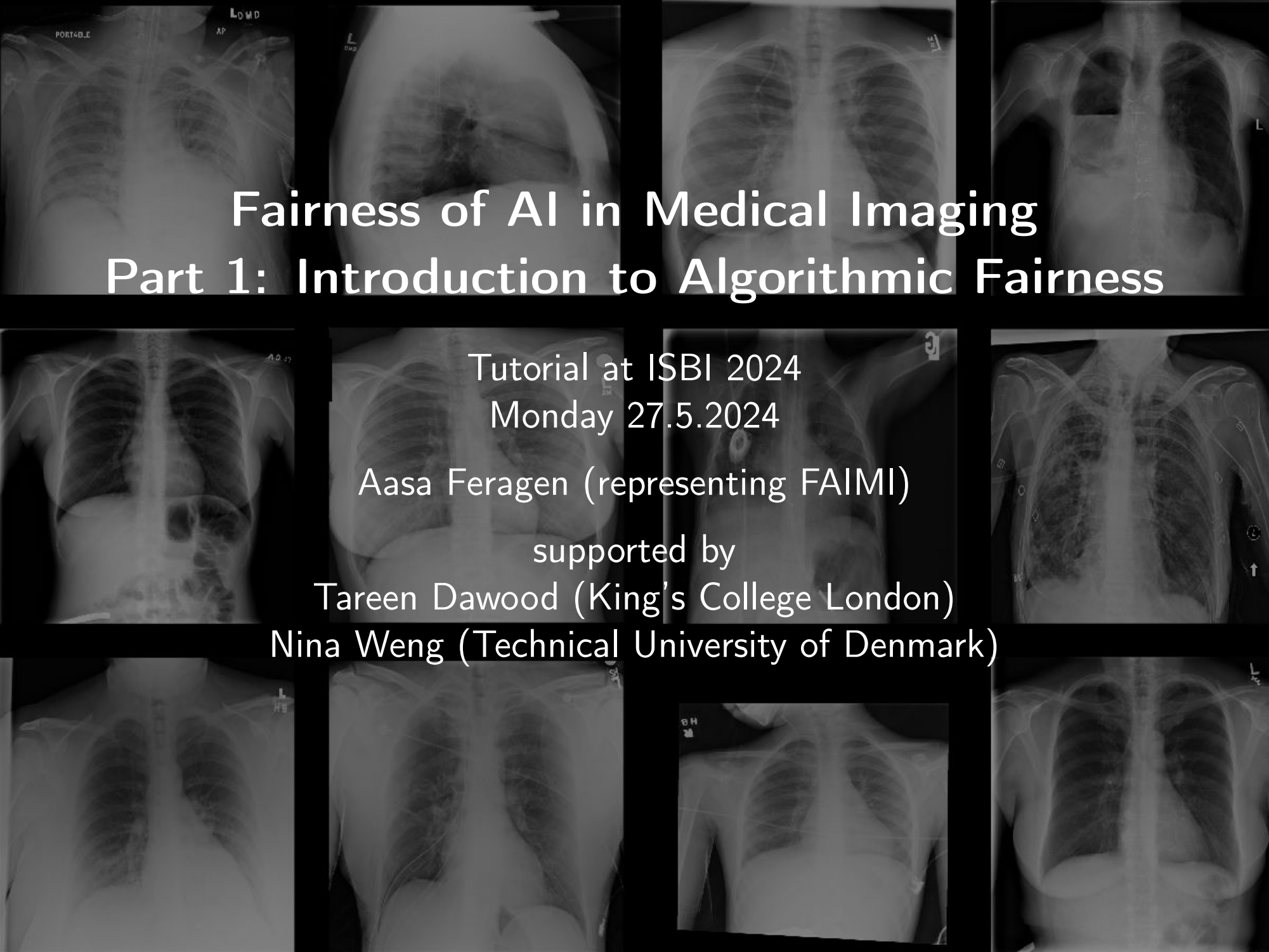
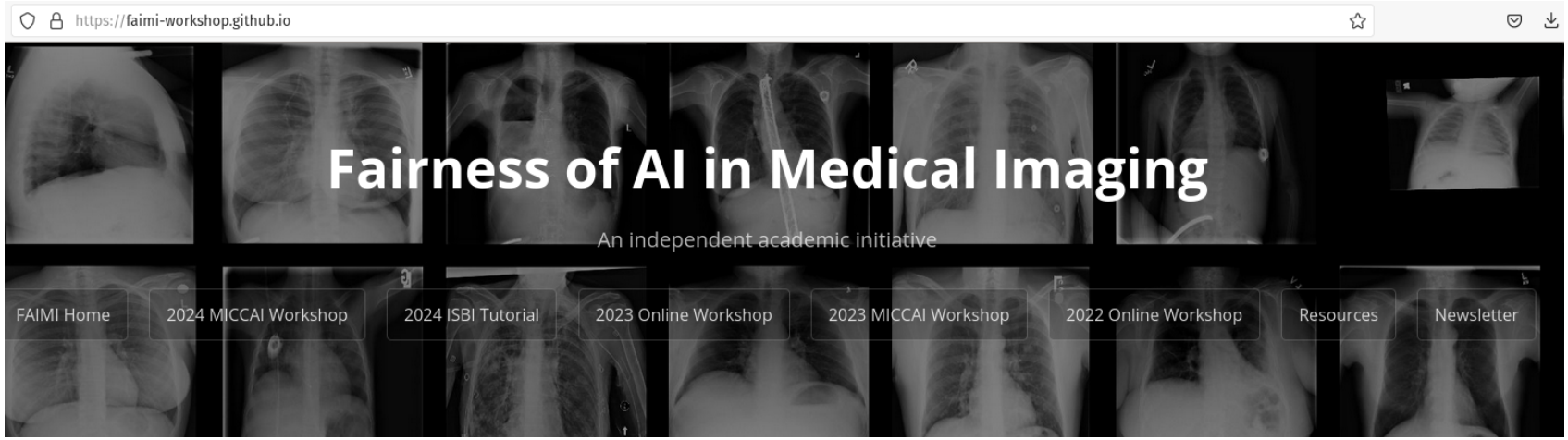


# Fairness of AI in Medical Imaging
# Part 1: Introduction to Algorithmic Fairness

Tutorial at ISBI 2024
Monday 27.5.2024

Aasa Feragen (representing FAIMI)

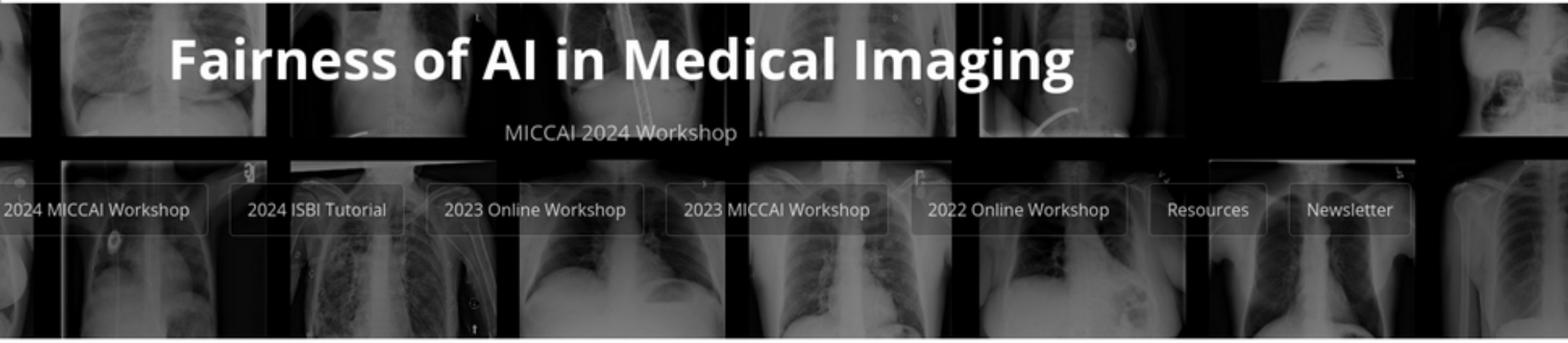
supported by
Tareen Dawood (King's College London)
Nina Weng (Technical University of Denmark)




Aasa Feragen
Technical University
of Denmark

Tareen Dawood
King's College London

Nina Weng
Technical University
of Denmark




## TL;DR

- Fairness & Medical Imaging workshop on **Oct 10th** at MICCAI 2024 (Morocco)
- The workshop will be only *fully* in-person (as in *no* remote live talk or *no* hybrid mode)
- Organized jointly with the workshop on Ethical and Philosophical Issues in Medical Imaging

## Call for Papers

We invite the submission of papers for

**FAIMI: The MICCAI 2024 Workshop on Fairness of AI in Medical Imaging**.

## Dates

*All dates are Anywhere on Earth.*

Full Paper Deadline: **June 24, 2024**

Notification of Acceptance: **July 15, 2024**

Camera-ready Version: **August 1, 2024**

Workshop: **October 10th, 2024**

For more information, check: https://faimi-workshop.github.io

# Tutorial overview:

Part 1 (13:00-14:30)

- Lecture 1: Introduction to Algorithmic Fairness
- Discussing case studies in groups

Part 2 (15:00-16:30)

- Wrapping up case studies (I will look into using mentimeter or similar)
- Lecture 2: Bias mitigation and pitfalls
- Lecture 3: Generative AI bias

# By the end of this first part of the tutorial, you should:

- Be familiar with different ways in which demographic bias can become embedded in machine learning models
- Be familiar with the three most common fairness criteria, and variants thereof
- Know how to turn fairness criteria into diagnostic criteria for algorithmic bias
- Be familiar with incompatibility results for fairness criteria

# Introduction to Algorithmic Fairness

Fairness definitions based on the classification chapter of `https://fairmlbook.org/`

# Fairness in AI

## Google apologizes after Photos app tags black couple as gorillas: Fault in image recognition software mislabeled picture

- **New York based computer programmer spotted images of himself and a friend had been labelled in an album marked 'gorillas' by Google Photos**
- **Google's image recognition software had automatically grouped the photos**
- **Google has said it is 'appalled' and 'genuinely sorry' for the mistake**
- **The company is working on ways to prevent similar errors in the future**

By RICHARD GRAY FOR MAILONLINE
**PUBLISHED:** 13:39 GMT, 1 July 2015 | **UPDATED:** 21:17 GMT, 1 July 2015

Share 4.9k shares 163 View comments

Google has been forced to apologise after its image recognition software mislabelled photographs of black people as gorillas.

The internet giant's new Google Photos application uses an auto-tagging feature to help organise images uploaded to the service and make searching easier.

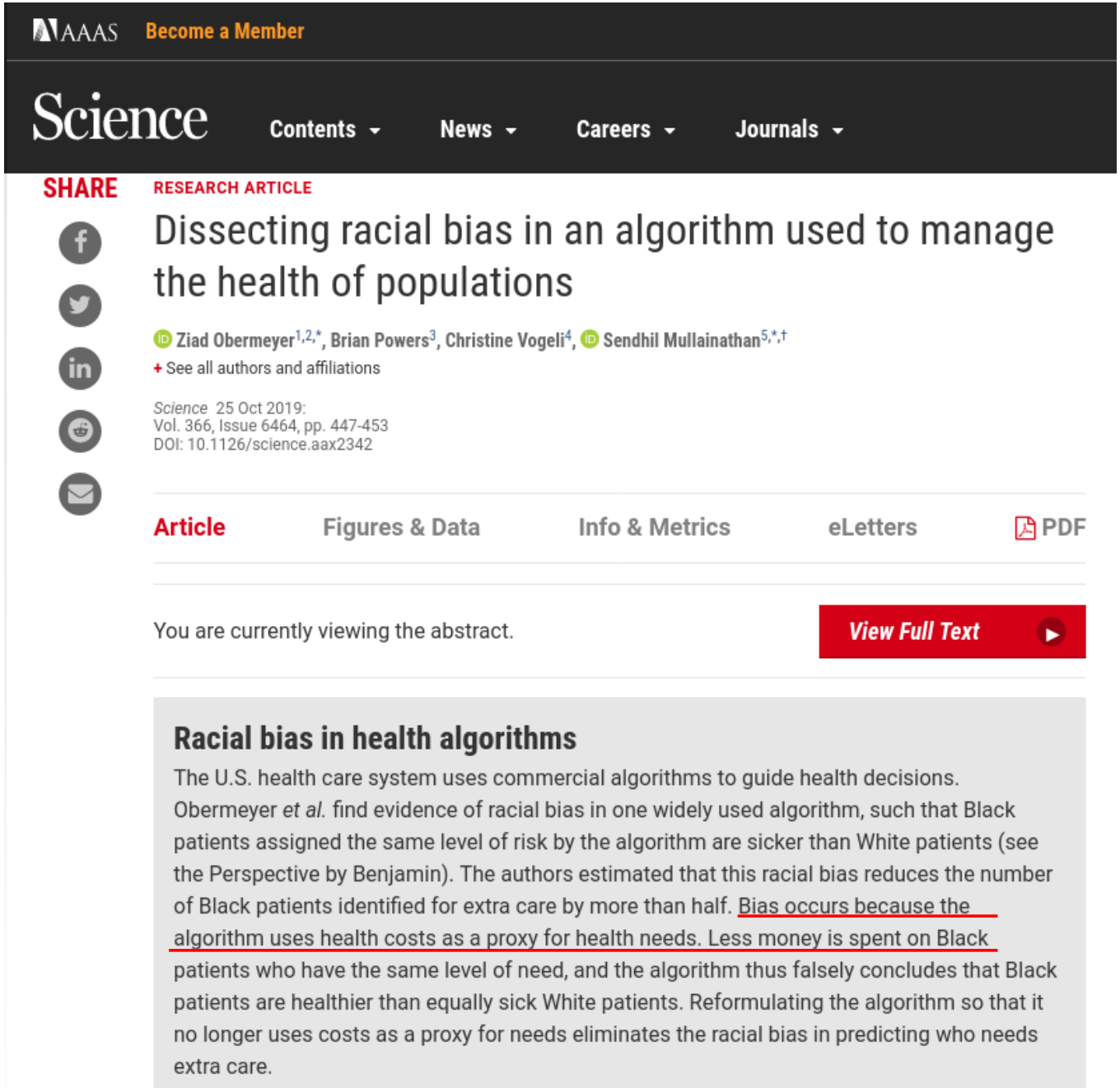
However the software has outraged users after it mislabelled images of a computer programmer and his friend as the great apes.


**Google has issued an apology after computer programmer Jacky Alcine, from New York, spotted photographs of himand a femalefriend hadbeen labelled as gorillas by Google Photos image recognition software. He sent aseries ofTweets to Google highlighting theproblem (like above) leading Google to issuea fix for the problem**

Google said it was 'appalled' and 'genuinely sorry' for the mistake.

The fault comes just over a month after Flickr's autotagging system placed potentially offensive tags on images including mislabelling concentration camps as 'jungle gyms' and people as apes.

# Fairness in AI



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

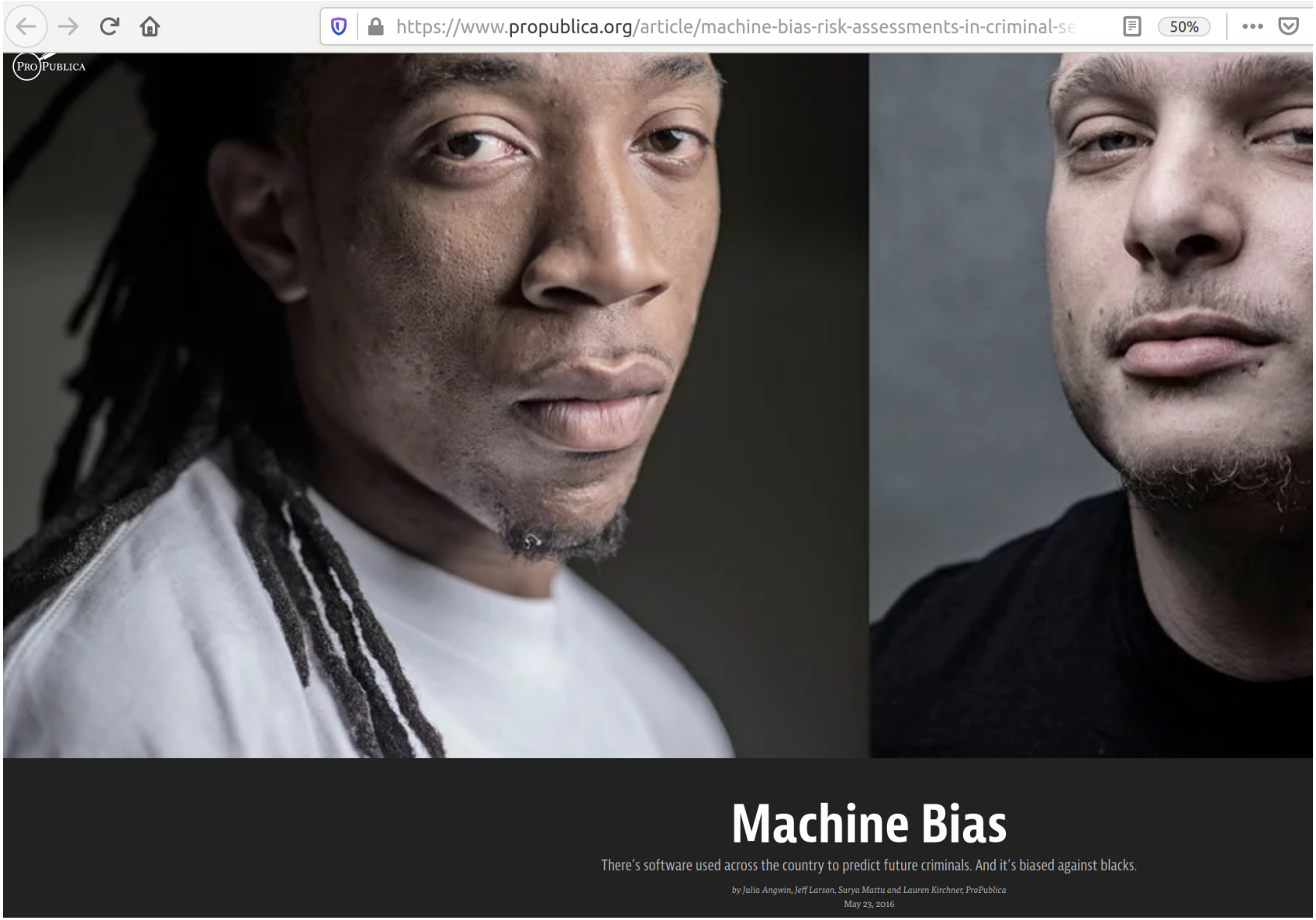
Article Figures & Data Info & Metrics eLetters PDF

You are currently viewing the abstract. View Full Text

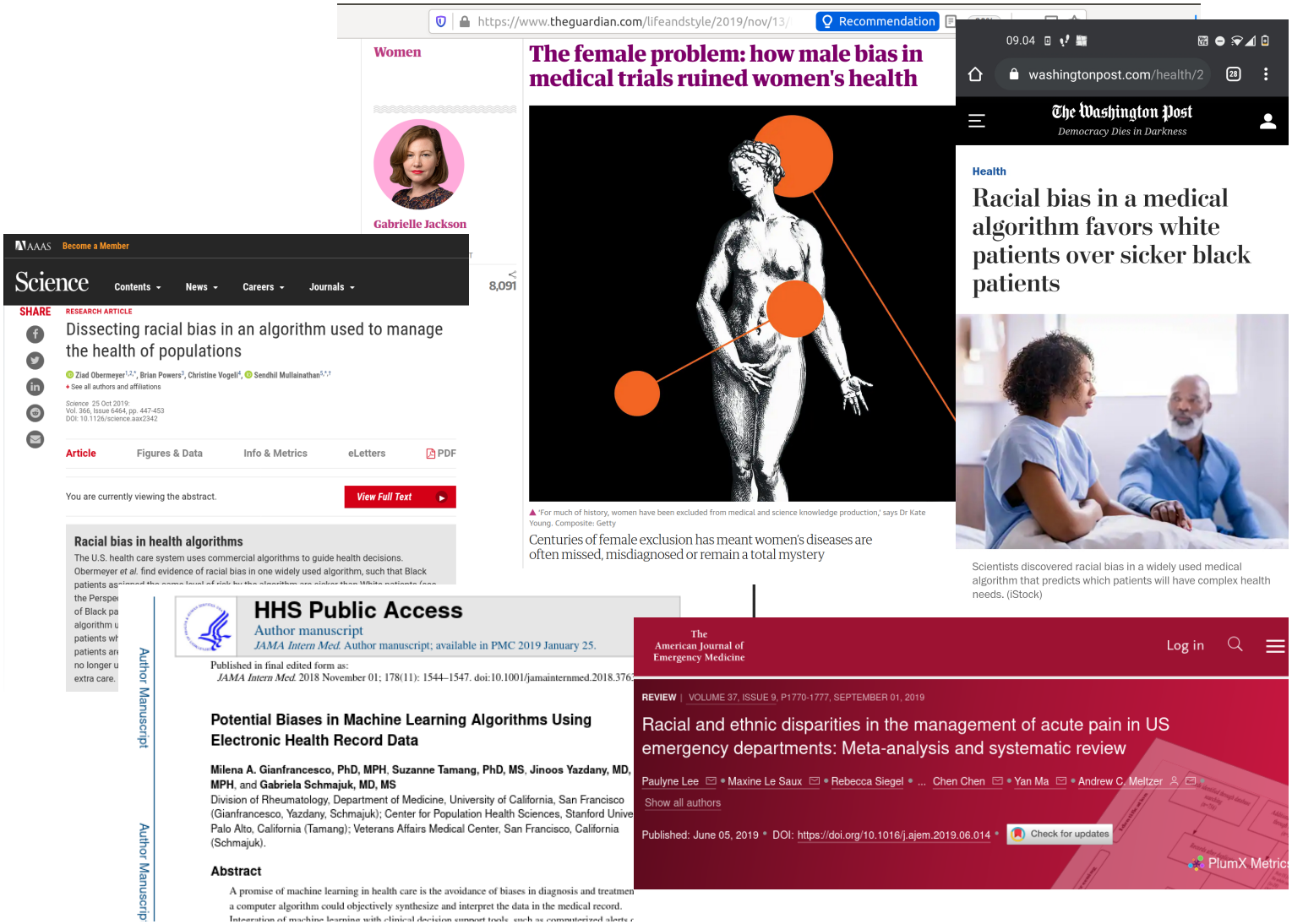
### Racial bias in health algorithms

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.

# Fairness in AI



# Fairness in (healthcare) AI



# Case 1. A simulated example

Imagine using predicted depression risk scores for prioritizing resources such as referral to a psychologist

# (It's not exactly hypothetical)

Forbes

FORBES > INNOVATION > AI

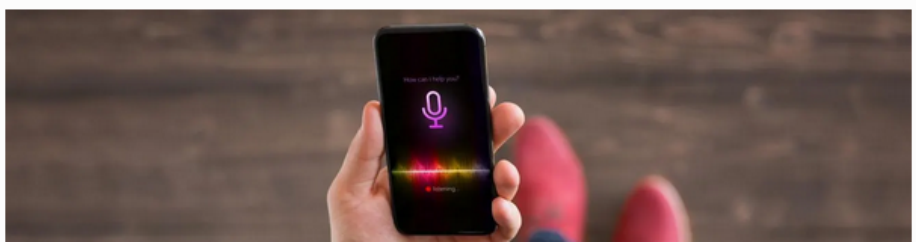
## AI Can Now Detect Depression From Your Voice, And It's Twice As Accurate As Human Practitioners

**Ganes Kesari** Contributor
*2X Founder & Chief Decision Scientist, TEDx Speaker, Adjunct Professor*

Follow

May 24, 2021, 06:50am EDT

This article is more than 2 years old.



# 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)

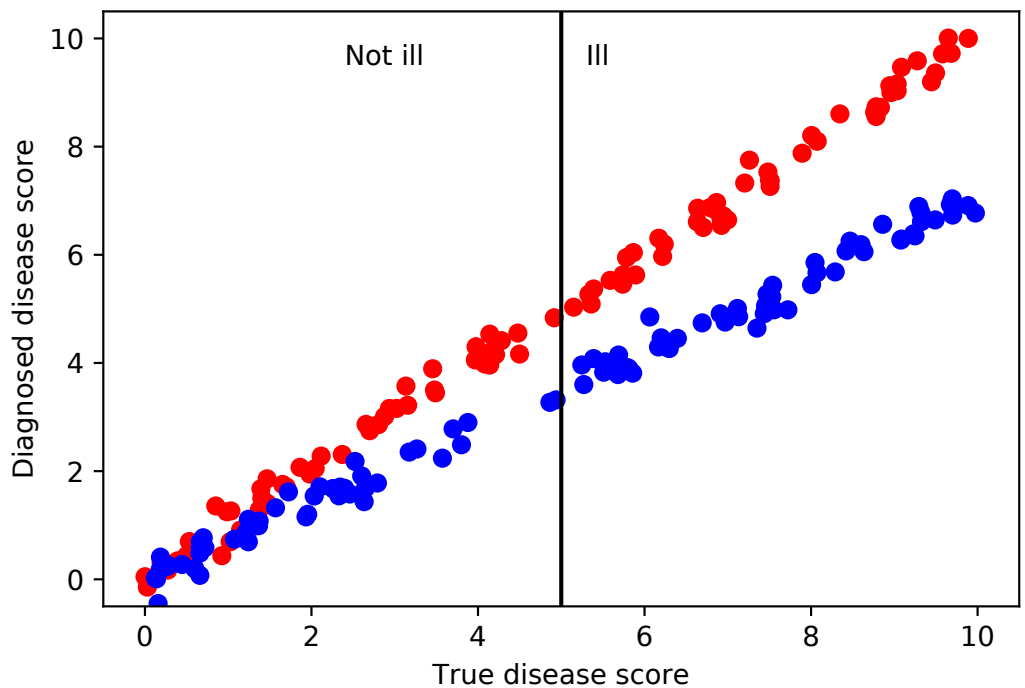
- If the diagnostic criteria are adapted to male symptoms, then the prevalence of depression among men increases (Martin et al, 2013)

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

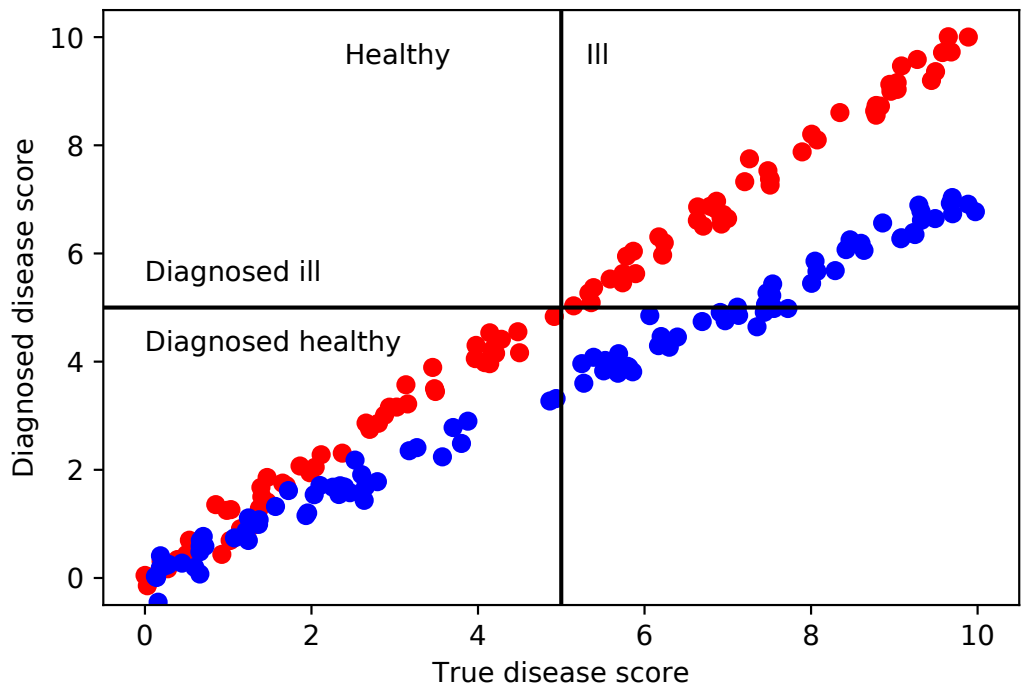
Imagine a disease model where

- 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

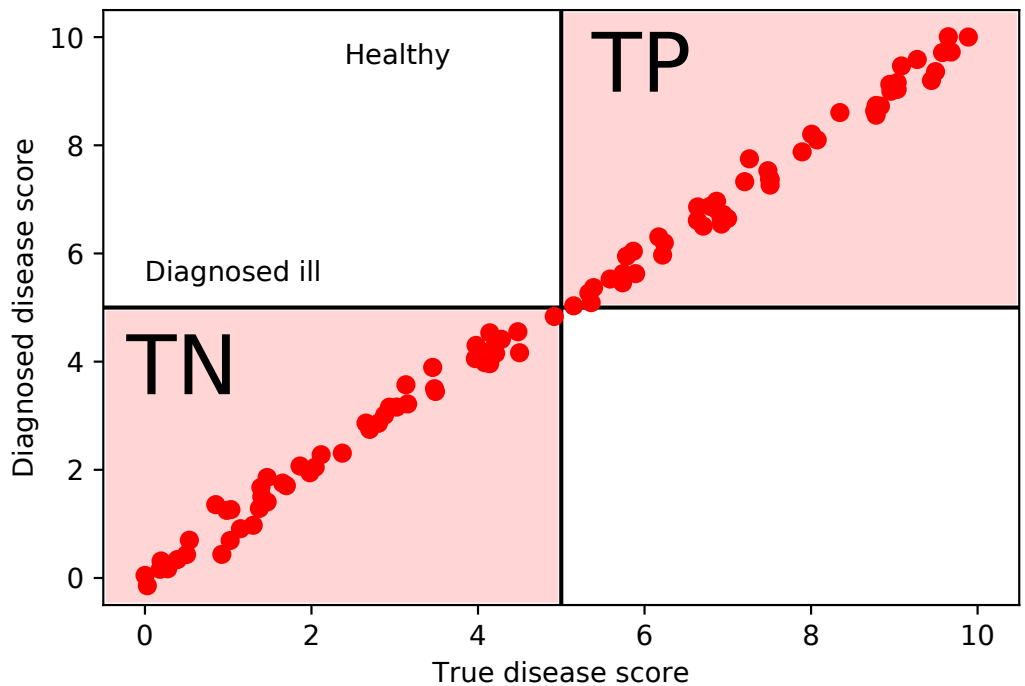
Setting a diagnostic threshold at diagnosed disease score = 5, we see that:



# Bias in algorithms: A toy illustration

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

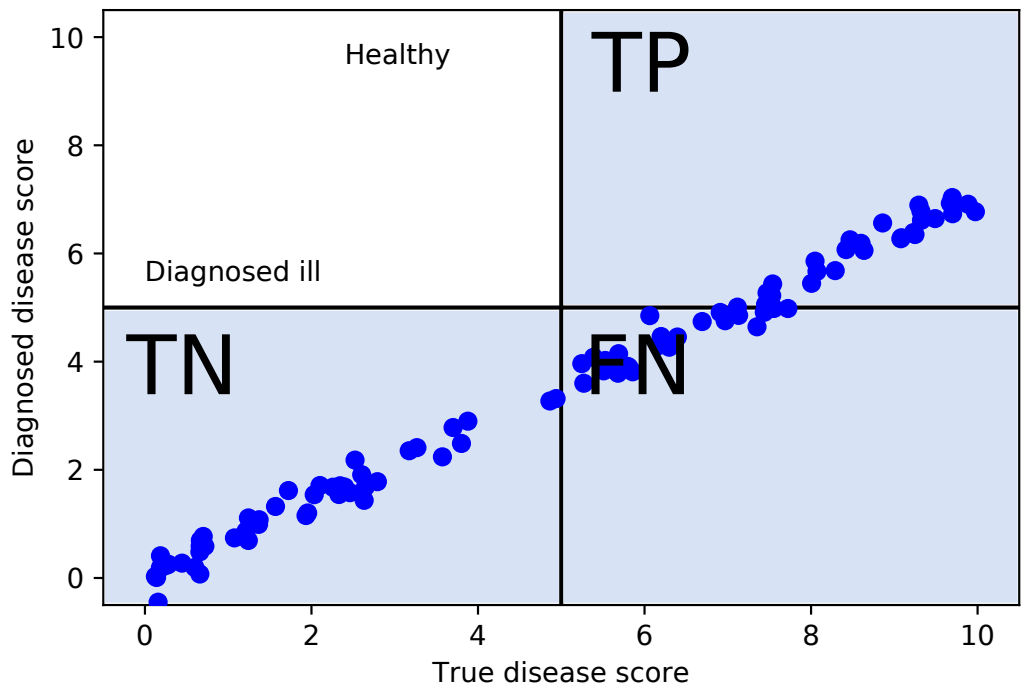
- For the red group, we have no false diagnoses



# Bias in algorithms: A toy illustration

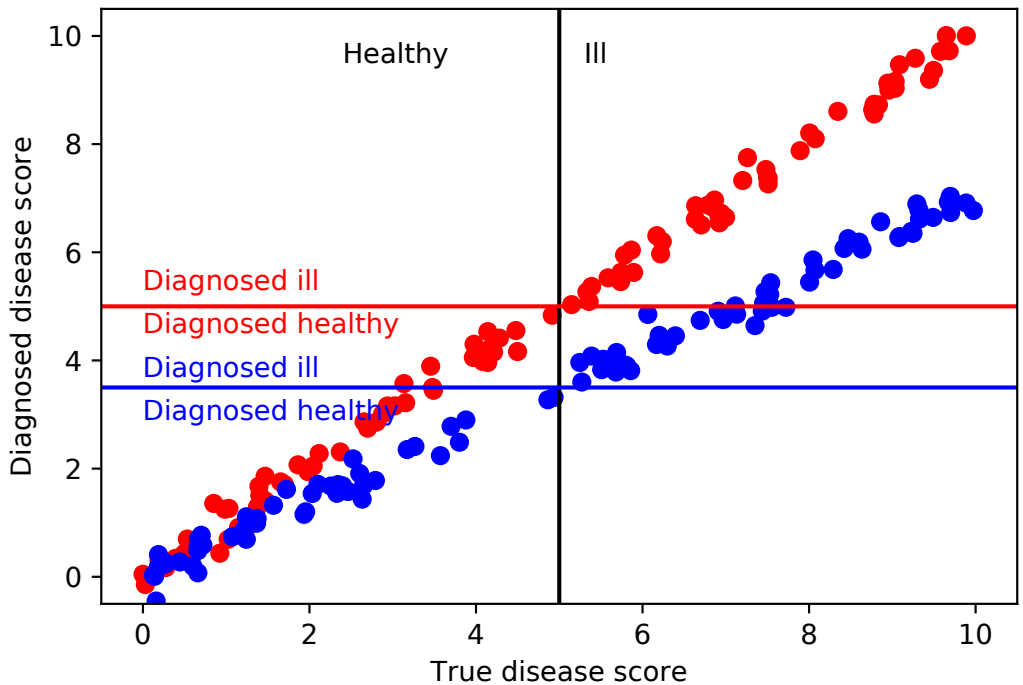
Setting a diagnostic threshold at diagnosed disease score = 5, we see that:

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



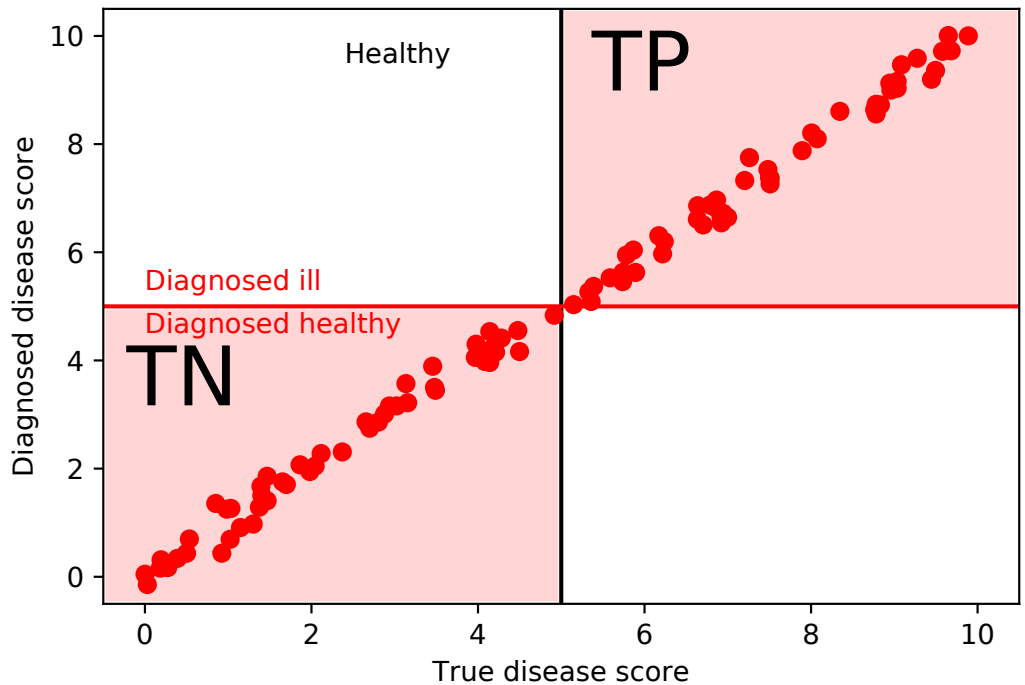
# Bias in algorithms: A toy illustration

**Solution:** Population-specific thresholds



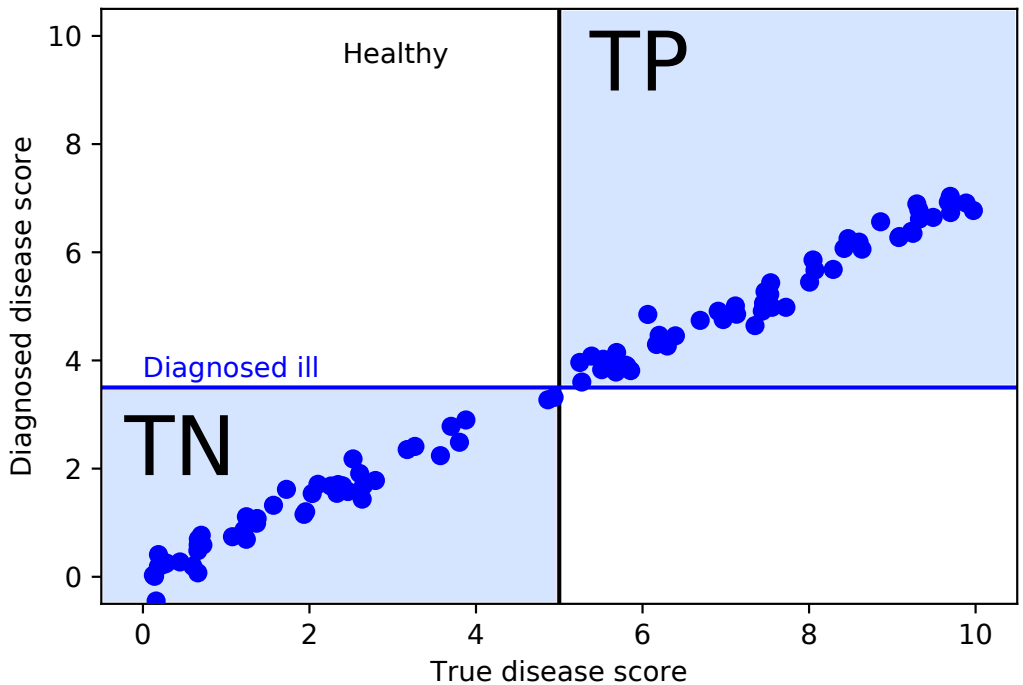
# Bias in algorithms: A toy illustration

**Solution:** Population-specific thresholds



# Bias in algorithms: A toy illustration

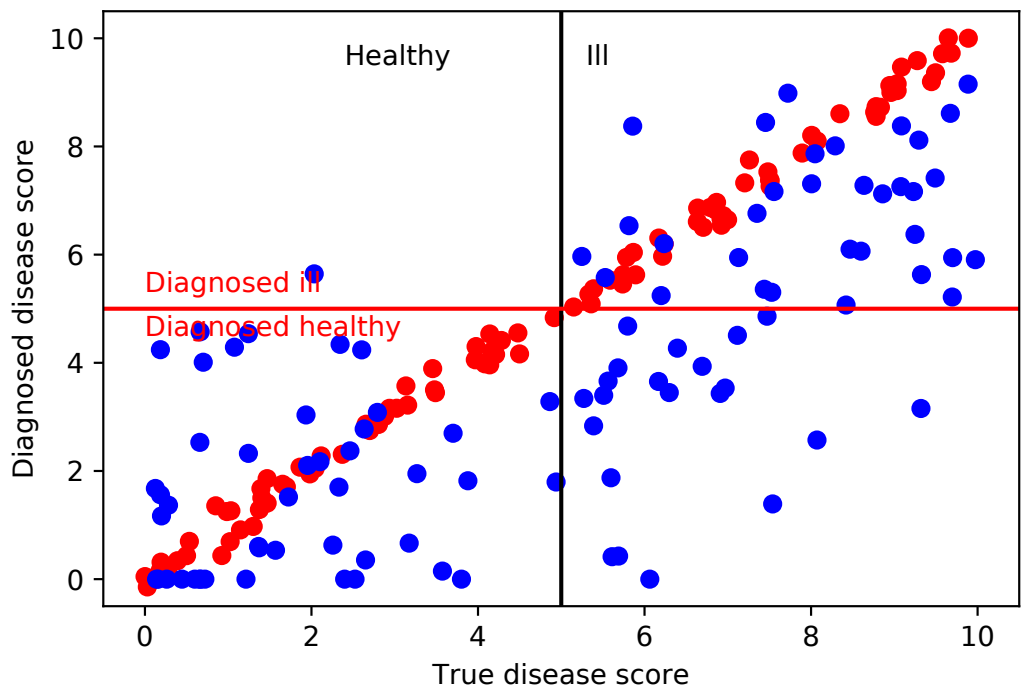
**Solution:** Population-specific thresholds



# 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)

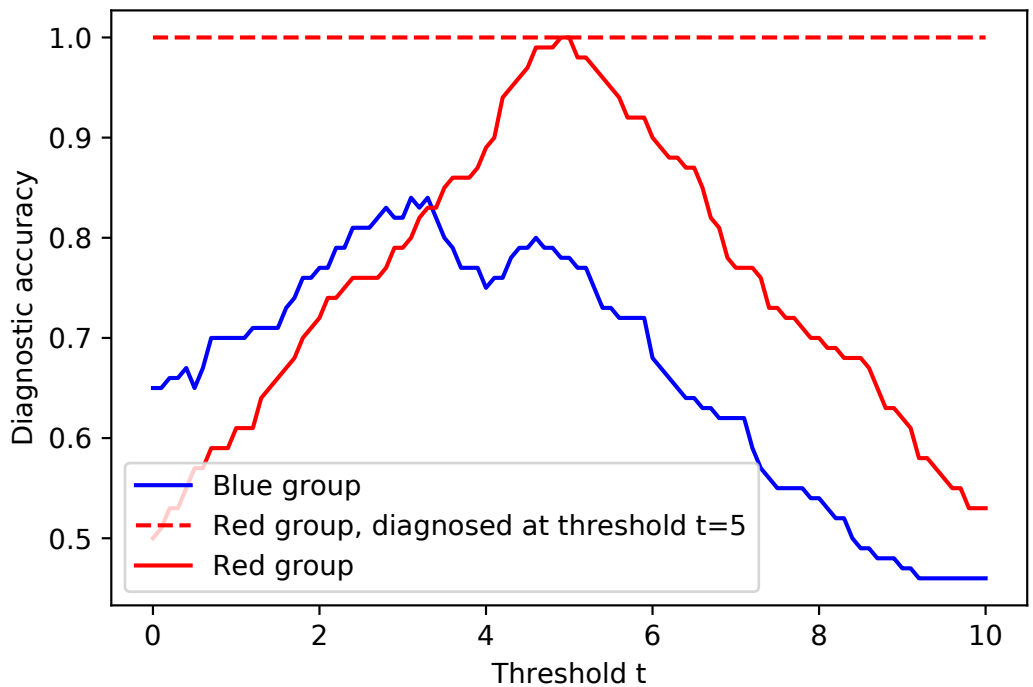
- Here, the score=5 threshold creates false positives and negatives in the blue group



# Bias in algorithms: A toy illustration

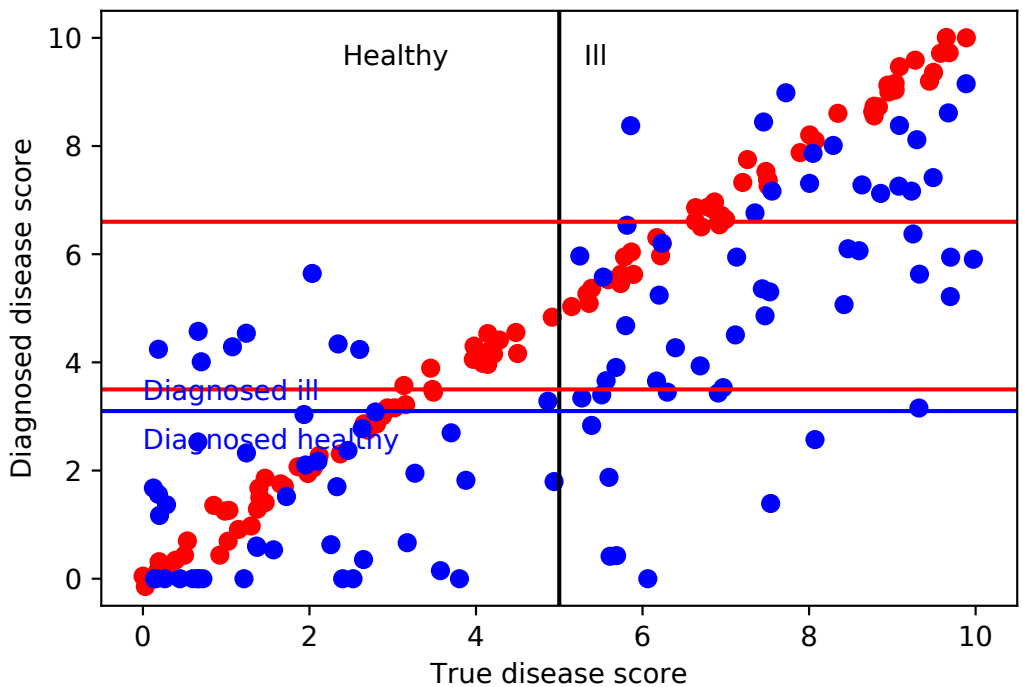
Below, see the group-wise diagnostic accuracy for the two different classes

- 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

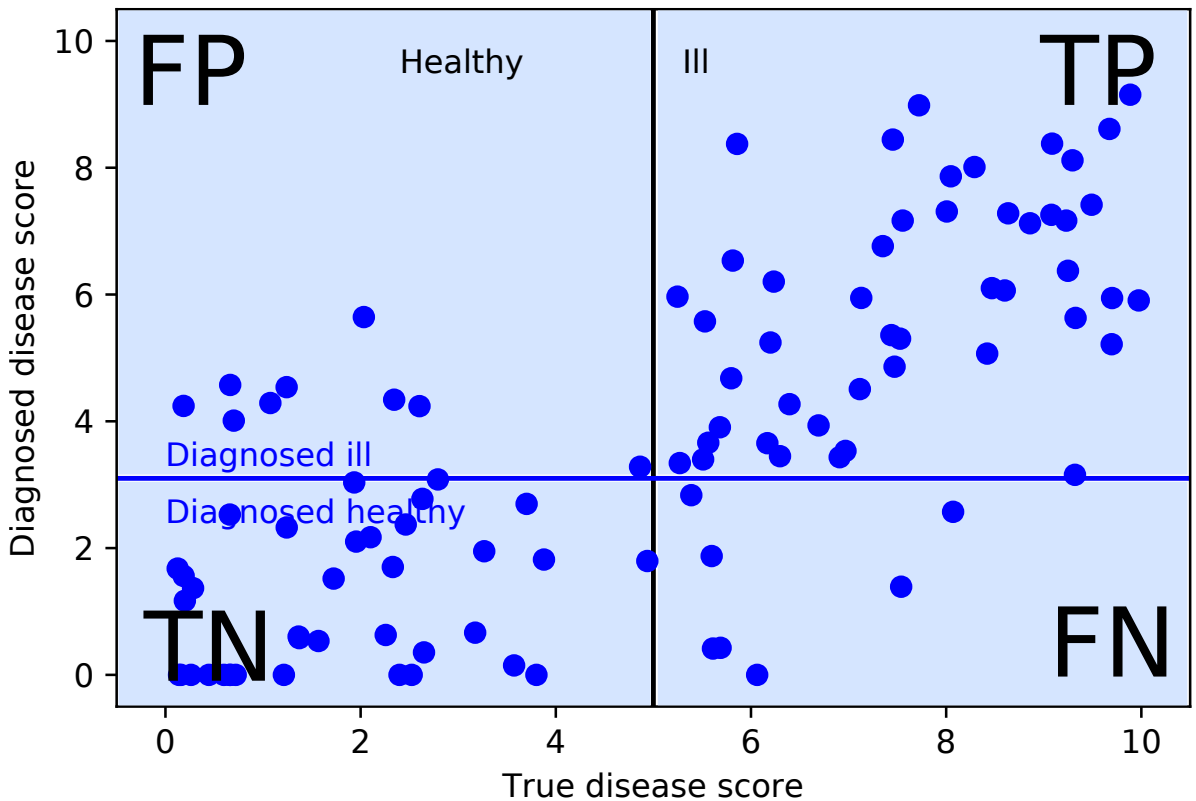
Let's see what those thresholds do:



# Bias in algorithms: A toy illustration

Let's see what those thresholds do:

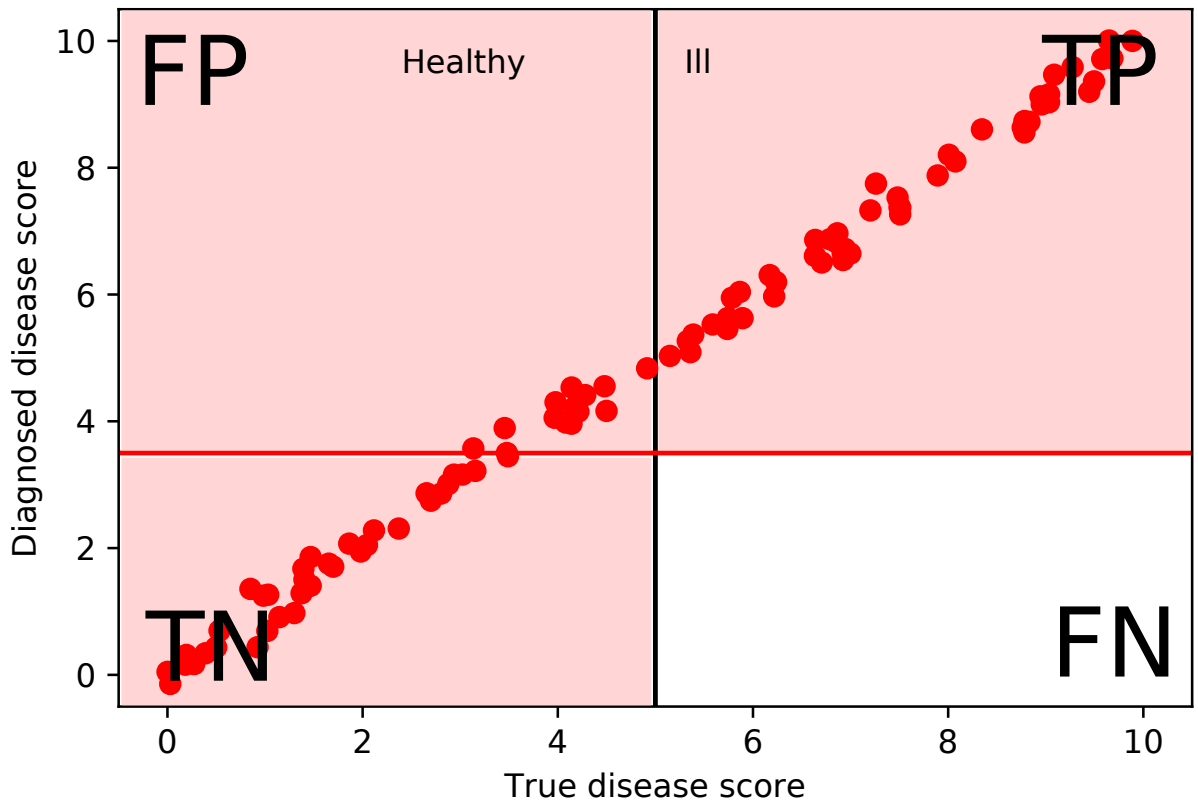
- Blue group has positive TP, TN, FP and FN



# Bias in algorithms: A toy illustration

Let's see what those thresholds do:

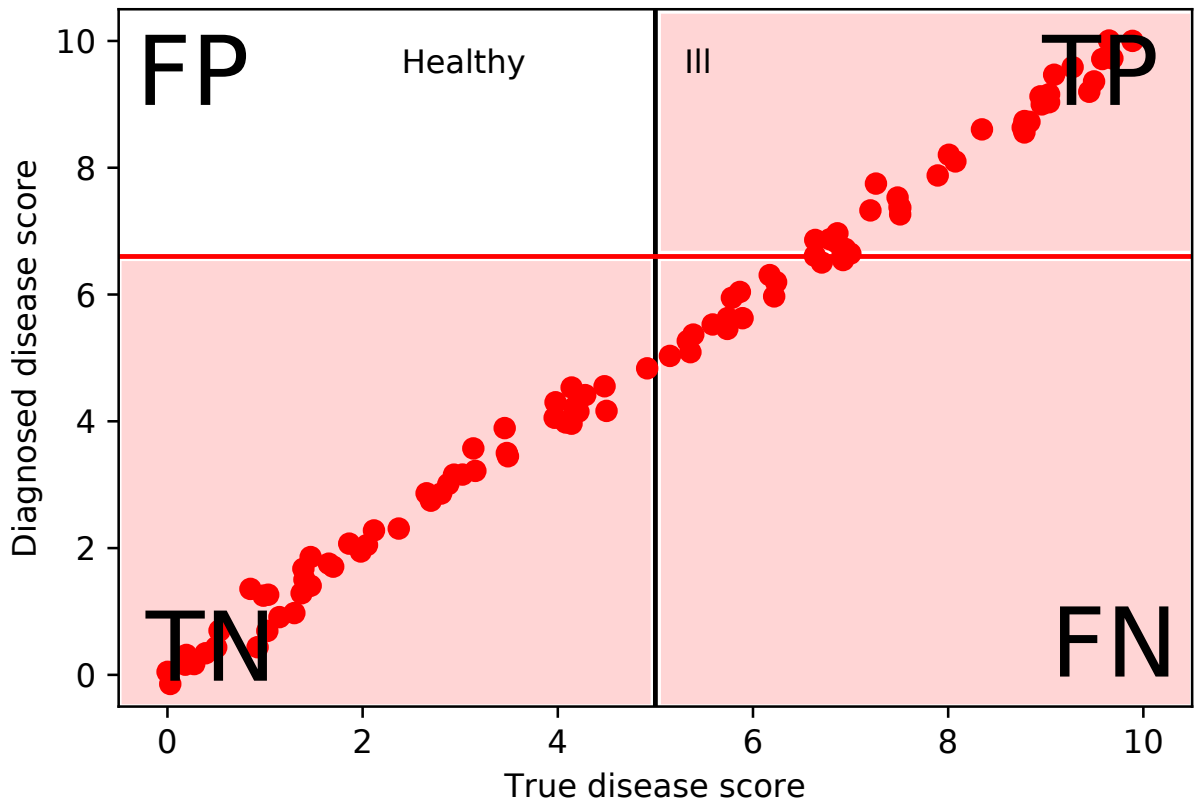
- Blue group has positive TP, TN, FP and FN
- Red group has positive TP, TN and FP, but no FN



# Bias in algorithms: A toy illustration

Let's see what those thresholds do:

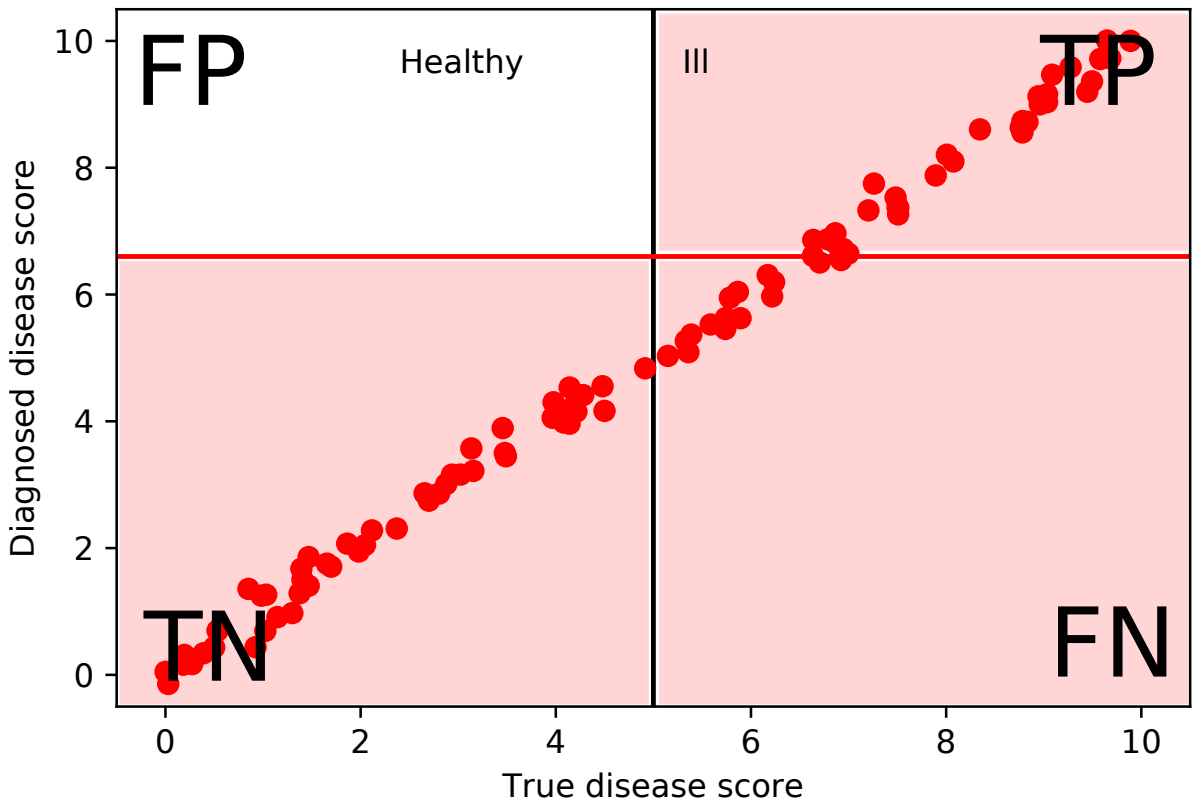
- Blue group has positive TP, TN, FP and FN
- Red group has positive TP, TN and FN, but no FP



# 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 (and a couple of tangents)

# Bias in image-based diagnosis of thoracic disorders

Gender bias in computer aided diagnosis based on chest X-ray

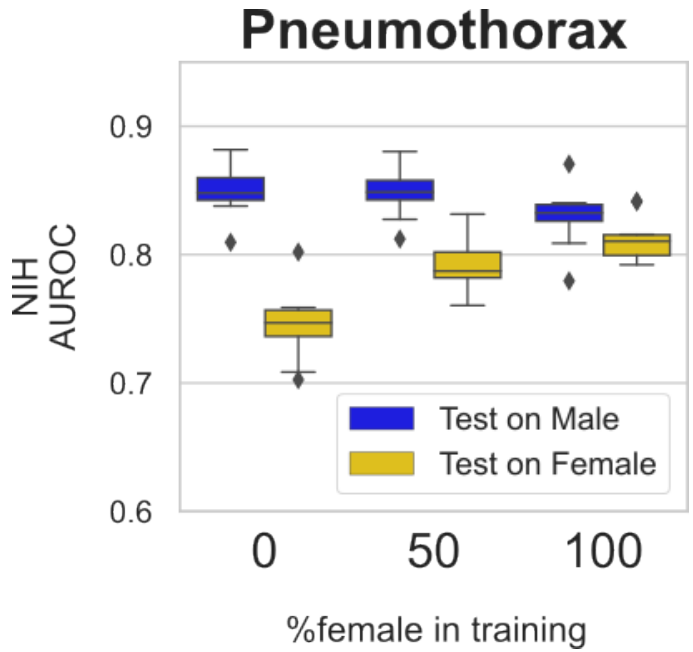



Figure: Diagnostic accuracy for Pneumothorax for men (left) and women (right) as a function of % women in the training set. Low representation gives low performance.

---

Figure by Weng et al, FAIMI'23; reproduced from Larrazabal et al, PNAS'20.

# Bias in image-based diagnosis of thoracic disorders

But it's not always that simple:

A replicated experiment diagnosing Alzheimer from brain MRI does not show the expected analogous effect

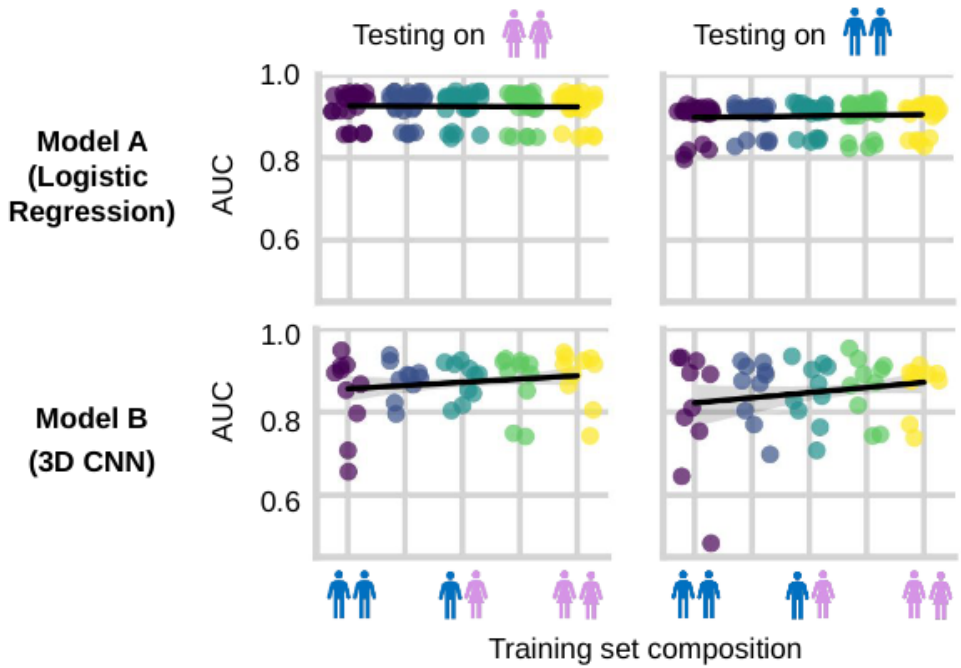


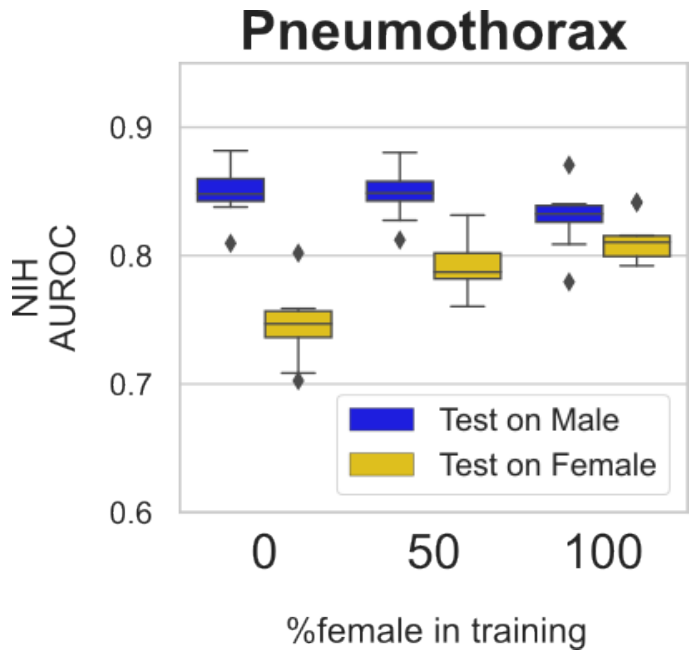

Figure: **Left:** Diagnostic accuracy for Alzheimer's disease for men (left) and women (right) as a function of % women in the training set.

---

Petersen et al, MICCAI 2022

# Bias in image-based diagnosis of thoracic disorders

Gender bias in computer aided diagnosis based on chest X-ray




**Note** that the best models for women are better for men!!

---

Figure by Weng et al, FAIMI'23; reproduced from Larrazabal et al, PNAS'20.

# Bias in image-based diagnosis of thoracic disorders

I thought I knew the cause of the bias!

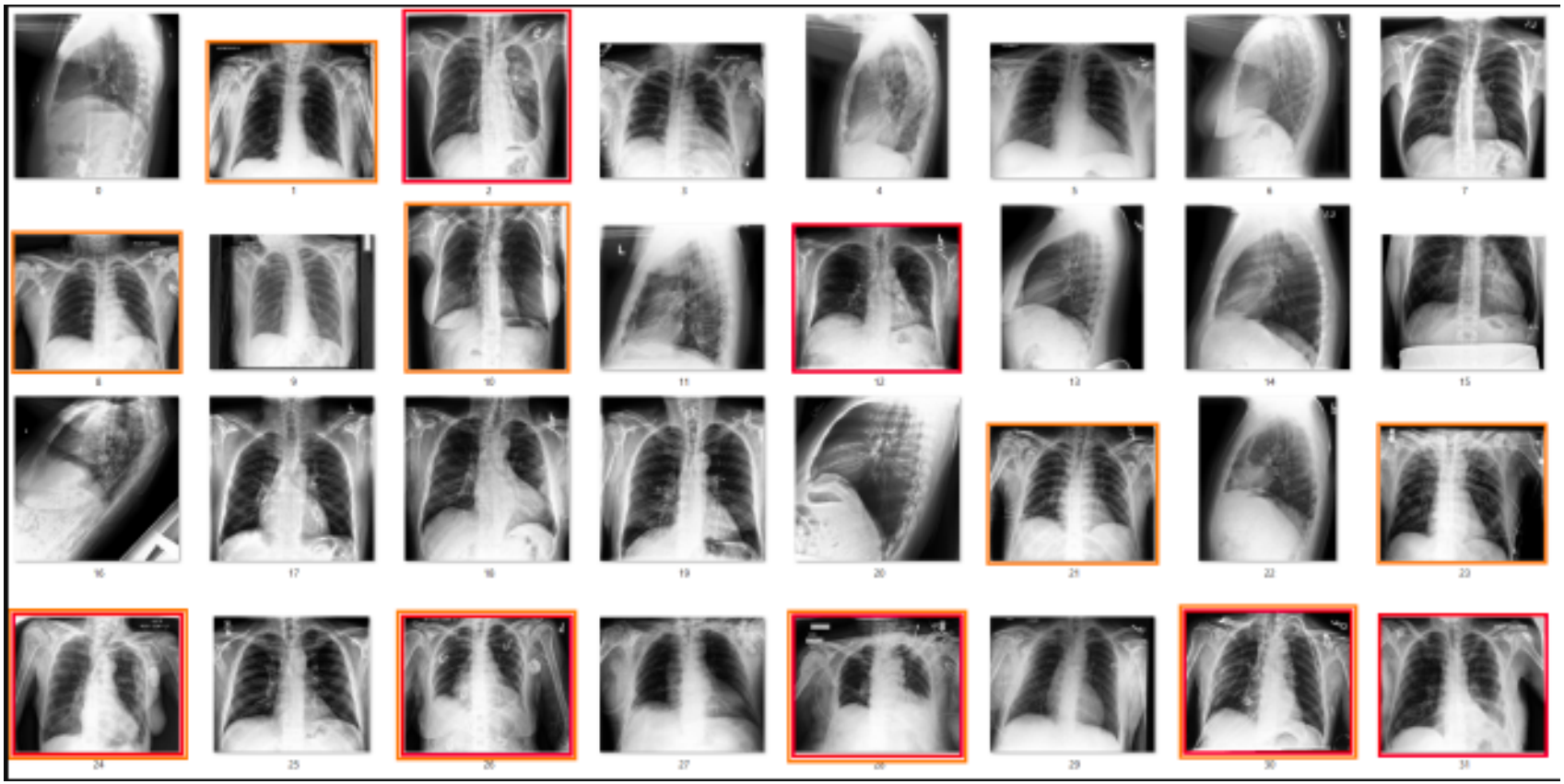


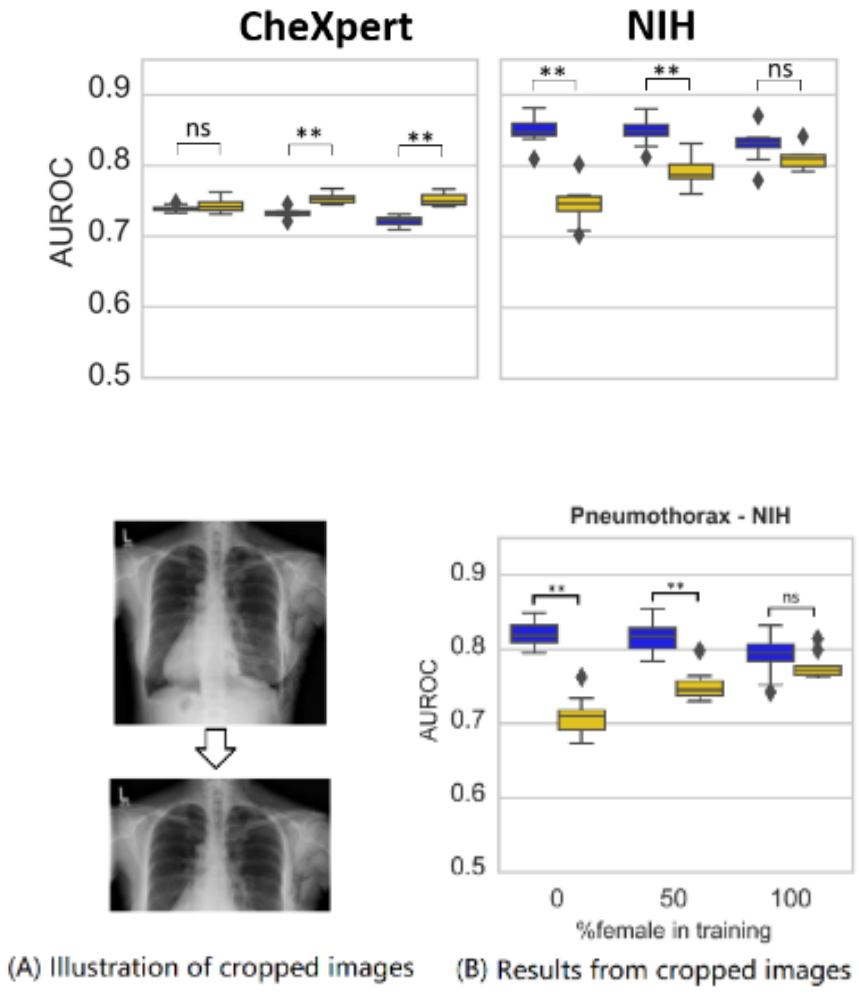
Figure: Selected chest X-rays

---

https://laurenoakdenrayner.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/

# Bias in image-based diagnosis of thoracic disorders

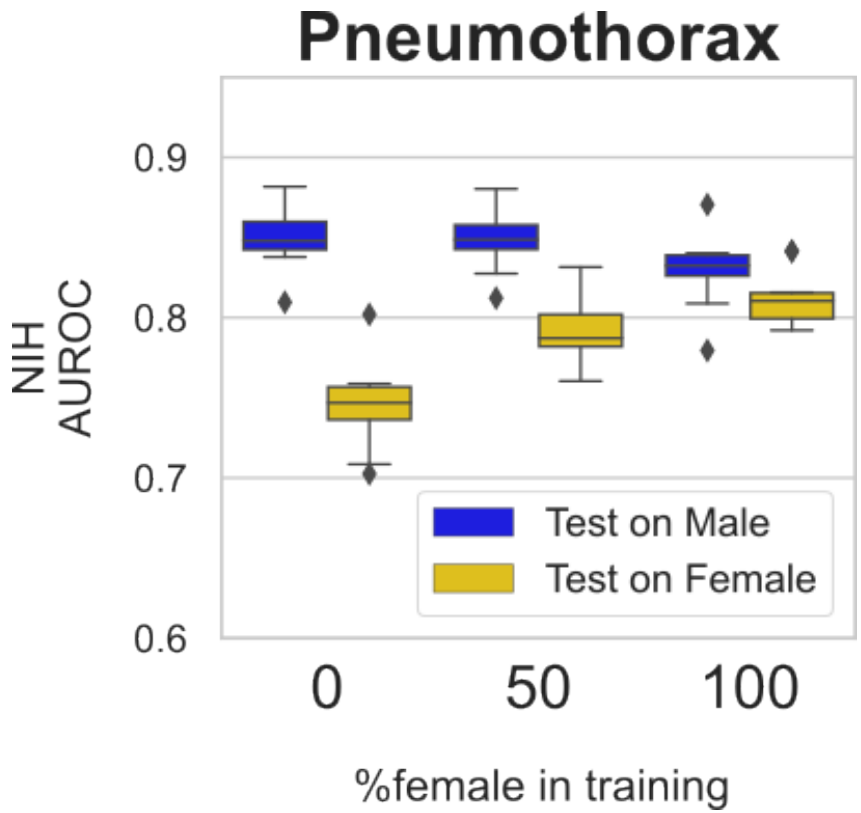
But it wasn't that simple...



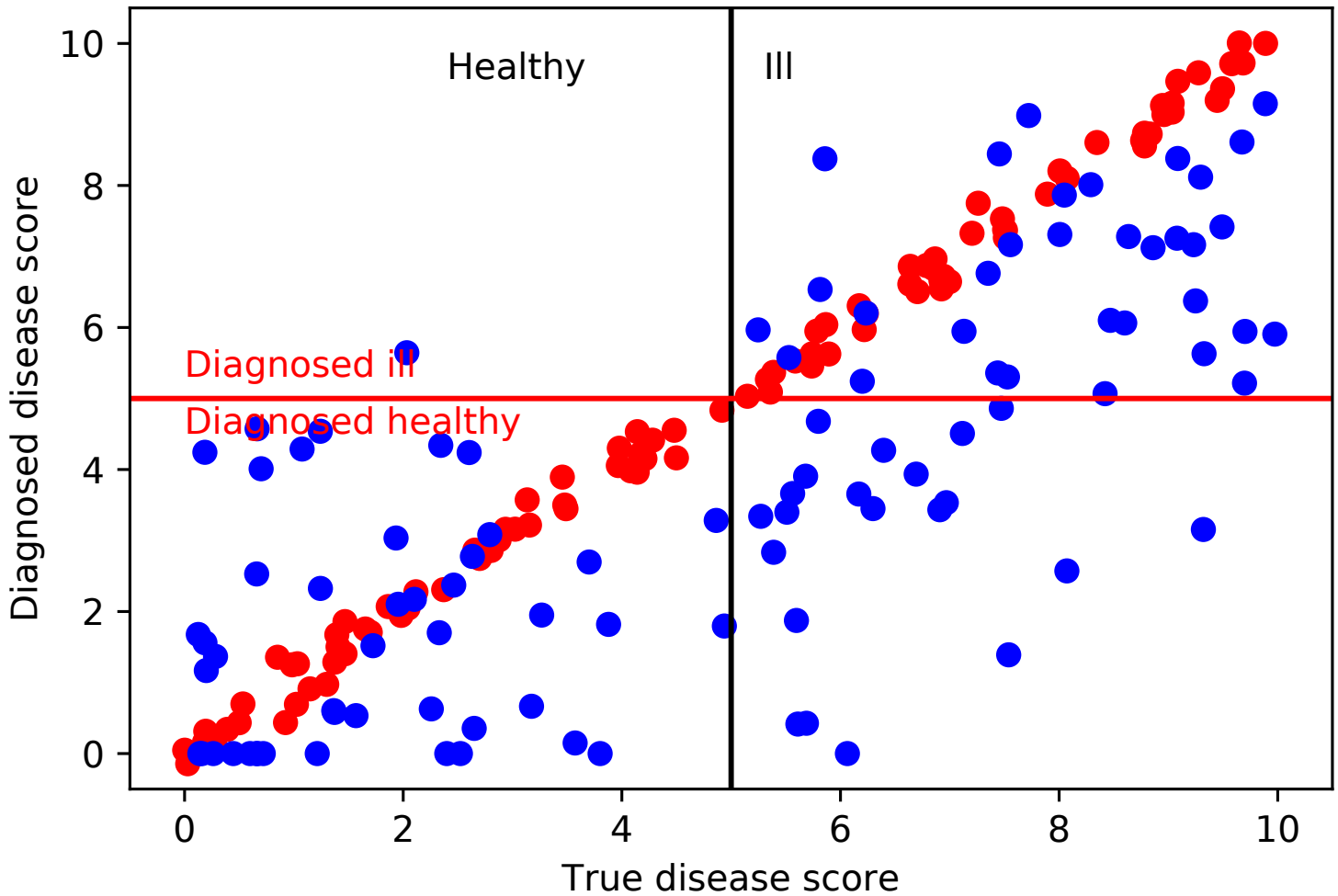

(A) Illustration of cropped images (B) Results from cropped images

Weng et al: Are Sex-based Physiological Differences the Cause of Gender Bias for Chest X-ray Diagnosis? FAIMI'23

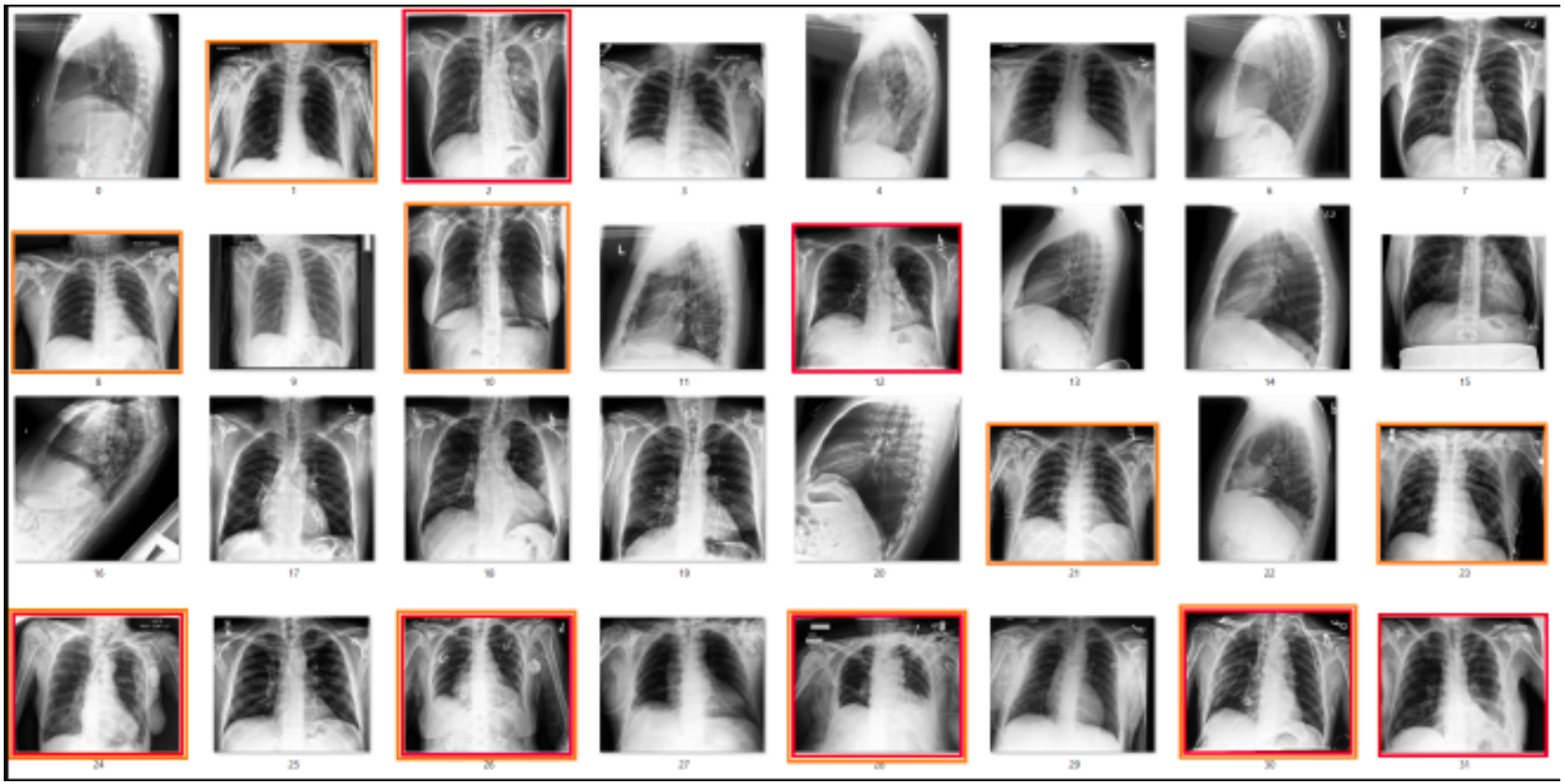
# Why would a model be unfair? Sampling bias



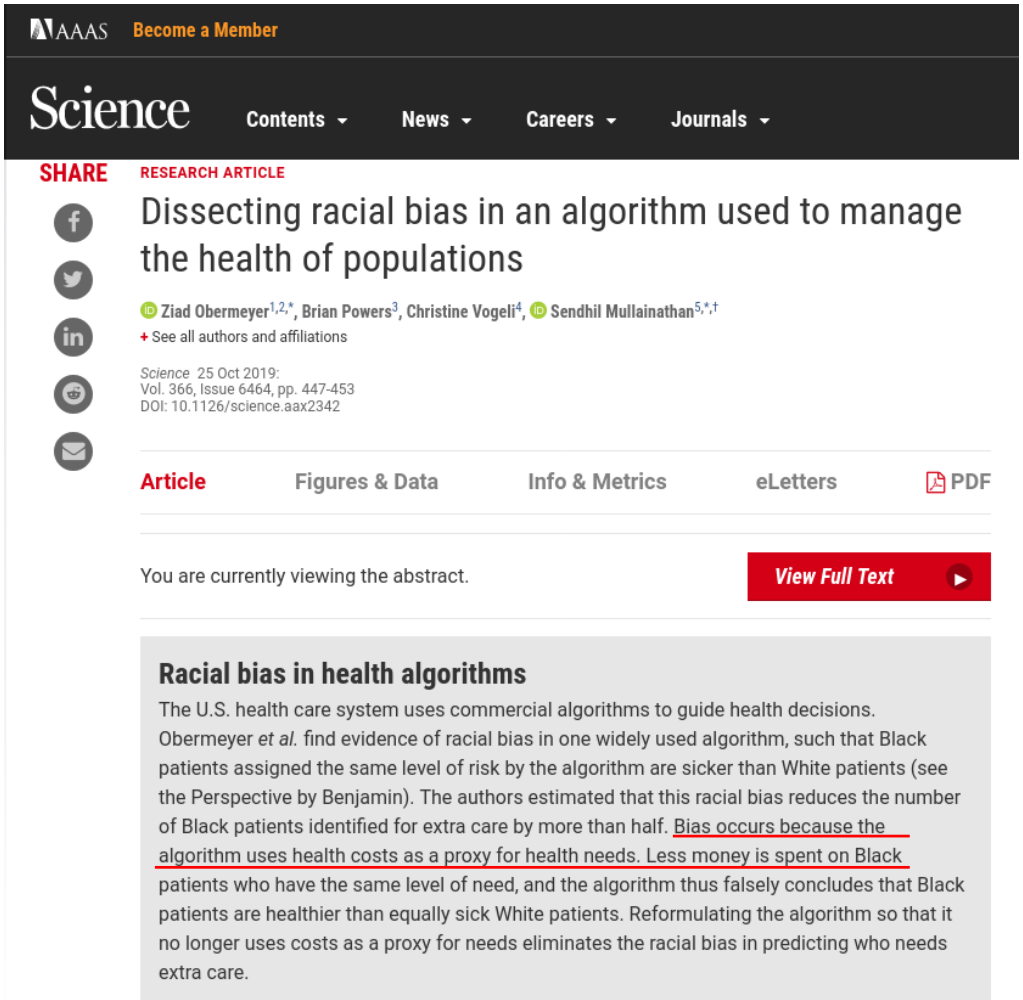
# Why would a model be unfair? Differences in task difficulty (e.g. diagnostic features fit some groups better than others)



# Why would a model be unfair? Differences in task difficulty (e.g. data quality difference between groups)



# Why would a model be unfair? Systematic underdiagnosis for certain groups



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

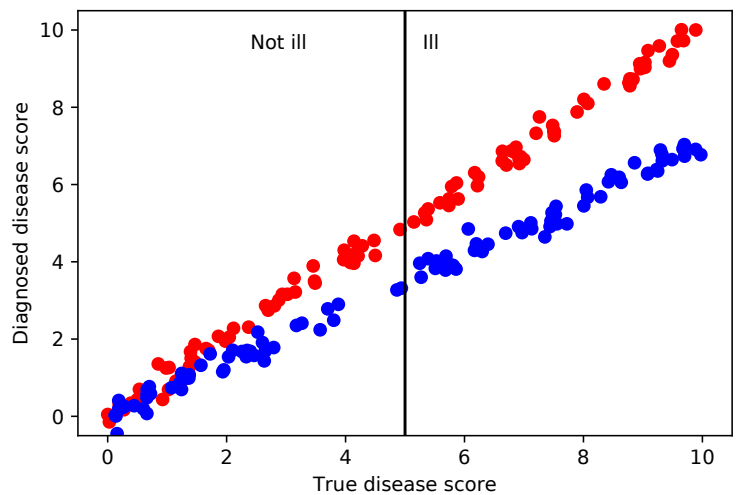
**Racial bias in health algorithms**

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.

Here, even *measuring* bias becomes difficult – we lack true diagnoses.

# 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.



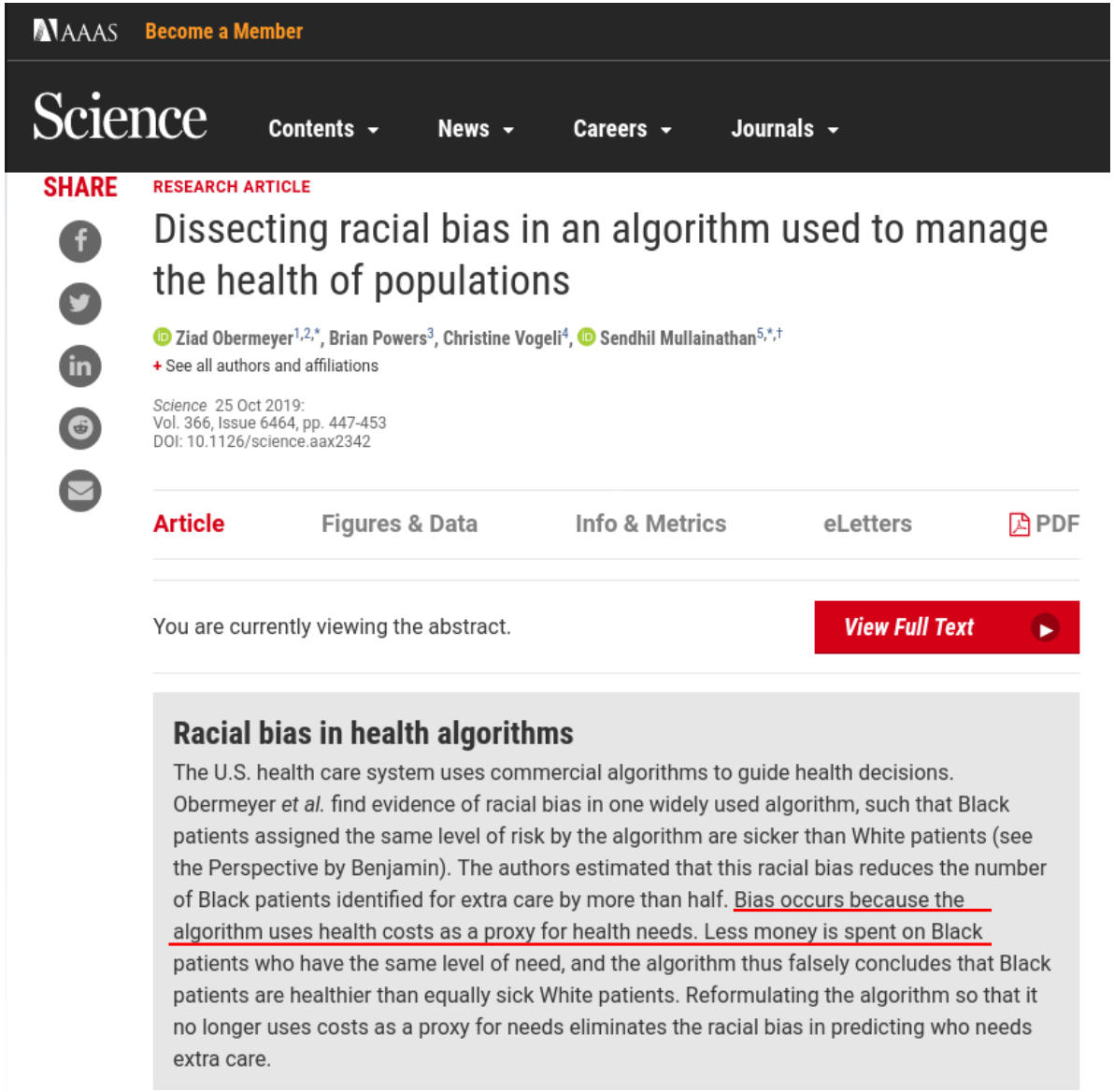
# Quality of labels:
# Proxy variables for bias detection and better training?

COMPAS case[1]: 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.

[1] 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

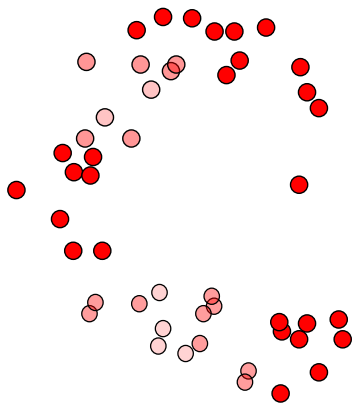
You are currently viewing the abstract. View Full Text

### Racial bias in health algorithms

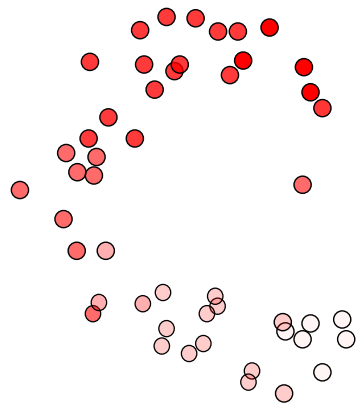
The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The 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?

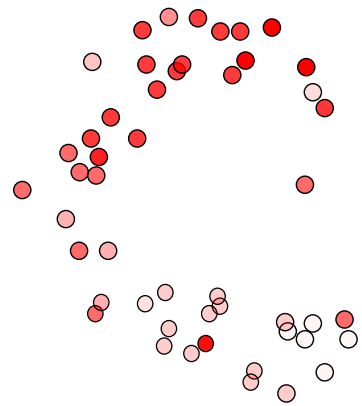
**Open problem:** Proxy variables for diagnosis?



Observed diagnosis



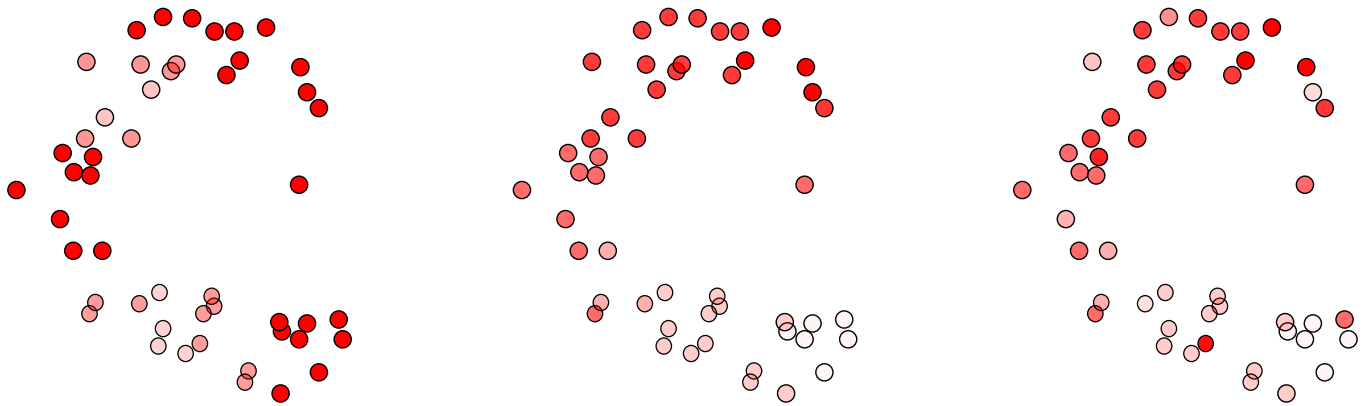
True diagnosis



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?

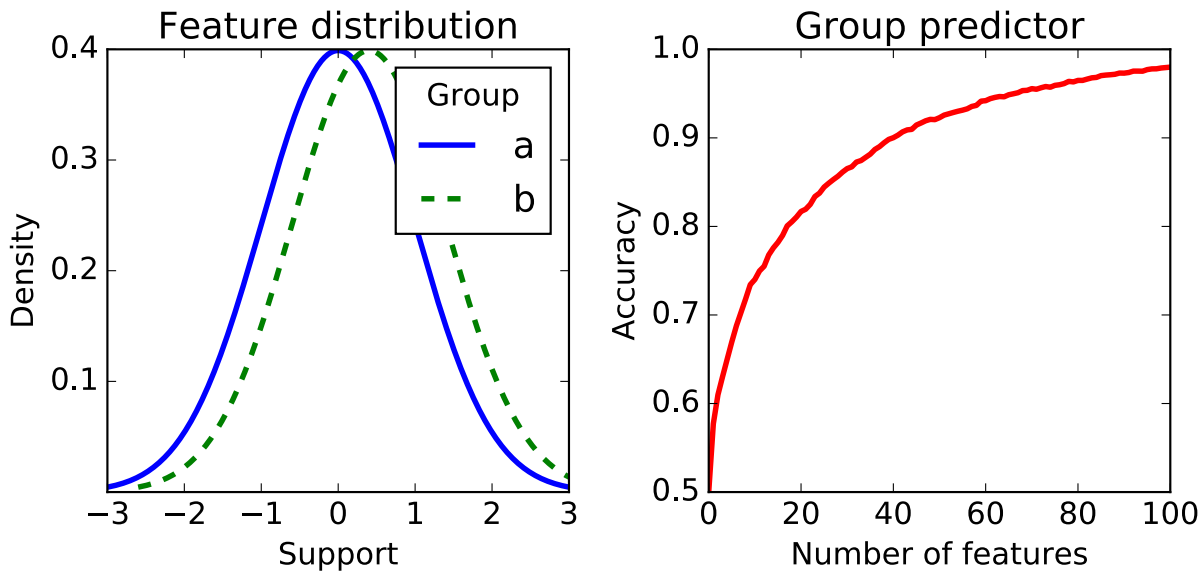
# Fairness criteria

# Classification: Fixing notation

- Input features (covariates) $X$, targets $Y$ – we consider binary classification where $Y \in \{0, 1\}$
- Classifier (e.g. neural network) $f(X) \approx Y$
- Models generally predict a probabilistic *score* $f(X) = R$, which can be thresholded to obtain a binary $\hat{Y}$

# Sensitive attributes - no fairness through unawareness

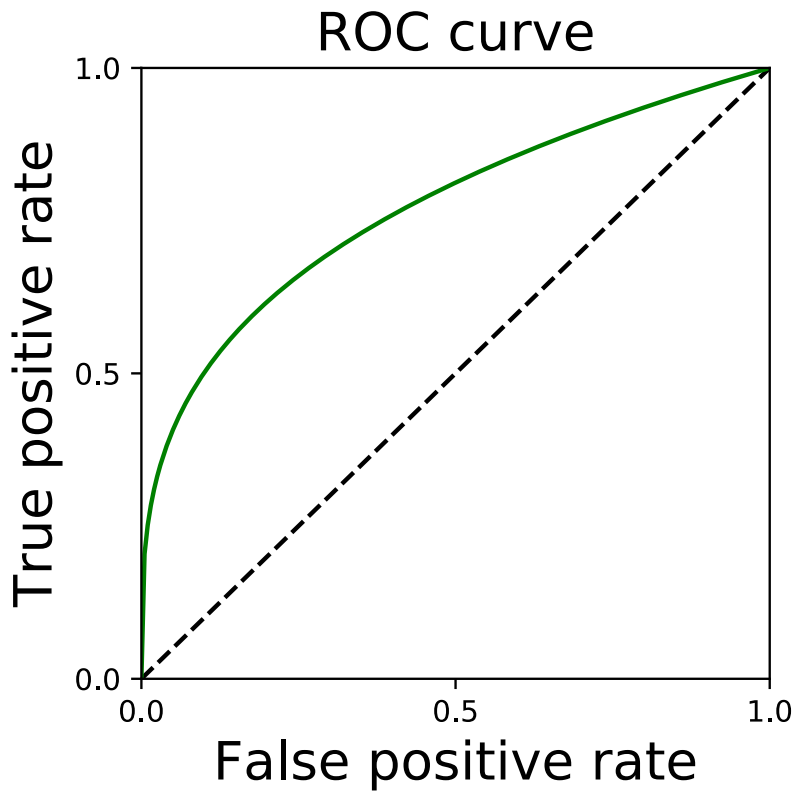
- Denote the sensitive group by $A$
- No fairness through unawareness! Deleting the sensitive attribute will not solve the problem
- Below, consider features that are distributed slightly differently with respect to group




- To formalize fairness of classifiers, we study how the predictions $\hat{Y}$ and the model scores $R$ that generated them depend on the sensitive group $A$.

# A tool for handling scores and their classifications: ROC curves

- The probabilistic *score* $R$ is typically thresholded to obtain a binary $\hat{Y}$
- Different thresholds give different classifiers with different TPR/FPR
- The entire spectrum of thresholds gives rise to the ROC curve:



# Classification metrics

- Input features (covariates) $X$, targets $Y$ – we consider binary classification where $Y \in \{0, 1\}$
- Classifier (e.g. neural network) $f(X) = R$ thresholded to give $\hat{Y}$ predicting $Y$

A range of different classification metrics can be used to define fairness:

Common classification criteria

| Event | Condition | Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$) |
|---|---|---|
| $\hat{Y} = 1$ | $Y = 1$ | True positive rate, recall |
| $\hat{Y} = 0$ | $Y = 1$ | False negative rate |
| $\hat{Y} = 1$ | $Y = 0$ | False positive rate |
| $\hat{Y} = 0$ | $Y = 0$ | True negative rate |

Additional classification criteria

| Event | Condition | Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$) |
|---|---|---|
| $Y = 1$ | $\hat{Y} = 1$ | Positive predictive value, precision |
| $Y = 0$ | $\hat{Y} = 0$ | Negative predictive value |

# Different notions of algorithmic fairness

https://fairmlbook.org/

Three different categories of "fairness"

- Independence
- Separation (Equalized odds)
- Sufficiency

These correspond to equalizing across groups

- Acceptance rate $\mathbb{P}\{\hat{Y} = 1\}$ of the classifier
- The error rates $\mathbb{P}\{\hat{Y} = 0 | Y = 1\}$ and $\mathbb{P}\{\hat{Y} = 1 | Y = 0\}$ of the classifier
- Outcome frequency $\mathbb{P}\{Y = 1 | R = r\}$ of a score $R$

# Independence

- **Definition:**
  The model $R$ satisfies *independence* if $R \perp A$.
- Applied to binary predictions $\hat{Y}$, this is equivalent to

$$\mathbb{P}\{\hat{Y} = 1 | A = a\} = \mathbb{P}\{\hat{Y} = 1 | A = b\}$$

  for all groups $a, b$
- That is, equalizing the acceptance rate $\mathbb{P}\{\hat{Y} = 1\}$ of the classifier across groups

## Question

*In which ways is this a good criterion? In which ways is it bad?*
*E.g. consider scenarios:*

- *Hiring female and male software developer*
- *Diagnosing cancer*

# Separation (Equalized Odds)

- **Definition:**
  The model $R$ satisfies *separation* if $R \perp A | Y$.
- **Intuition:** The target $Y$ quantifies underlying *merit*
- Applied to binary predictions $\hat{Y}$, this is becomes:

$$\mathbb{P}\{\hat{Y} = 1 | Y = 1, A = a\} = \mathbb{P}\{\hat{Y} = 1 | Y = 1, A = b\}$$

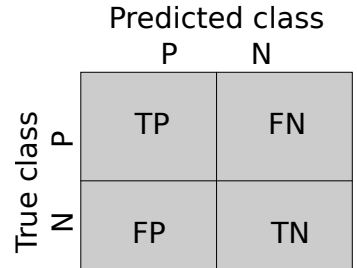
$$\mathbb{P}\{\hat{Y} = 1 | Y = 0, A = a\} = \mathbb{P}\{\hat{Y} = 1 | Y = 0, A = b\}$$

- Equalizing *true positive rate* and *false positive rate*
- Equivalent: Separation equalizes the error rates $\mathbb{P}\{\hat{Y} = 0 | Y = 1\}$ and $\mathbb{P}\{\hat{Y} = 1 | Y = 0\}$ of the classifier across groups

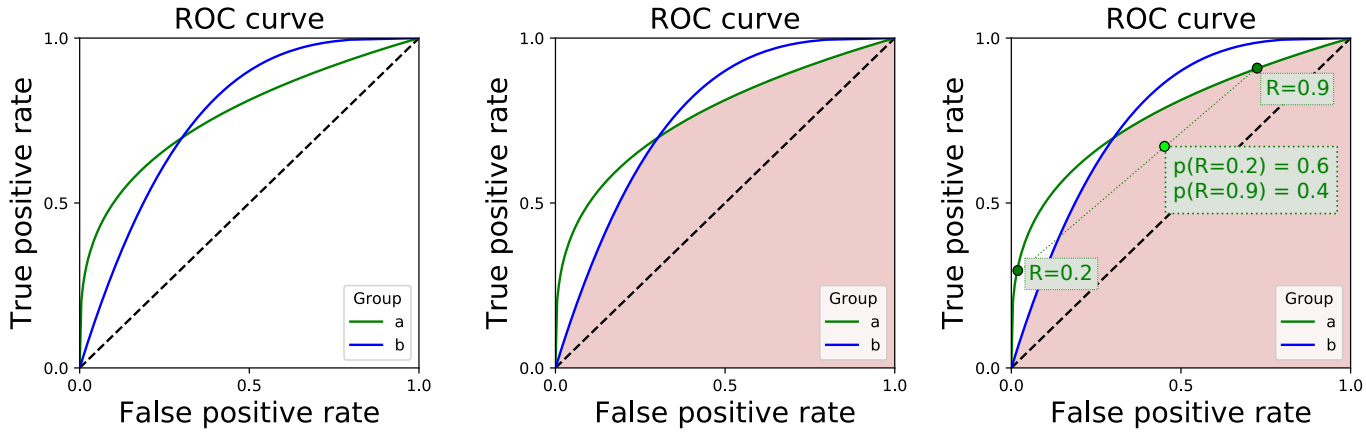
## Question

*In which ways is this a good or bad criterion? E.g.:*

- *Hiring female and male software developers*
- *Diagnosing heart attacks*
- *Diagnosing cancer*



# Visualizing and Obtaining Separation




Note that we can obtain separation via *randomization*, visualized on the ROC curve:

- A point on the green ROC curve corresponds to a classifier obtained by applying a given threshold to the model $R$.
- We can linearly interpolate between two points (classifiers) by sampling with different probabilities from the corresponding thresholds.
- **Relaxation:** *Equal opportunity* – only conditioning on positive outcome

$$\mathbb{P}\{\hat{Y} = 1 | Y = 1, A = a\} = \mathbb{P}\{\hat{Y} = 1 | Y = 1, A = b\}$$

# Sufficiency

- The model $R$ satisfies *sufficiency* if $Y \perp A | R$
- That is, for all groups $a, b$ and all $r$, we have

$$\mathbb{P}\{Y = 1 | R = r, A = a\} = \mathbb{P}\{Y = 1 | R = r, A = b\}$$

- This corresponds to equalizing the outcome frequency $\mathbb{P}\{Y = 1 | R = r\}$ of the score $R$ across groups $A$ for all scores $r \in R$.

## Question

*In which ways is this a good criterion? In which ways is it bad?*
*Again consider the scenarios:*

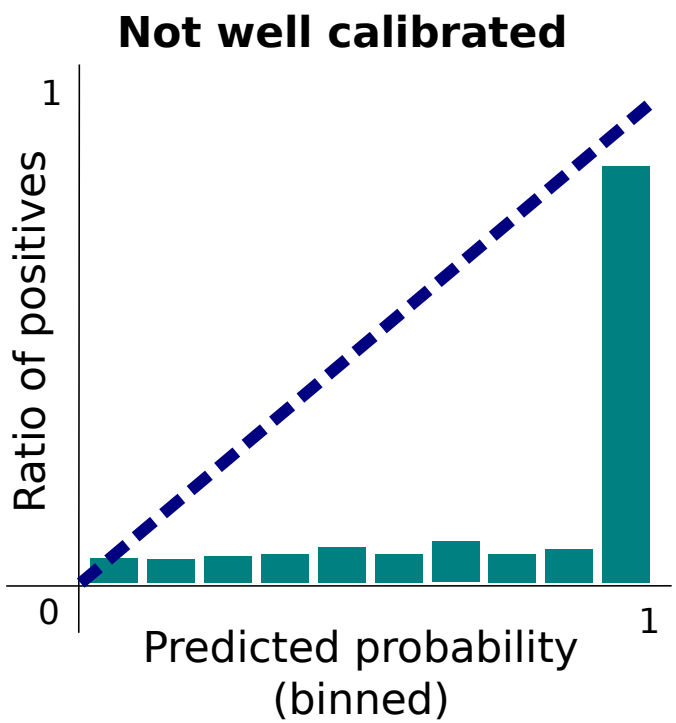
- *Hiring female and male software developers*
- *Diagnosing heart attacks*
- *Diagnosing cancer*

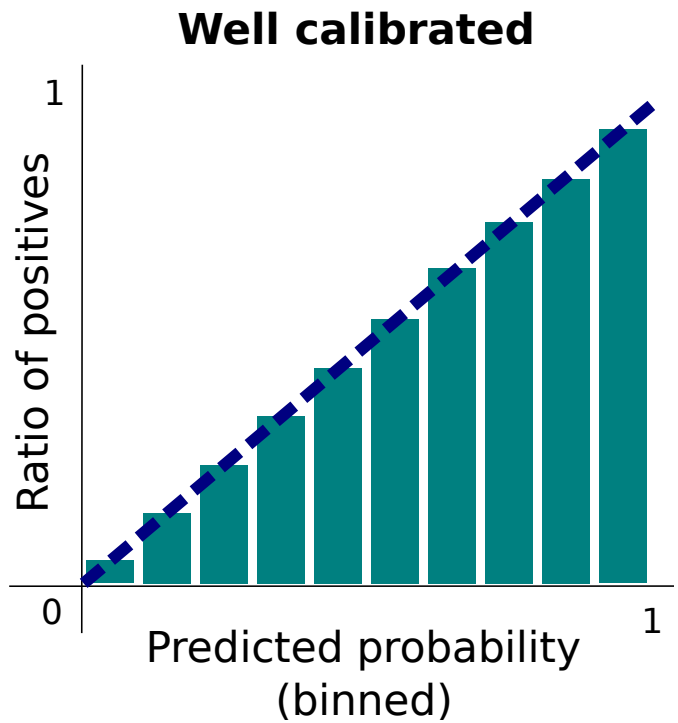
# How sufficiency relates to calibration

- Sufficiency is closely related to *calibration*.
- A model $R$ is *calibrated* with respect to the target $Y$ if for all scores $r \in [0, 1]$, we have

$$\mathbb{P}\{Y = 1 | R = r\} = r.$$

- Equivalently: on a population level, the predicted foreground probability equals the empirical one.



# How sufficiency relates to calibration

**Recall:**

- Sufficiency:
  $\mathbb{P}\{Y = 1 | R = r, A = a\} = \mathbb{P}\{Y = 1 | R = r, A = b\}$
- Calibration: $\mathbb{P}\{Y = 1 | R = r\} = r$

## Theorem

- *Calibration by group implies sufficiency*
- *If a model $R$ satisfies sufficiency, then there exists a calibrating function $l : [0, 1] \to [0, 1]$ so that $l(R)$ satisfies calibration by group.*

# Mutually exclusive: Independence ($R \perp A$) vs Sufficiency ($Y \perp A|R$)

**Assumption:** $A$ and $Y$ are not independent

**Claim:** Then sufficiency and independence cannot both hold

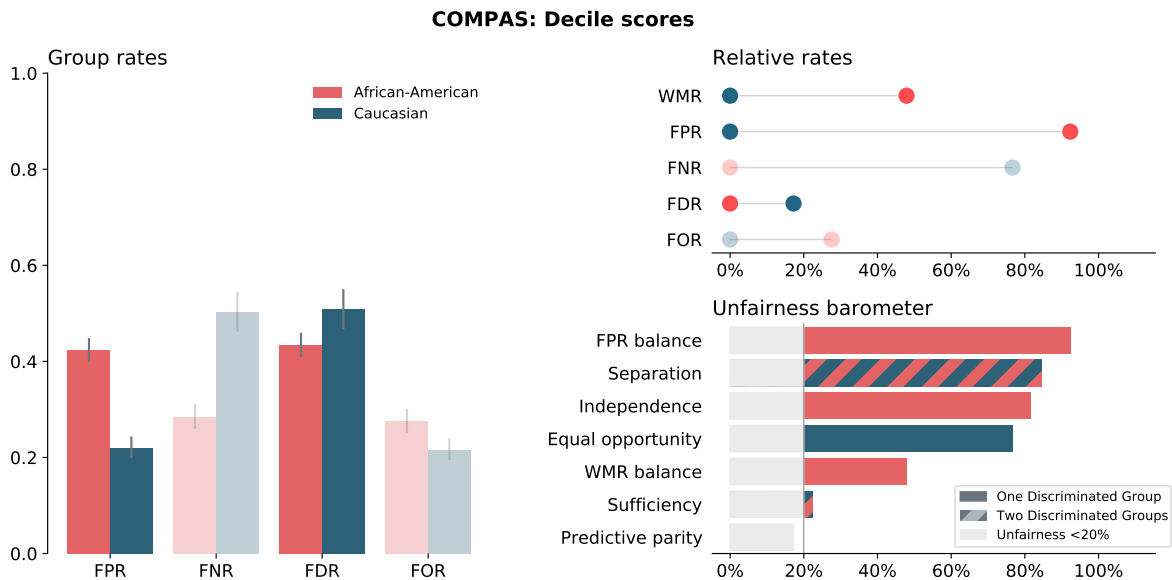



Figure: Figure from MSc thesis of CF Damgaard and E Zinck, 2022

# Mutually exclusive: Independence ($R \perp A$) vs Separation ($R \perp A|Y$)

**Assumption:** $Y$ is binary, $A$ is not independent of $Y$, $R$ is not independent of $Y$.
**Claim:** Then, independence and separation cannot both hold

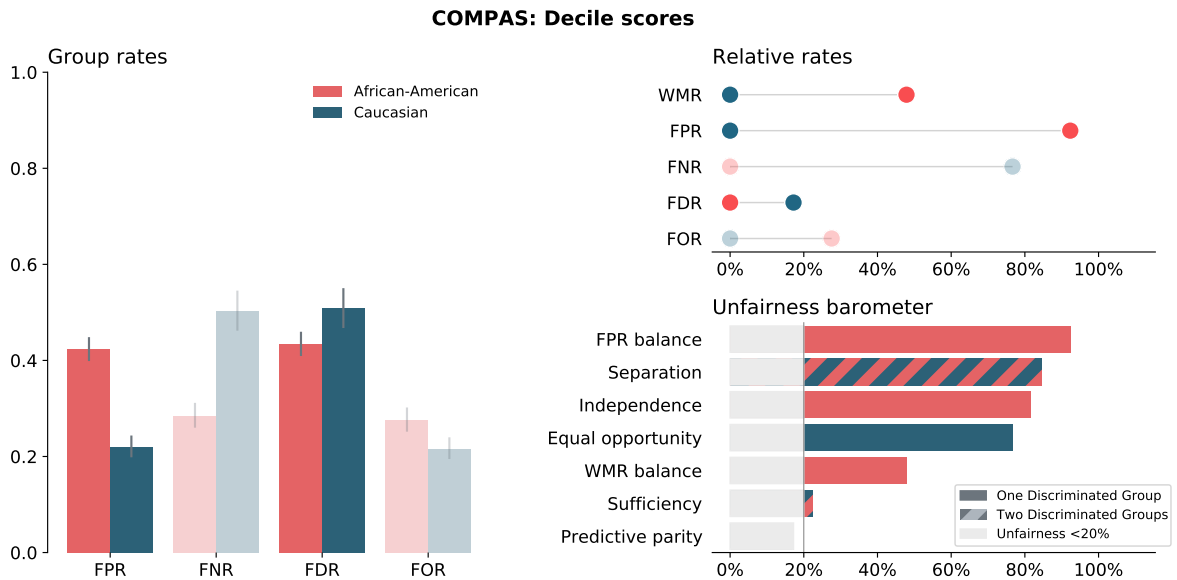



Figure: Figure from MSc thesis of CF Damgaard and E Zinck, 2022

# Mutually exclusive: Sufficiency ($Y \perp A|R$) vs Separation ($R \perp A|Y$)

**Assumption:** All events in the joint distribution of $(A, R, Y)$ have positive probability, and $A$ is independent of $Y$. Then, separation and sufficiency cannot both hold.

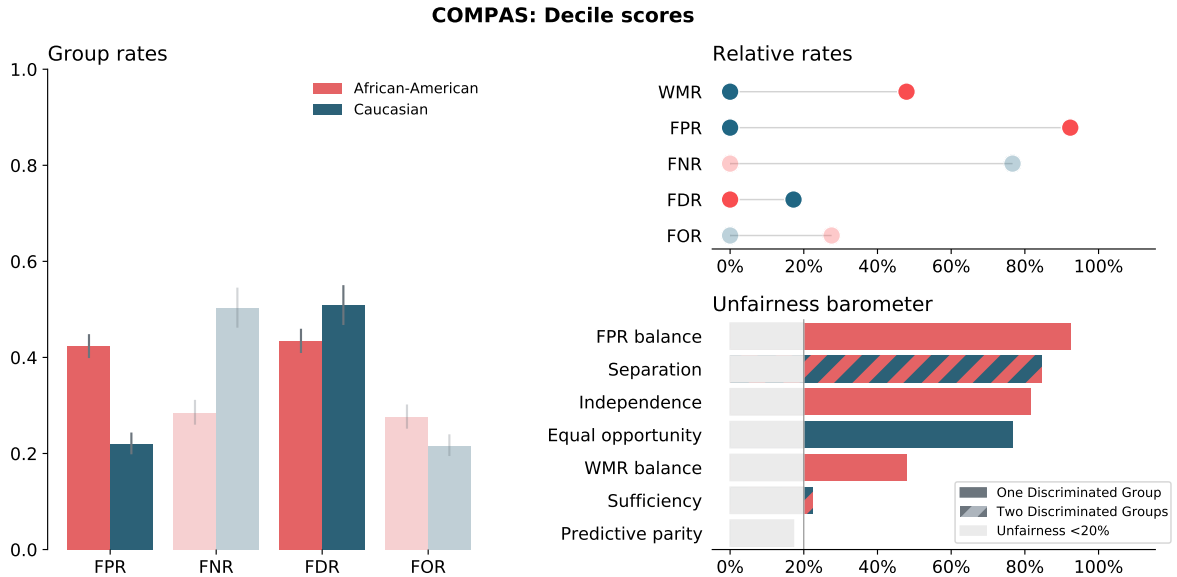



Figure: Figure from MSc thesis of CF Damgaard and E Zinck, 2022

Next:
Bias diagnostics – from fairness criterion to quantification

# Diagnostics based on algorithmic fairness criteria

**Recall:** The three categories of "fairness" correspond to equalizing across groups

- **Independence:** Acceptance rate $\mathbb{P}\{\hat{Y} = 1\}$ of the classifier
- **Separation:** The error rates $\mathbb{P}\{\hat{Y} = 0 | Y = 1\}$ and $\mathbb{P}\{\hat{Y} = 1 | Y = 0\}$ of the classifier
- **Sufficiency:** Outcome frequency $\mathbb{P}\{Y = 1 | R = r\}$ of a score $R$

**Diagnosing bias is easy!** A simple diagnostic of "unfairness" can be derived by looking at either difference or fractions of the associated rates or frequencies between groups.

# Summary

By now, you should:

- Be familiar with different ways in which demographic bias can become embedded in machine learning models
- Be familiar with the three most common fairness criteria, and variants thereof
- Know how to turn fairness criteria into diagnostic criteria for algorithmic bias
- Be familiar with incompatibility results for fairness criteria


## TL;DR

- Fairness & Medical Imaging workshop on **Oct 10th** at MICCAI 2024 (Morocco)
- The workshop will be only *fully* in-person (as in *no* remote live talk or *no* hybrid mode)
- Organized jointly with the workshop on Ethical and Philosophical Issues in Medical Imaging

## Call for Papers

We invite the submission of papers for

**FAIMI: The MICCAI 2024 Workshop on Fairness of AI in Medical Imaging**.

## Dates

*All dates are Anywhere on Earth.*

Full Paper Deadline: **June 24, 2024**

Notification of Acceptance: **July 15, 2024**

Camera-ready Version: **August 1, 2024**

Workshop: **October 10th, 2024**

For more information, check: https://faimi-workshop.github.io