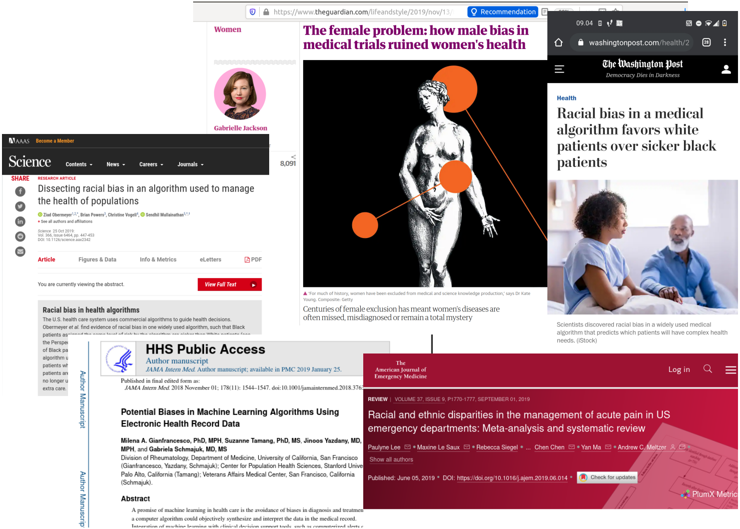
Bias and Fairness in Medicine

Sustainable AI in practice
25.8.2020

Aasa Feragen,
Section for Image Analysis and Computer Graphics
DTU Compute

Bias in healthcare AI



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)

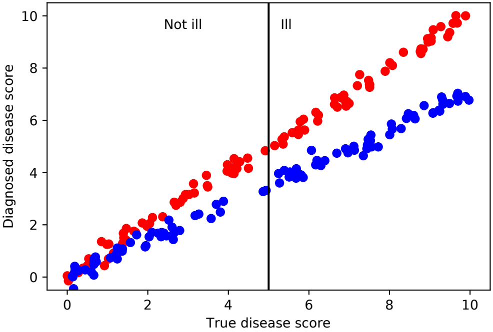
- If the diagnostic criteria are adapted to male symptoms, then the prevalence of depression among men increases (Martin et al, 2013)

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

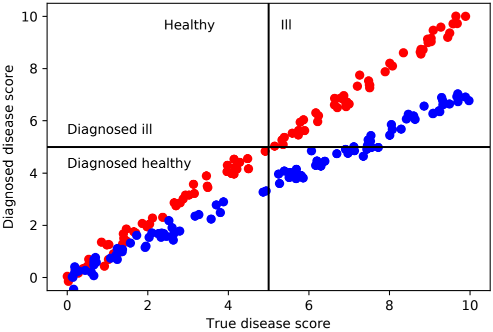
Imagine a disease model where

- 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

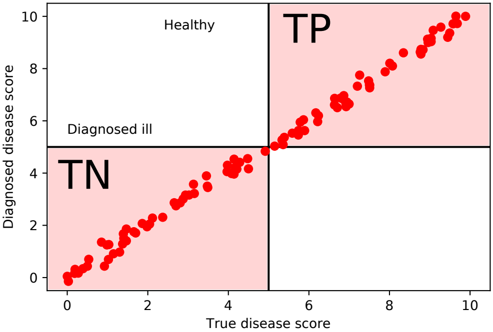
Setting a diagnostic threshold at diagnosed disease score = 5, we see that:



Bias in algorithms: A toy illustration

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

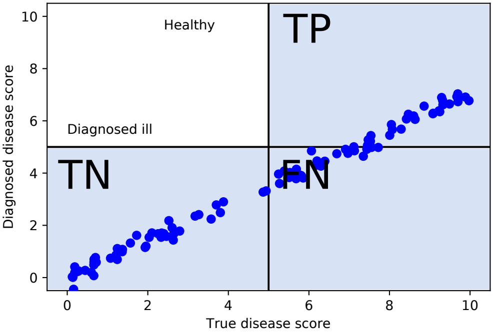
- For the red group, we have no false diagnoses



Bias in algorithms: A toy illustration

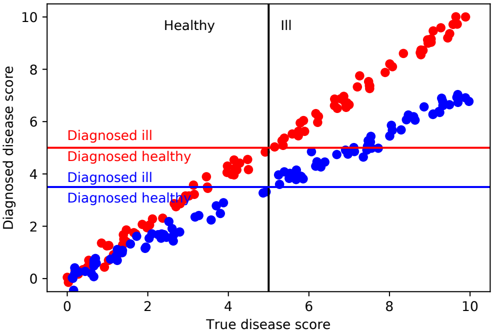
Setting a diagnostic threshold at diagnosed disease score = 5, we see that:

- For the red group, we have no false diagnoses
- For the blue group, false negative diagnoses are made



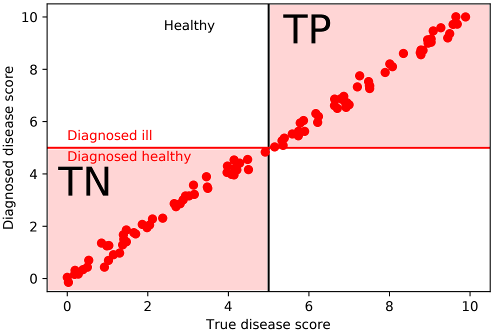
Bias in algorithms: A toy illustration

Solution: Population-specific thresholds



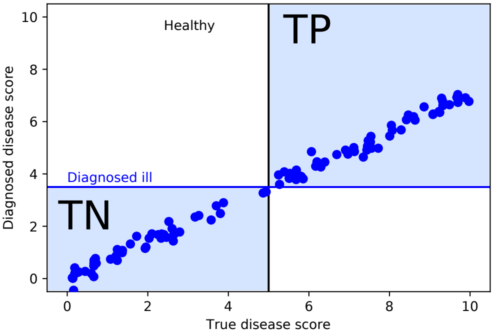
Bias in algorithms: A toy illustration

Solution: Population-specific thresholds



Bias in algorithms: A toy illustration

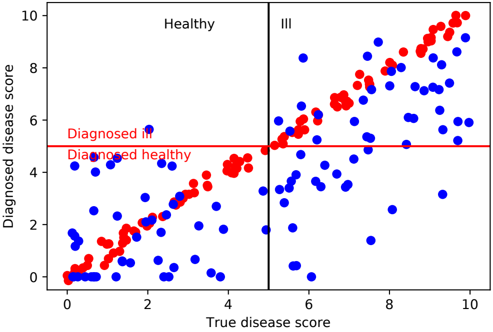
Solution: Population-specific thresholds



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)

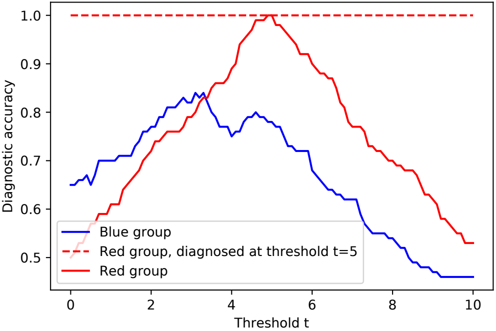
- Here, the score=5 threshold creates false positives and negatives in the blue group



Bias in algorithms: A toy illustration

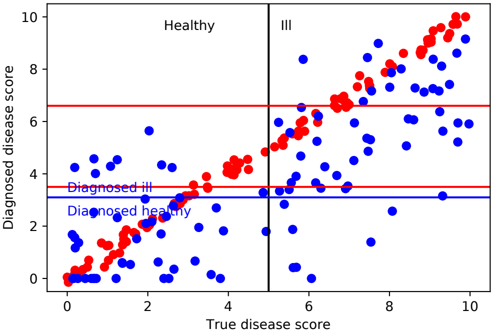
Below, see the group-wise diagnostic accuracy for the two different classes

- We are uncapable 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

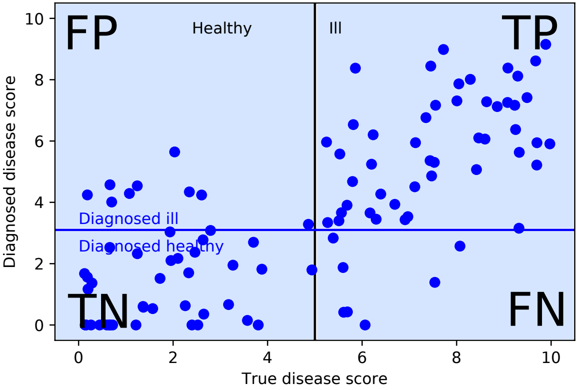
Let's see what those thresholds do:



Bias in algorithms: A toy illustration

Let's see what those thresholds do:

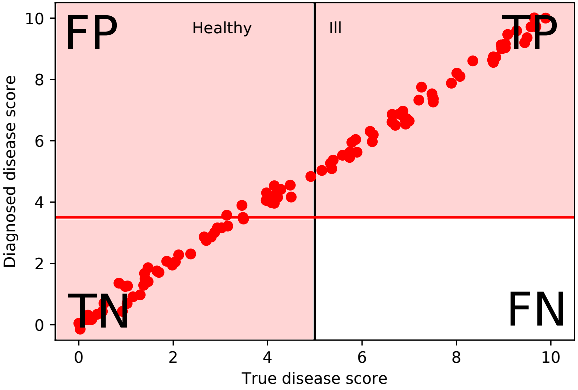
- Blue group has positive TP, TN, FP and FN



Bias in algorithms: A toy illustration

Let's see what those thresholds do:

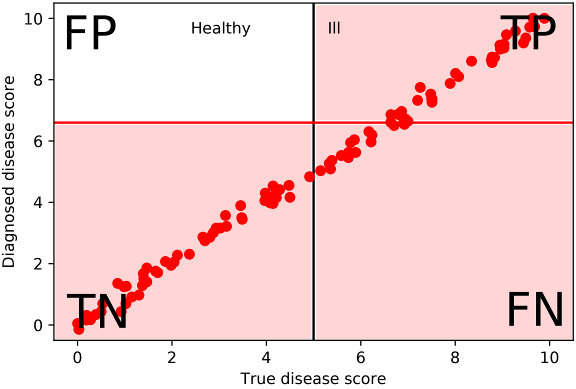
- Blue group has positive TP, TN, FP and FN
- Red group has positive TP, TN and FP, but no FN



Bias in algorithms: A toy illustration

Let's see what those thresholds do:

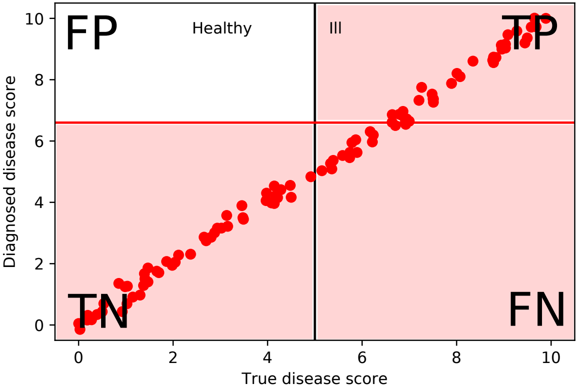
- Blue group has positive TP, TN, FP and FN
- Red group has positive TP, TN and FN, but no FP



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 ⇒ improved female test diagnosis

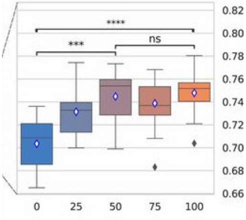


Figure: Diagnostic accuracy of Pneumothorax for female test subjects as a function of % females in training set

[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 ⇒ improved female test diagnosis
- Increased % females ⇒ decreased male test diagnosis

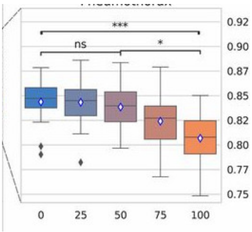


Figure: Diagnostic accuracy of Pneumothorax for male test subjects as a function of % females in training set

[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[1]
- Increased % females $\Rightarrow$ improved female test diagnosis
- Increased % females $\Rightarrow$ decreased male test diagnosis
- *Predictor trained only on females performs better on men*

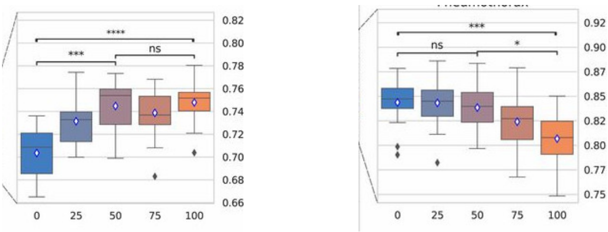
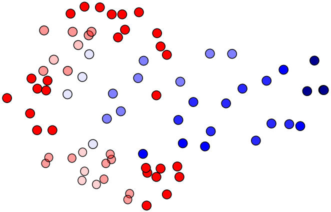


Figure: Diagnostic accuracy of Pneumothorax for female (left) and male (right) test subjects as a function of % females in training set

[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

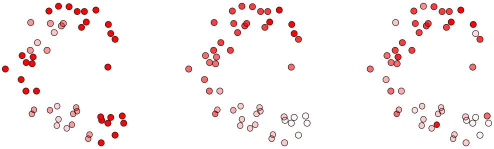
Algorithms are new, bias is not

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.



Observed diagnosis — True diagnosis — Measured well-being

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?

AAAS Become a Member

Science Contents News Careers Journals

SHARE

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Article | Figures & Data | Info & Metrics | eLetters | PDF

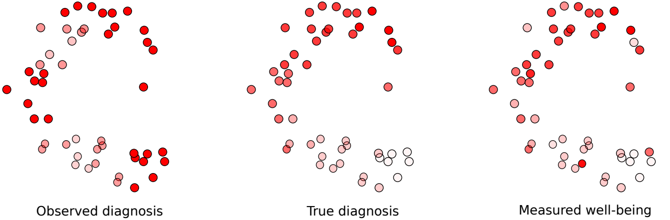
You are currently viewing the abstract. View Full Text

Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

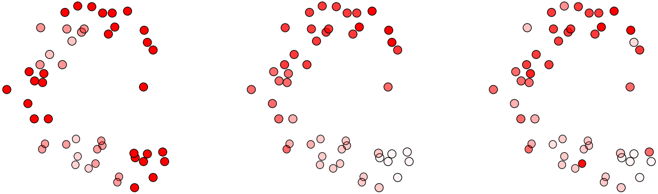
Quality of labels:
Proxy variables for bias detection and better training?

Open problem: Proxy variables for diagnosis?



Quality of labels:
Proxy variables for bias detection and better training?

Open problem: Proxy variables for diagnosis?



Observed diagnosis　　True diagnosis　　Measured well-being

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**

J Abnorm Psychol. Author manuscript; available in PMC 2015 Mar 25.
Published in final edited form as:
J Abnorm Psychol. 2007 Feb; 116(1): 166–175.
doi: 10.1037/0021-843X.116.1.166

PMCID: PMC4372614
NIHMSID: NIHMS670356
PMID: 17324027

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

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

Abstract

Go to:

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, 1997). Using item response theory methods, the authors compared data from 215 men and

Algorithmic fairness

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

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

J Abnorm Psychol. Author manuscript; available in PMC 2015 Mar 25.
Published in final edited form as:
J Abnorm Psychol. 2007 Feb; 116(1): 166–175.
doi: 10.1037/0021-843X.116.1.166

PMCID: PMC4372614
NIHMSID: NIHMS670356
PMID: 17324027

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

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

▸ Author information ▸ Copyright and License information Disclaimer

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

Go to:

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, 1997). Using item response theory methods, the authors compared data from 215 men and

- **Individual fairness: Similar subjects should get similar predictions**

Algorithmic fairness

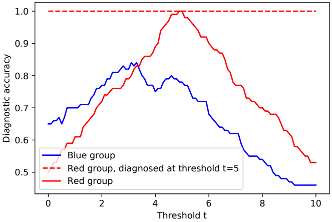
A number of candidate definitions for a “fair ML algorithm” have been proposed:

- **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**

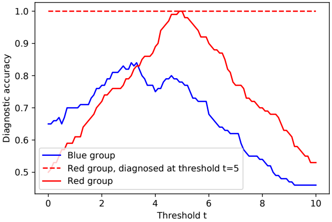


Do we allow lowering diagnostic performance for women when it is hard to diagnose men?

Algorithmic fairness

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

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



  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**

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

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