Bias and Fairness in Medicine

Sustainable AI in practice
25.8.2020

Aasa Feragen,
Section for Image Analysis and Computer Graphics
DTU Compute
Bias in healthcare AI

The female problem: how male bias in medical trials ruined women’s health

Centuries of female exclusion has meant women’s diseases are often missed, misdiagnosed or remain a total mystery.

Racial bias in a medical algorithm favors white patients over sicker black patients

Scientists discovered racial biases in a widely used medical algorithm that predicts which patients will have complex health needs. (iStock)
Case 1. A simulated example

Imagine using predicted depression risk scores for prioritizing resources such as referral to a psychologist
Bias in algorithms: A toy illustration

It is well known that:

▶ Depression is diagnosed more frequently in women than in men
▶ This can partially be explained by different cultural perceptions of women and men (Sigmon et al, 2005)
▶ If the diagnostic criteria are adapted to male symptoms, then the prevalence of depression among men increases (Martin et al, 2013)

If the data used for training ML algorithms to predict depression risk is skewed, then the trained algorithm will produce skewed predictions – it will be unfair. Let’s simulate this.
Bias in algorithms: A toy illustration

Imagine a disease model where

- Disease is scored from 0=healthy to 10=severe
- A true diagnosis corresponds to true score > 5
- Blue people (e.g. men) are systematically underdiagnosed due to differences in cultural perceptions of gender (e.g. as with depression, Sigmon et al. 2005)
Bias in algorithms: A toy illustration

Setting a diagnostic threshold at diagnosed disease score = 5, we see that:

- For the red group, we have no false diagnoses.
- For the blue group, false negative diagnoses are made.
Bias in algorithms: A toy illustration

Setting a diagnostic threshold at diagnosed disease score = 5, we see that:

- For the red group, we have no false diagnoses.

![Graph showing bias in algorithms with true disease score and diagnosed disease score axes.]
Bias in algorithms: A toy illustration

Setting a diagnostic threshold at diagnosed disease score $= 5$, we see that:

- For the red group, we have no false diagnoses
- For the blue group, false negative diagnoses are made
Bias in algorithms: A toy illustration

**Solution:** Population-specific thresholds
Bias in algorithms: A toy illustration

**Solution:** Population-specific thresholds
Solution: Population-specific thresholds
Bias in algorithms: A toy illustration

In a different disease model, the diagnostic criteria are more appropriate for the red group than for the blue, as in (Martin et al., 2013)

- Here, the score=5 threshold creates false positives and negatives in the blue group
Bias in algorithms: A toy illustration

Below, see the group-wise diagnostic accuracy for the two different classes

- We are incapable of reaching perfect accuracy for the blue group
- Two thresholds for the red group give the same accuracy as the best seen for the blue group
Bias in algorithms: A toy illustration

Let’s see what those thresholds do:

Note: Although we have sacrificed performance in the red group, we still have a bias in our errors.
Bias in algorithms: A toy illustration

Let’s see what those thresholds do:

- Blue group has positive TP, TN, FP and FN

---

Note: Although we have sacrificed performance in the red group, we still have a bias in our errors.
Bias in algorithms: A toy illustration

Let’s see what those thresholds do:

- Blue group has positive TP, TN, FP and FN
- Red group has positive TP, TN and FP, but no FN
Bias in algorithms: A toy illustration

Let’s see what those thresholds do:

▶ Blue group has positive TP, TN, FP and FN
▶ Red group has positive TP, TN and FN, but no FP

![Graph showing the relationship between true disease score and diagnosed disease score, with TP, TN, FP, and FN regions highlighted.]
Bias in algorithms: A toy illustration

Let’s see what those thresholds do:

- Blue group has positive TP, TN, FP and FN
- Red group has positive TP, TN and FN, but no FP
- **Note:** Although we have *sacrificed performance* in the red group, we still have a *bias* in our errors.
Case 2: Image-based diagnosis of thoracic disorders
Bias in algorithms: A computer assisted diagnosis example

- State-of-the-art CNN diagnosing thoracic diseases from X-ray\(^1\)

\(^1\)Larrazabal et al, PNAS 2020
Bias in algorithms: A computer assisted diagnosis example

- State-of-the-art CNN diagnosing thoracic diseases from X-ray\(^1\)
- Increased % females $\Rightarrow$ improved female test diagnosis

Figure: Diagnostic accuracy of Pneumothorax for female test subjects as a function of % females in training set

---

\(^1\)Larrazabal et al, PNAS 2020
Bias in algorithms: A computer assisted diagnosis example

- State-of-the-art CNN diagnosing thoracic diseases from X-ray\(^1\)
- Increased % females $\implies$ improved female test diagnosis
- Increased % females $\implies$ decreased male test diagnosis

Figure: Diagnostic accuracy of Pneumothorax for male test subjects as a function of % females in training set

\(^1\)Larrazabal et al, PNAS 2020
Bias in algorithms: A computer assisted diagnosis example

- State-of-the-art CNN diagnosing thoracic diseases from X-ray\textsuperscript{1}
- Increased % females $\Rightarrow$ improved female test diagnosis
- Increased % females $\Rightarrow$ decreased male test diagnosis
- *Predictor trained only on females performs better on men*

![Figure: Diagnostic accuracy of Pneumothorax for female (left) and male (right) test subjects as a function of % females in training set](Image)

\textsuperscript{1}Larrazabal et al, PNAS 2020
Sources of bias in ML algorithms

- Discrimination embedded in training data
- Imbalanced training data
- Different levels of label noise (diagnosis errors) give different training conditions for different groups
- Different feature distributions in different groups (different disease patterns and/or anatomical features) give different training conditions for different groups

Additionally:
- In medicine, our entire knowledge base is based on the white, male anatomy
Algorithms are new, bias is not
Algorithms are new, bias is not

Algorithms come with potential for early discovery of bias
What *is* bias?

- Over- or under-representation is not a discriminating bias in itself – for instance, breast cancer *is* more prevalent in women than in men.
- Data- and algorithmic bias refers to *systematic errors* that differ between groups.
- In order to detect this bias, we need to access the true labels (e.g. true diagnosis).
- This is often impossible – thus, our analysis depends on finding a reliable *proxy* for the true label.
Quality of labels: Proxy variables for bias detection and better training?

COMPAS case\(^2\): Racial bias in predicting risk of re-offense among US criminals.

Proxy variable for criminality used in COMPAS: previous verdicts; in analysis that documented unfairness: 2-year re-offense.

\(^2\)https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
Quality of labels:
Proxy variables for bias detection and better training?
Quality of labels: Proxy variables for bias detection and better training?

Open problem: Proxy variables for diagnosis?

Observed diagnosis    True diagnosis    Measured well-being
Quality of labels: 
Proxy variables for bias detection and better training?

Open problem: Proxy variables for diagnosis?

Survival? Perceived quality of life? Continued need for treatment?
Algorithmic fairness

A number of candidate definitions for a “fair ML algorithm” have been proposed:

- **Predicted outcome should be independent of sensitive variables**
Algorithmic fairness

A number of candidate definitions for a “fair ML algorithm” have been proposed:

▷ Predicted outcome should be independent of sensitive variables
Algorithmic fairness

A number of candidate definitions for a “fair ML algorithm” have been proposed:

▶ Predicted outcome should be independent of sensitive variables

Gender Bias in Diagnostic Criteria for Personality Disorders: An Item Response Theory Analysis

J. Serena Jane, Thomas F. Oltmanns, Susan C. South, and Eric Turkheimer

The publisher’s final edited version of this article is available at J Abnorm Psychol
See other articles in PMC that cite the published article.

Abstract

The authors examined gender bias in the diagnostic criteria for Diagnostic and Statistical Manual of Mental Disorders (4th ed., text revision; American Psychiatric Association, 2000) personality disorders. Participants (N = 599) were selected from 2 large, nonclinical samples on the basis of information from self-report questionnaires and peer nominations that suggested the presence of personality pathology. All were interviewed with the Structured Interview for DSM-IV Personality (B. Pfohl, N. Blum, & M. Zimmerman, 1995). Using item response theory methods, the authors compared data from 315 men and

▶ Individual fairness: Similar subjects should get similar predictions
Algorithmic fairness

A number of candidate definitions for a “fair ML algorithm” have been proposed:

▶ **Group fairness**: Different groups should have same predictive performance
Algorithmic fairness

A number of candidate definitions for a “fair ML algorithm” have been proposed:

- **Group fairness**: Different groups should have same predictive performance

Do we allow lowering diagnostic performance for women when it is hard to diagnose men?
Algorithmic fairness

A number of candidate definitions for a “fair ML algorithm” have been proposed:

- **Group fairness**: Different groups should have same predictive performance

![Graph showing diagnostic accuracy for Blue and Red groups at threshold t=5]

Do we allow lowering diagnostic performance for women when it is hard to diagnose men?

- **Equalized odds/Equality of opportunity**: Different groups should have similar rates of specific error types
What is fairness?

- Fairness is more than accuracy and error rates – access to resources
- Current “fair” algorithms would likely be considered unethical, possibly illegal
Fairness for healthcare AI: An open problem

What is fairness?
- Fairness is more than accuracy and error rates – access to resources
- Current “fair” algorithms would likely be considered unethical, possibly illegal

AI/ML can be part of the solution
- **Important:** Bias did not come with the algorithms – it was already there in the data
- Trained ML algorithms come with a potential for discovering bias before a single real prediction is made – as opposed to with biased human operators