生物医学工程学院
School of Biomedical Engineering

上海科技大学
ShanghaiTech University

Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 6: Bias Issues in Medical AI and Algorithmic Fairness

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)

Assistant Professor of Biomedical Engineering

ShanghaiTech University

October 17, 2025

Learning Objectives of This Lecture

- **Understand the definition and types of Bias**
- Understand 8 algorithmic fairness metrics
- Know 3 types of bias mitigation methods

Simpson’s Paradox

- Berkley gender bias in the 1970s.
- The university admitted men at higher rates.
- All departments admitted women at higher rates.
- Who is correct? Does gender bias exist?

	Men		women
Dept. A	0/10	<	50/100
Dept. B	70/100	<	10/10
Total	70/110	>	60/110

Policing

- Predictive policing uses AI to forecast crime likelihood and proactively police areas.
- Data is typically drawn from prior-arrest databases.
- This creates a feedback loop.
- Potential bias in arrests.

Definition of Bias

- 1. Prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.
- 2. A systematic distortion of a statistical result due to a factor not allowed for in its derivation.

Bias from a Data Perspective

- **Sampling Bias**: Occurs when the sample data is not representative of the population intended to be analyzed
- **Survivorship Bias**: Focusing on data from “survivors” of a process while ignoring those that did not make it through
- **Data Collection Bias**: Bias introduced during the data collection process due to inconsistent or flawed methodologies
- **Reporting Bias**: Arises when only certain outcomes or data points are reported, often those that support a particular hypothesis

Bias from a Data Perspective

- **Social Desirability Bias**: Respondents provide answers that are more socially acceptable than their true thoughts or behaviors
- **Publication Bias**: Studies with significant/positive results are more likely published, skewing perception of research outcome
- **Historical Bias**: Results from biases present in historical data that are perpetuated in current models
- **Algorithmic Bias**: Bias introduced by the design and functioning of algorithm itself

Sampling Bias

- **Scenario**: A tech company is developing an AI-based facial recognition system for gender and uses a dataset predominantly composed of images from public figures and celebrities.
- **Bias**: This dataset is likely to underrepresent older individuals, people of varying attractiveness, and ethnic minorities. As a result, the AI model trained on this dataset may perform poorly when recognizing faces outside these demographic groups.
- **Implication**: The facial recognition system may exhibit significant inaccuracies and higher error rates for underrepresented groups, leading to biased and unreliable results in practical applications.

Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency 2018 Jan 21 (pp. 77-91). PMLR.

Survivorship Bias

- **Scenario**: During WWII, returning aircraft were analyzed for where to add amor. They observed damage on wings and fuselage, and thus suggested reinforcing these areas.
- **Bias**: This analysis only included planes that survived and returned from missions. The missing data were from planes that were shot down and did not return, which might have been hit critical areas like the engines.
- **Implication**: Focusing on the surviving aircraft led to incorrect conclusions. The real vulnerabilities were in the parts that, when hit, caused planes to be lost.

Mangel M, Samaniego FJ. Abraham Wald's work on aircraft survivability. Journal of the American Statistical Association. 1984 Jun 1;79(386):259-67.

Social Desirability Bias

- **Scenario**: A tech company is developing a sentiment analysis AI to gauge public opinion on sensitive topics by collecting survey data on controversial issues like racial discrimination or political views.
- **Bias**: Respondents may provide socially acceptable answers rather than their true opinions to avoid judgement or backlash, leading to social desirability bias.
- **Implication**: people are going to report what they think is the right answer as opposed to what they truly believe, especially in something like customer survey or sentiment analysis.

Krumpal I. Determinants of social desirability bias in sensitive surveys: a literature review. Quality & quantity. 2013 Jun;47(4):2025-47.

Historical Bias

- **Scenario**: A company develops a hiring algorithm designed to screen resumes and predict job performance based on historical hiring data.
- **Bias**: Training data predominantly includes resumes of employees who were hired and performed well in the past, which may reflect historical biases favoring certain demographics. The algorithm may favor resumes that resemble those of historically preferred candidates, while disadvantaging equally qualified candidates from underrepresented groups.
- **Implication**: It's very difficult when you're developing a selection tool to use your existing population.

Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 conference on fairness, accountability, and transparency 2020 Jan 27 (pp. 469-481).

Algorithmic Bias

- **Scenario**: An AI system predicts the likelihood of patients developing complications after surgery, using preoperative health data
- **Bias**: The algorithm is **trained** on data where certain demographic groups (e.g., younger patients or those with fewer comorbidities) are overrepresented. If the **model** relies heavily on these characteristics, it may inaccurately predict lower risk for older patients or those with more complex medical histories, leading to under-preparation and potentially poorer outcomes.
- **Implication**: The AI system may fail to predict complications for diverse patient groups that are not like it was trained on.

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine. 2018 Dec 18;169(12):866-72.

Ways to Mitigate Bias

- Ensure the sample is representative and randomly selected.
- Use validated and reliable measurement instruments.
- Train data collectors thoroughly to minimize observer bias.
- Collect data from multiple sources and contexts.
- Transparently report all data, including null and negative results.
- Regularly **audit** and **evaluate** data and algorithms for bias.
- Include diverse perspectives in the data collection and analysis process.

Biases in Computational Medicine Studies

- Examples
  - Associations between Framingham risk factors and cardiovascular events are significantly different across ethnic groups.
  - Video stream analysis algorithms are challenging for Asian individuals.
  - Undiagnosed silent hypoxemia, detected from pulse oximetry, occurred three times in Black people due to their dark skin.

Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, Wang F. Algorithmic fairness in computational medicine. EBioMedicine. 2022 Oct 1;84.

Computational bias

- Data bias
    - Patients of low socioeconomic status may have limited access to health care
    - Sampling bias (Selection bias)
        - Melanoma detection algorithms based on classification of skin lesion images may perform poorly on dark-pigmented skin if the training images contain predominantly lighter skin.
        - Face2Gene, a machine learning algorithm to recognize Down syndrome based on facial images, performed much better in Caucasian than in African.
    - Allocation bias
        - Emulate clinical trials with real world data such as EHRs

Computational bias

- Data bias
  - Attrition bias
    - It can occur if there are systematic differences in the way different groups of participants are recruited or are dropped from a study.
  - Publication bias
    - It occurs when the decision to publish a study depends on its own results.
    - It makes people overestimate the effectiveness of specific treatments or models.
- Measurement bias
  - When the data are labeled inconsistently
  - When Diseases are collected or measured inaccurately

Computational bias

- Measurement bias
  - When the data are labeled inconsistently
  - When Diseases are collected or measured inaccurately
  - Response bias
    - When respondents tend to give inaccurate or even wrong answers on self-reported questions.
    - Example 1: People might tend to always rate themselves favorably or feel pressured to provide socially acceptable answers.
    - Example 2: Misleading questions can lead to biased answers.
    - Example 3: Demographic groups who are willing to answer survey questions are sometimes different from those who are not.
  - Algorithm bias

A case study

- Build an alerting algorithm in ICU setting (e.g., for developing sepsis)
- Machine learning algorithm based on the patient's EHR and the patient's race.
- Consider only two demographic groups (e.g., Black or white)
  - A in {0, 1}: Protected attribute
  - X: Observable attributes
  - U: Relevant latent attributes not observed
  - Y in {0, 1}: Outcome to be predicted
  - $\hat{Y}$ in {0,1}: Prediction

Fairness metrics

1. Unawareness
   - No protected attribute A is explicitly used in the decision-making
   - *A*: Protected attribute (e.g., race)
   - $\hat{Y}=f(X, A)=f(X)$

2. Demographic Parity
   - The outcomes must be equal
   - $P(\hat{Y} = y | A=0) = P(\hat{Y} = y | A=1)$, y in {0,1}
   - *P*: Proportion or Percentage

Fairness metrics

3. Equalized Odds
    - Different groups deal with similar odds
    - $P(\hat{Y} = 1 | A = 0, Y = y) = P(\hat{Y} = 1 | A = 1, Y = y)$, y in $\{0,1\}$
    - The true positive rates (of those who actually developed sepsis, how many were correctly predicted to be positive) and false positive rates in both demographic groups are equal

4. Equal Opportunity
    - The true positive rates in both groups are equal.
    - $P(\hat{Y} = 1 | A = 0, Y = 1) = P(\hat{Y} = 1 | A = 1, Y = 1)$

Fairness metrics

5. Individual Fairness
   - Similar individuals have similar predictions.
   - Individuals *i* and *j*, if distance *d*(*i*, *j*) is small, then $|\hat{Y}(i) - \hat{Y}(j)|$ is small.

6. Counterfactual Fairness
   - The predicted outcome does not change if a patient from one demographic group is assigned to the other demographic group
   - $P(\hat{Y} = y | A=0, X=x) = P(\hat{Y} = y | A=1, X=x)$ for all *x* and *y*
   - Counterfactual reasoning may negatively affect the process of causality identification (e.g., *Y* is dependent on *A*)

Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. *Advances in neural information processing systems*. 2017;30.

Fairness metrics

7. Average odds difference (AOD)
   - AOD=1/2(Average TPR Difference + Average FPR Difference)

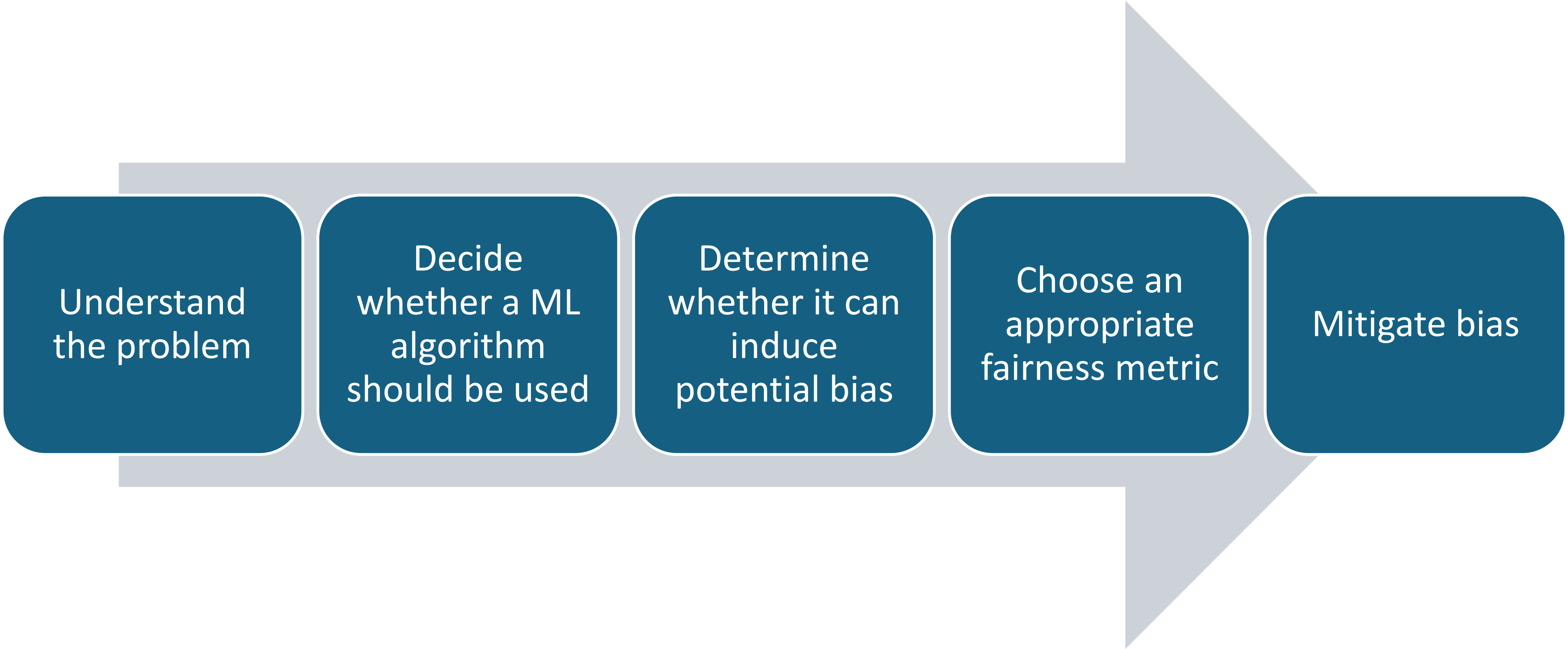
     = ½(|$P(\hat{Y}$ =1|A=0, Y=1) - P($\hat{Y}$ =1|A=1, Y=1)|

     +|$P(\hat{Y}$ =1|A=0, Y=0) - $P(\hat{Y}$ =1|A=1, Y=0)|)

8. Disparate impact (DI)
   - $DI_{ij} = \min\left(\frac{P\left(\hat{Y} = 1 \middle| A = i, Y = 1\right)}{P\left(\hat{Y} = 1 \middle| A = j, Y = 1\right)}, \frac{P\left(\hat{Y} = 1 \middle| A = j, Y = 1\right)}{P\left(\hat{Y} = 1 \middle| A = i, Y = 1\right)}\right), \mathrm{i}, \mathrm{j} = 0{,}1, \mathrm{i} \neq \mathrm{j}$
   - $DI = \max DI_{ij}$

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining 2015 Aug 10 (pp. 259-268).

Fairness-aware problem solving



Understand the problem

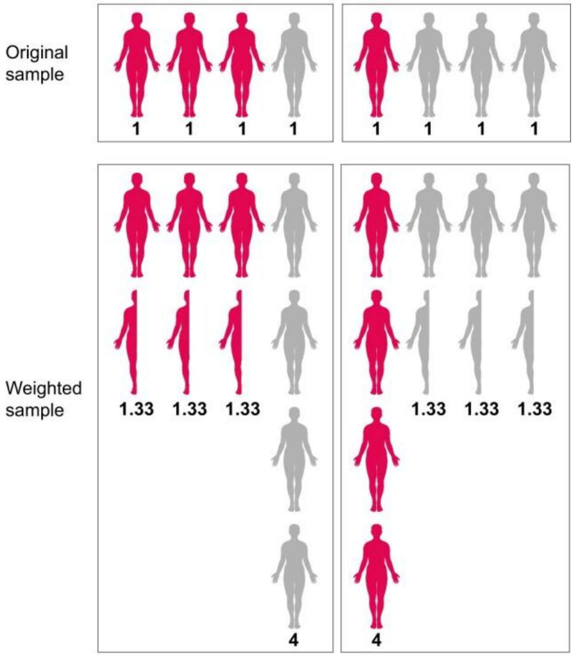
Decide whether a ML algorithm should be used

Determine whether it can induce potential bias

Choose an appropriate fairness metric

Mitigate bias

Bias mitigation



1. Pre-processing
   - ○ Choice of sampling (Resampling)
     - Ensure that all demographic groups are properly and proportionately represented in the training dataset
     - Under-sample the majority group or oversample the minority group
     - Collecting more data is better
   - ○ Reweighting
     - Inverse propensity score weighting
     - w(1,1)=1/P(A=1|Y=1)=1.25
     - w(1,0)=1/P(A=1|Y=0)=1.5
     - w(0,1)=1/P(A=0|Y=1)=5
     - w(0,0)=1/P(A=0|Y=0)=3

Y

A

	Case(1)	Control(0)
White(1)	80	200
Black(0)	20	100

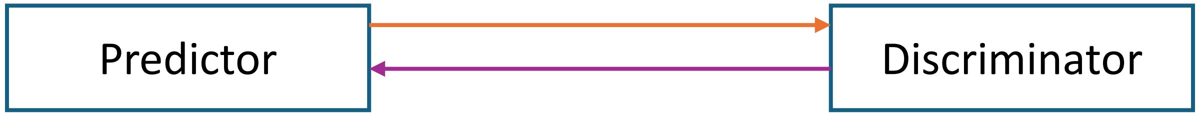
Y

A

	Case(1)	Control(0)
White(1)	100	300
Black(0)	100	300

Bias mitigation

2. In-processing
   - Prejudice remover
     - Make predictions be independent from the protected attribute
   - Adversarial learning



Loss function: prediction error

Loss function: equalized odds bias

   - Interpretable models: reveals biased decision-making process
   - Independent learning
     - Trains a model for each protected group → Reduces the performance
     - Transfer learning → Align the sample distributions

Bias mitigation

3. Post-processing
   - o Equalized odds post-processing
     - Changing output labels to achieve the equalized odds objective
   - o Adjust the risk scores of the instances in the disadvantaged group
   - o Adjust the ranking order of the samples across different protected groups
   - o Causal analysis approach

Popular software libraries

Project	Developer	Year	Description	Publication
FairMLHealth	KenSci	2020	Tools and tutorials for evaluating bias in healthcare machine learning.	GitHub
AIF360	IBM	2019	Fairness metrics for datasets and machine learning algorithms, interpretation of the metrics, and approaches for reducing bias in datasets and models. It is available in both Python and R.	IBM Journal of Research and Development
Fairlearn	Microsoft	2020	A Python package to evaluate fairness and mitigate any observed inequities.	Microsoft Tech
Fairness-comparison	Sorelle et al.	2019	Compare fairness-aware machine learning techniques. It aims to facilitate benchmarking of fairness-aware machine learning algorithms.	ACM FAccT
MEASURES	Cardoso et al.	2019	A benchmark framework for assessing discrimination-aware models.	AAAI/ACM CAES
Fairness Indicators	Google	2024	A suite of tools built on top of TensorFlow Model Analysis that enable regular evaluation of fairness metrics in product pipelines.	Google Colab
ML-fairness-gym	Google	2020	A general framework for studying and exploring long-term equity effects in carefully constructed simulation scenarios where learning subjects interact with the environment over time.	Google Blog
Themis-ml	Niels Bantilan	2017	A Python library built on top of pandas and sklearn that implements fairness-aware machine learning algorithms.	J. of Technology in Human Services
FairML	Julius Adebayo	2017	A Python toolkit for auditing machine learning model deviations.	Github

Readings Due on October 22

- Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, Wang F. Algorithmic fairness in computational medicine. EBioMedicine. 2022 Oct 1;84.
  - https://www.thelancet.com/pdfs/journals/ebiom/PIIS2352-3964(22)00432-7.pdf
- Optional
  - Kearns M, Roth A. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press; 2019 Oct 4. (Ch.2)
  - 《Ethics of medical AI》 pp. 117-132.
  - Dunkelau J, Leuschel M. Fairness-aware machine learning: An extensive overview. 2019. https://stups.hhu-hosting.de/downloads/pdf/fairness-survey.pdf
  - Molnar, Christoph. Interpretable machine learning. 2020. (Ch. 5) https://christophm.github.io/interpretable-ml-book/
  - Lundberg, S. M., & Lee, S. I. A unified approach to interpreting model predictions. NeurIPS. 2017 (Original SHAP paper).

Feedback Survey

- One thing you learned or felt was valuable from today’s class & reading
- Muddiest point: what, if anything, feels unclear, confusing or “muddy”
- https://www.wjx.cn/vm/hX0mIro.aspx