattenberg et al. at Google. Read through the article.
> 2. Play around with setting different thresholds for the algorithm to better understand the tradeoffs between the different metrics, such as accuracy, positive rate, and profit.
> 3. Simulate different loan decisions and different thresholds for two different groups: the blue and orange subpopulations. What happens when we use different conditions to set the different thresholds, like maximizing the bank's profit, using a group-unaware strategy, and ensuring equal opportunity?

In the above example, we looked at distributing loans. However, we can imagine a similar, more clinically relevant scenario: our own clinical algorithms for determining if a patient needs supplemental oxygen. For simplicity, let's assume that this decision is solely based on the patient's O2 saturation. Here are two potential strategies we can use:

> Does a patient need supplemental oxygen?
>
> **Important Context**: O_2 saturation measurements are less accurate for Black individuals.[6] More specifically, pulse oximeters often *overestimate* true O_2 saturation for

Black patients.
Two Strategies to Consider:

1. Group-Unaware Strategy: For all patients irregardless of skin color, we only start supplemental oxygen if SpO2 < 92%.
2. Equal Opportunity Strategy: The same fraction of Black and White patients should be on SpO2, so we will use the SpO2 < 92% cutoff for White patients and use a separate SpO2 < 94% cutoff for Black patients.

Which strategy is more biased? Is using race as an input into algorithms (as in the equal opportunity strategy) "good"? Why or why not?

Discussion Questions

1. Who do you agree with: ProPublica or the COMPAS developers? In other words, do you believe that the COMPAS algorithm is biased based on the evidence presented? Would you be comfortable having it used to determine the outcomes of the judicial system for close friends or family?
2. In the hands-on tutorial, we explored how even simple binary classifiers can be biased in scenarios such as determining loan repayment and supplemental oxygen requirements. What other analogous "binary classifier" clinical situations have you encountered? How did you decide your strategy on how to set your own "threshold" for positive and negative labels? Did your strategy vary between different patients?

How would your answer to this question change (or not) if instead we were deciding whether a patient was a lung transplant candidate? How about if we were deciding whether to start losartan therapy for newly diagnosed hypertension?

[6] Al-Halawani R, Charlton PH, Qassem M, et al. A review of the effect of skin pigmentation on pulse oximeter accuracy. Physiol Meas 44(5): 05TR01. (2023). doi: 10.1088/1361-6579/acd51a. PMID: 37172609

Summary

Bias is defined based on (1) protected attribute(s) and (2) a definition of harm. Because our definition of harm can vary from person-to-person, bias is often subjective. The impact of bias depends on the clinical scenario and the real-world implications of the definition of harm. It is important to recognize our own internal sources of bias in addition to the biases of clinical and computational algorithms.

Additional Readings

1. Nicoletti L and Bass D. Humans are biased. Generative AI is even worse. Bloomberg. (2023). Link to article
2. Kearns M, Roth A. Responsible AI in the wild: Lessons learned at AWS. Amazon Science Blog. (2023). Link to article
3. Evaluating Model Fairness. Arize Blog. (2023). Accessed 19 May 2024. Link to article
4. List JM, Palevsky P, Tamang S, et al. Eliminating algorithmic racial bias in clinical decision support algorithms: Use cases from the Veterans Health Administration. Health Equity 7(1): 809-16. (2023). doi: 10.1089/heq.2023.0037. PMID: 38076213
5. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: Challenges and mitigation strategies. npj Digit Med 6(113). (2023). doi: 10.1038/s41746-023-00858-z. PMID: 37311802