Human and Machine Intelligence
in
Medicine

Sendhil Mullainathan
University of Chicago

Story 1

A common dilemma

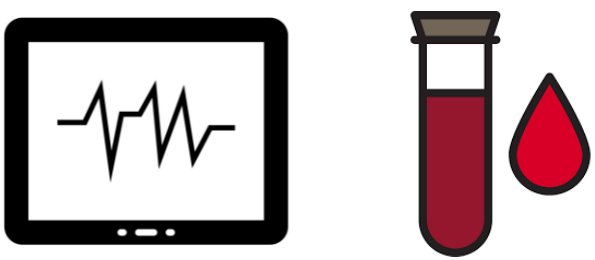
Chest pain

Trouble breathing

Nausea

...

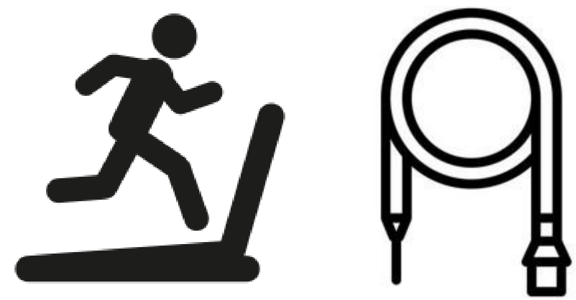
Easy but inconclusive tests



Electro-cardiogram

Labs

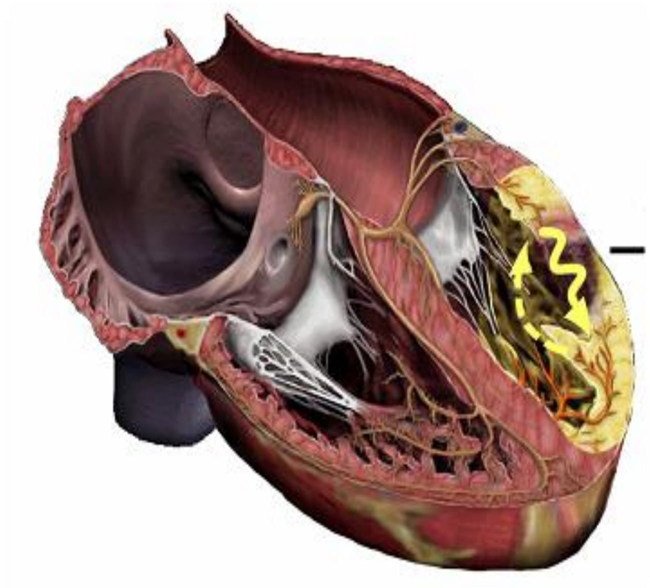
Test for heart attack?



Stress test

Catheterization

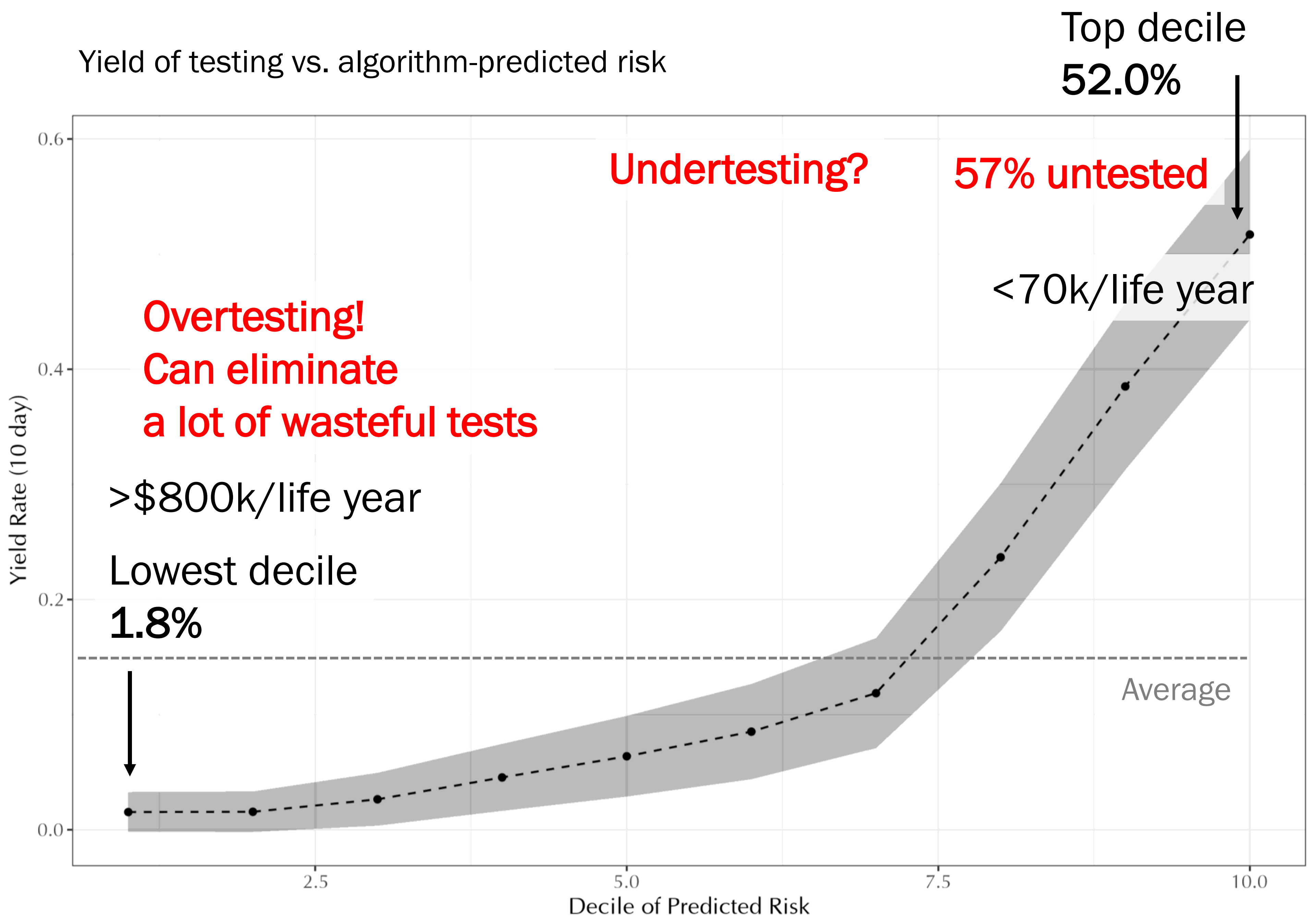
Why test for recent heart attack?



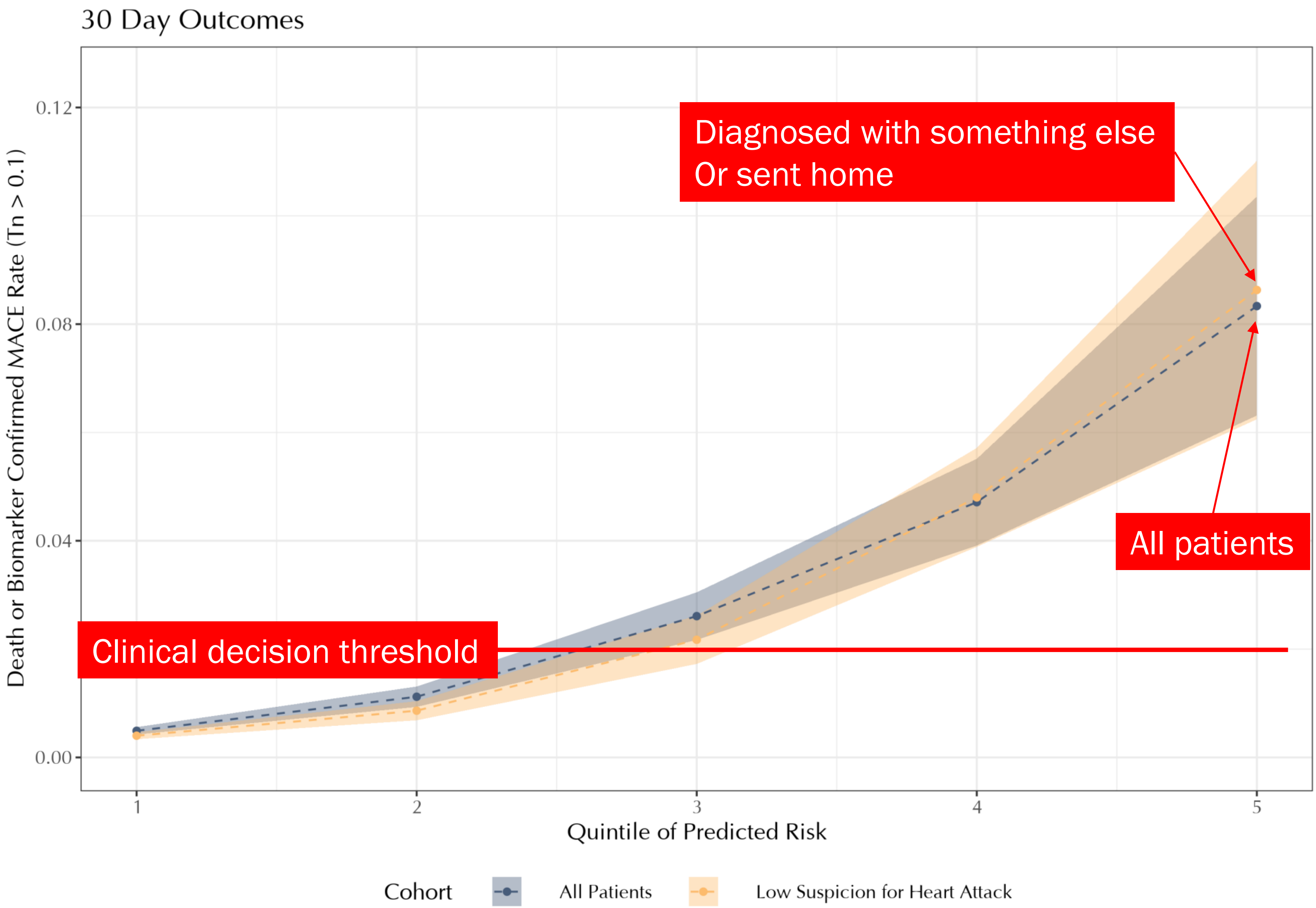
- Immediate and delayed consequences
  - Is 10-20% fatal
  - Long-term complications
- Treatment is effective
  - Stenting, bypass surgery
  - RCTs: ~50% reduction in mortality and sequelae

Can we build an algorithm? A decision-aid

- Predict given (some of) what physician sees…
- Whether a patient will have a positive stress test or catheterization
- Will help us understand mistakes in testing


Yield of testing vs. algorithm-predicted risk
Top decile
52.0%
Undertesting?
57% untested
<70k/life year
Overtesting!
Can eliminate
a lot of wasteful tests
>$800k/life year
Lowest decile
1.8%
Average
0.6
0.4
0.2
0.0
Yield Rate (10 day)
2.5
5.0
7.5
10.0
Decile of Predicted Risk

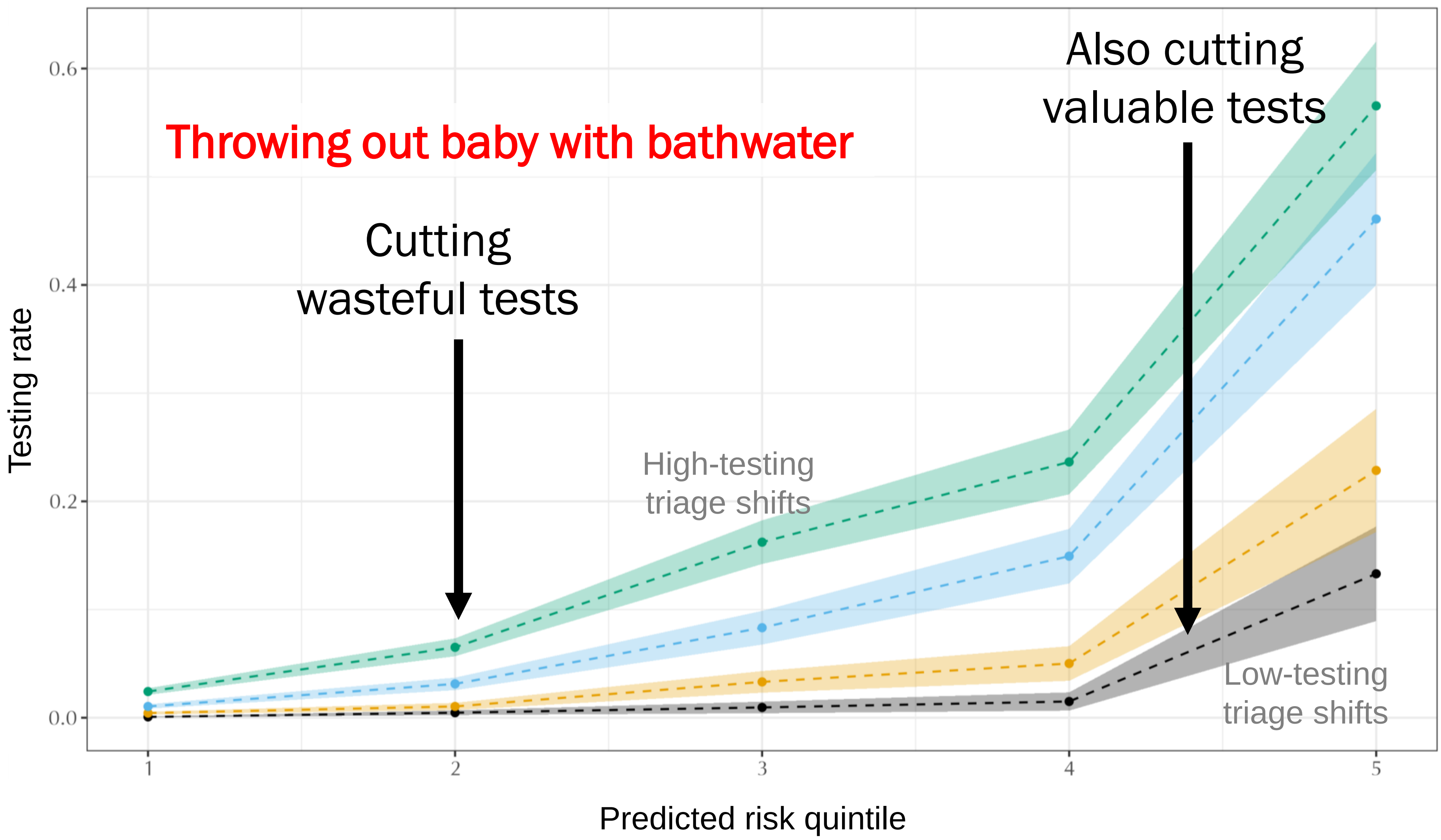
Adverse events + death in the untested (30 days after visit)



Diagnosis

- Conversation is around over-use
- In actuality there's a lot of <u>both</u> over and under use
- Algorithms can see things we can't

What happens when we change behavior?



Diagnosis

- Conversation is around over-use
- In actuality there's a lot of <u>both</u> over and under use
- Algorithms can see things we can't
- Can even lead us to rethink the underlying source of the problem
- Decision-aids can make a big difference

Story *2*

Care Coordination Programs

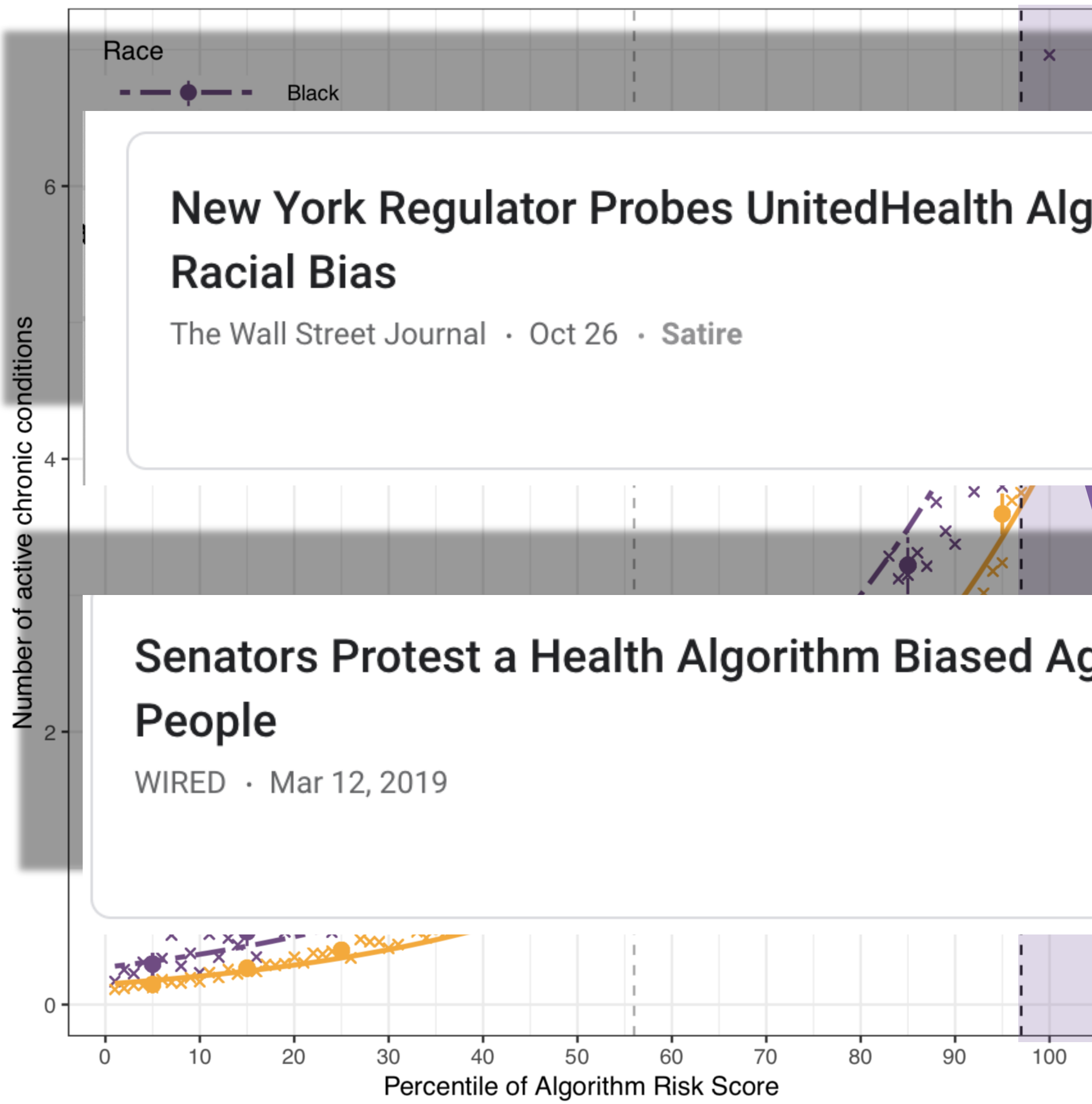
- Patients with many chronic condition are an epicenter of costs
- Programs target them with extra resources
- To target patients an algorithm is used
  - Already at scale, > hundred million patients

Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.*

Measuring Racial Inequity

- Access to one live, scaled private *sector* algorithm
  - One of the largest of a group of providers (~10 mill)
- Consequences for who gets in the program
  - What kinds of *Whites* and *Black* would be chosen for program (in terms of health)
- Since program allocated by level of score *S*, we can ask…
  - Two patients, same risk score: one back and one white
  - Who is sicker?

We studied Racial 'Bias'



- Principle: Same score

should not matter

New York Regulator Probes UnitedHealth Algorithm for Racial Bias

The Wall Street Journal · Oct 26 · Satire

Senators Protest a Health Algorithm Biased Against Black People

WIRED · Mar 12, 2019

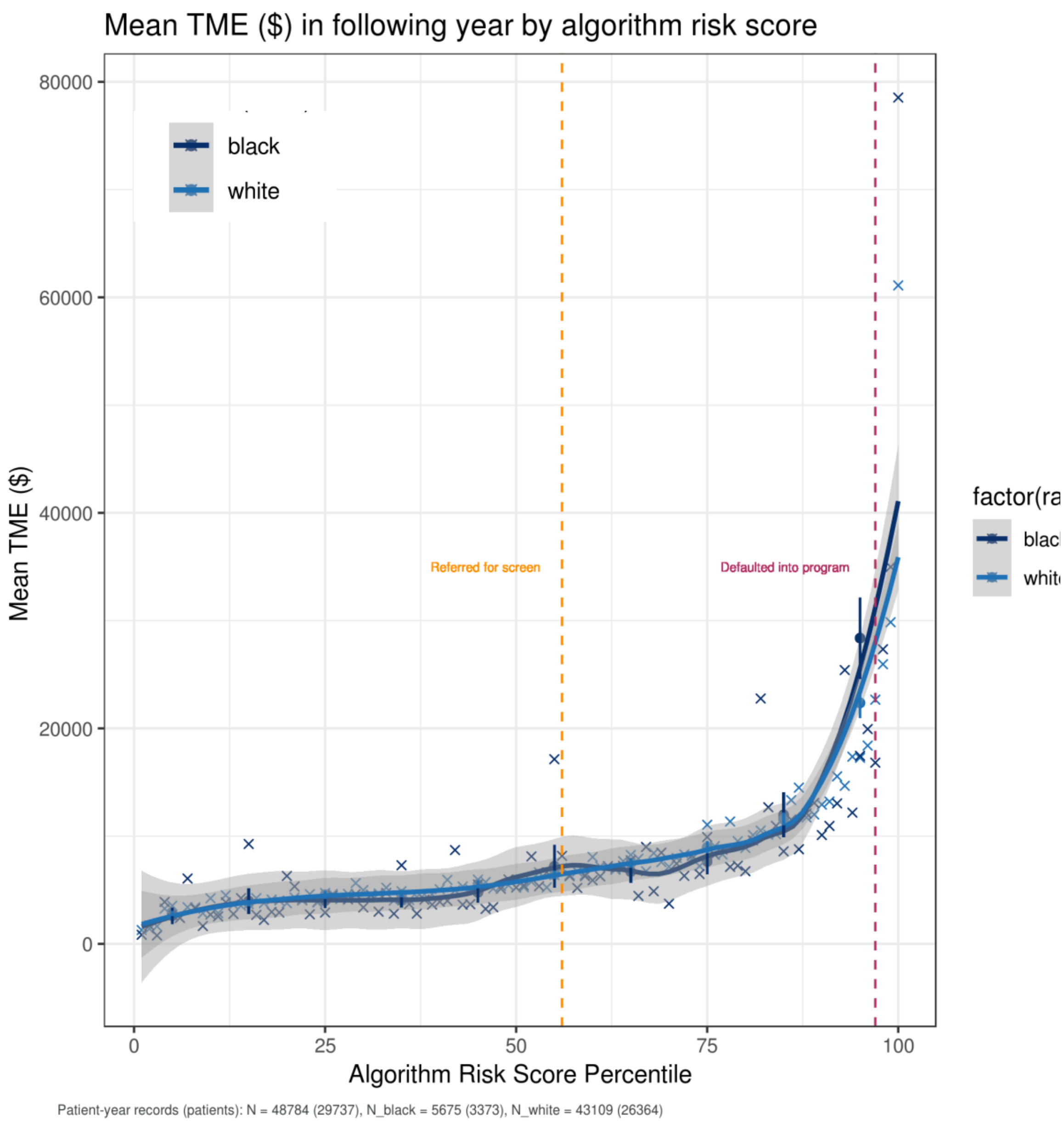
Obermeyer, Powers, Vogeli, Mullainathan, *Science* 2019

Dissecting the Inequity

- Where is the algorithm going wrong?

- One clue is in where it is going right

Where is algorithm going wrong?



Algorithm well calibrated by race for total health utilization

The Problem of Predicting Utilization

- Blacks and whites do not have same relation between health status and utilization

- Whites have better access to health care

- At every level of health blacks utilize less health care

- So accurate utilization prediction = biased health prediction

Dissecting The Problem

- Proximal cause: the *Label*
  - Algorithm optimized objective it was given
  - But that's not our full objective

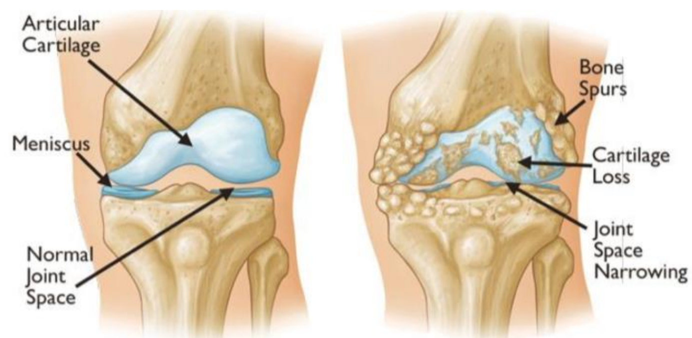
- Deeper cause –
  - Why was costs chosen and not health?

- Utilization and Health are often used synonymously

Story 3

Knee Pain

- Osteoarthritis most common joint disorder in US
- 10% of men over 60 and 13% of women over 60 have knee osteoarthritis



Pierson et al. "An algorithmic approach to reducing unexplained pain disparities in underserved populations," Nature: Medicine (2021)

Disadvantaged patients experience more pain…

	Pain gap
Race	**10.6** (8.3, 12.9)
Income	**4.2** (2.8, 5.6)
Education	**5.3** (3.7, 6.7)

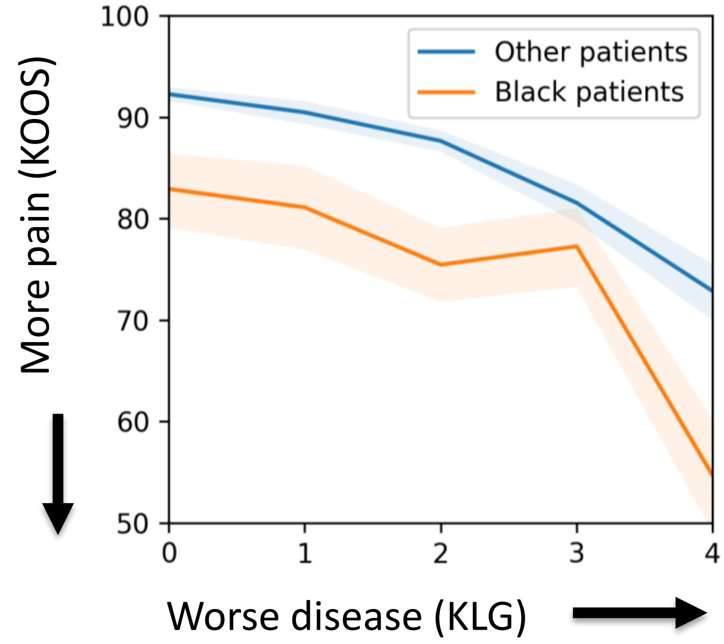
Why is there a Pain Gap?

- "Inside their knees"
  - Physical ailments more extreme

- "Outside their knees" - non-knee-related factors mean that the same physical knee problem results in more pain in some groups
  - Life stress (eg, tough bus-driving job)
  - Less access to pain medication
  - Different pain-coping strategies
  - Less social support

Disadvantaged experience more pain….

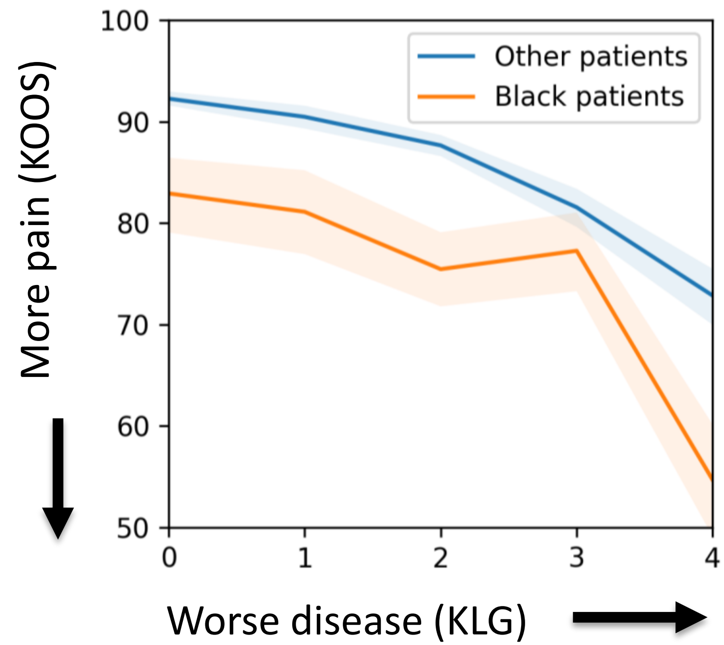
	Pain gap
Race	**10.6** (8.3, 12.9)
Income	**4.2** (2.8, 5.6)
Education	**5.3** (3.7, 6.7)



Even when controlling for severity

Disadvantaged experience more pain….

	Pain gap
Race	**10.6** (8.3, 12.9)
Income	**4.2** (2.8, 5.6)
Education	**5.3** (3.7, 6.7)



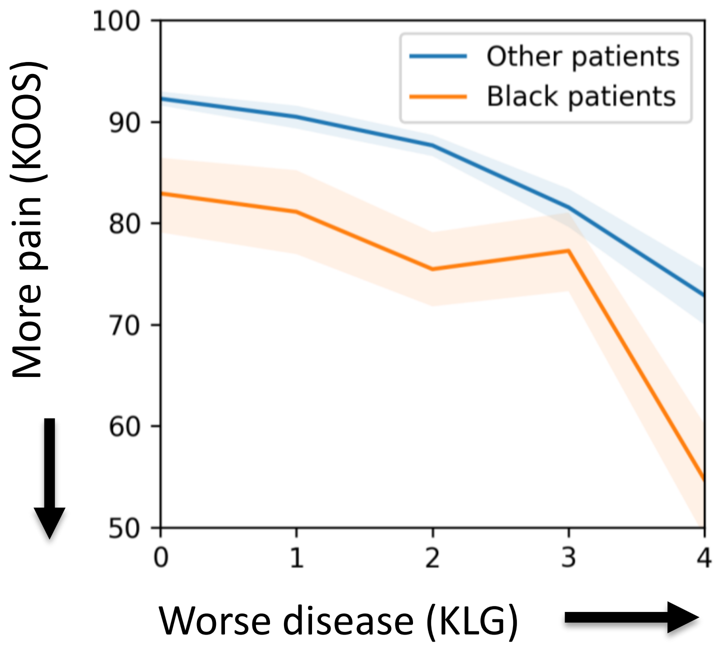
$$pain \sim race + KLG$$

regress pain on race and KLG
pain gap = race coefficient when controlling for KLG

Even when controlling for severity

Disadvantaged experience more pain….

	Pain gap (no controls)	Pain gap (control for KLG)
Race	**10.6** (8.3, 12.9)	**9.7** (7.4, 11.9)
Income	**4.2** (2.8, 5.6)	**3.5** (2.3, 4.9)
Education	**5.3** (3.7, 6.7)	**4.9** (3.5, 6.2)



$$pain \sim race + KLG$$

regress pain on race and KLG
pain gap = race coefficient when controlling for KLG

Even when controlling for severity

Does this settle the matter?

- Medical knowledge is, for most thing, still in flux
  - It's why we are still doing the science

- We know we don't understand pain that well. KLG doesn't explain pain well ($R^2 = 0.10$).

- Maybe there is something in the knees *we don't know about*?

- Are there overlooked physical features in the knee which explain the higher pain levels in disadvantaged groups?

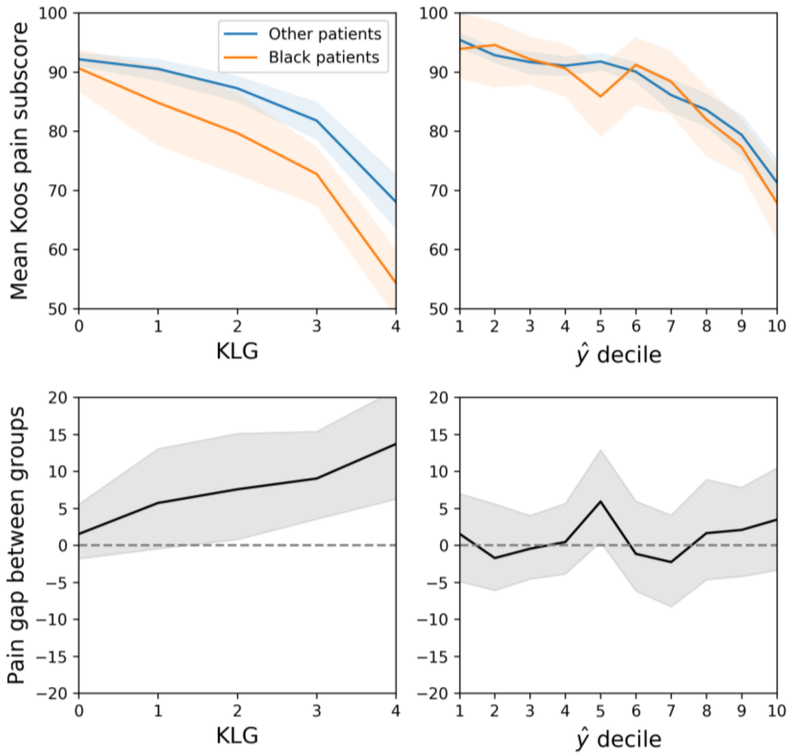
Using ML for Discovery

- Train convolutional net to predict pain from knee x-rays
- Input: image of both knees
- Output: predict Koos pain score in the knee

Pain score: 83.3

- Key: Algorithm only sees x-rays
  - Does not have access to other pieces of information that may predict pain e.g. lab values that signal inflammatory measure

Algorithm finds signal that reduces gap



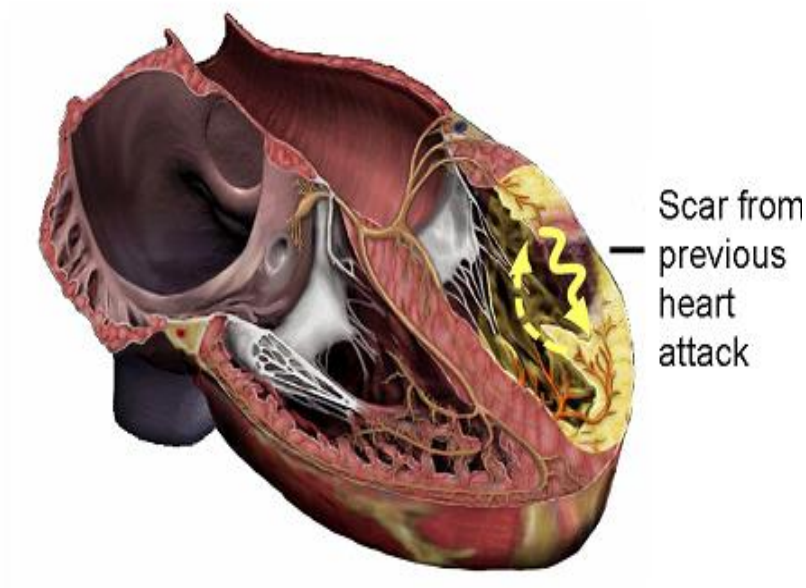
Implication:
Overlooked signal in knee x-ray which helps explain disadvantaged patients' higher pain

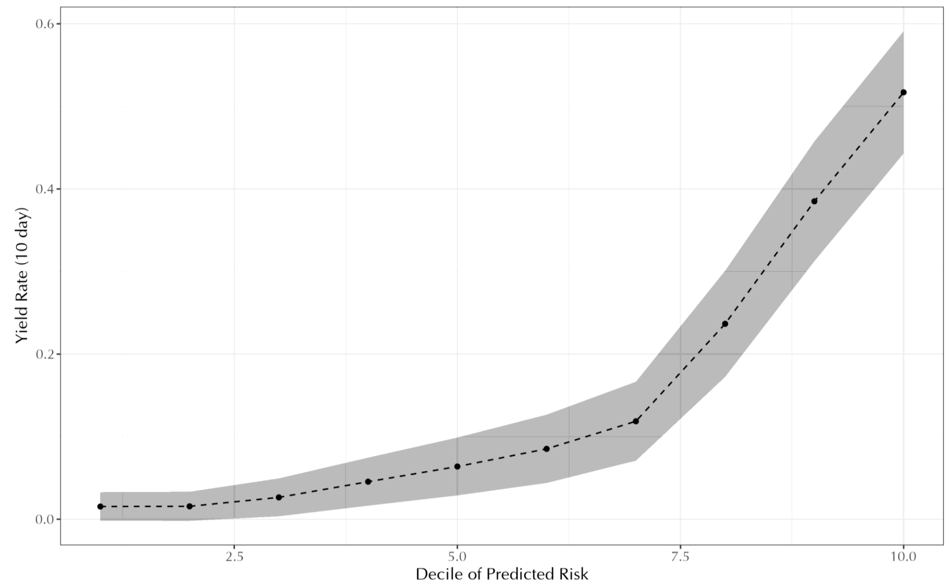
In their knees
What patients have been trying to tell us all along!

Algorithms a force for *equity*

Lessons

1. Data not algorithms the scarce resource





This is more and more commoditized

Very little difference in performance by skill

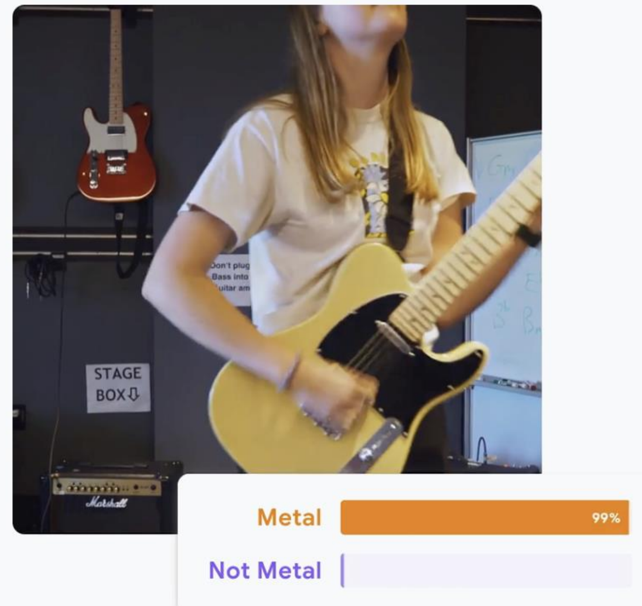
Auto ML

About FAQ Get Started

Teachable Machine

Train a computer to recognize your own images, sounds, & poses.

A fast, easy way to create machine learning models for your sites, apps, and more – no expertise or coding required.



Get Started

p5.js Coral

The scarce resource

- Technical skill still important in some cutting edge problem
- But it's sufficiently diffused that the real edge here goes to…
- Finding the right problem
- Having the right data
  - In medicine, this is the biggest bottleneck

Lessons

1. Data not algorithms the scarce resource

2. AI breaks because the data is broken

Google team - Image Pathology Model

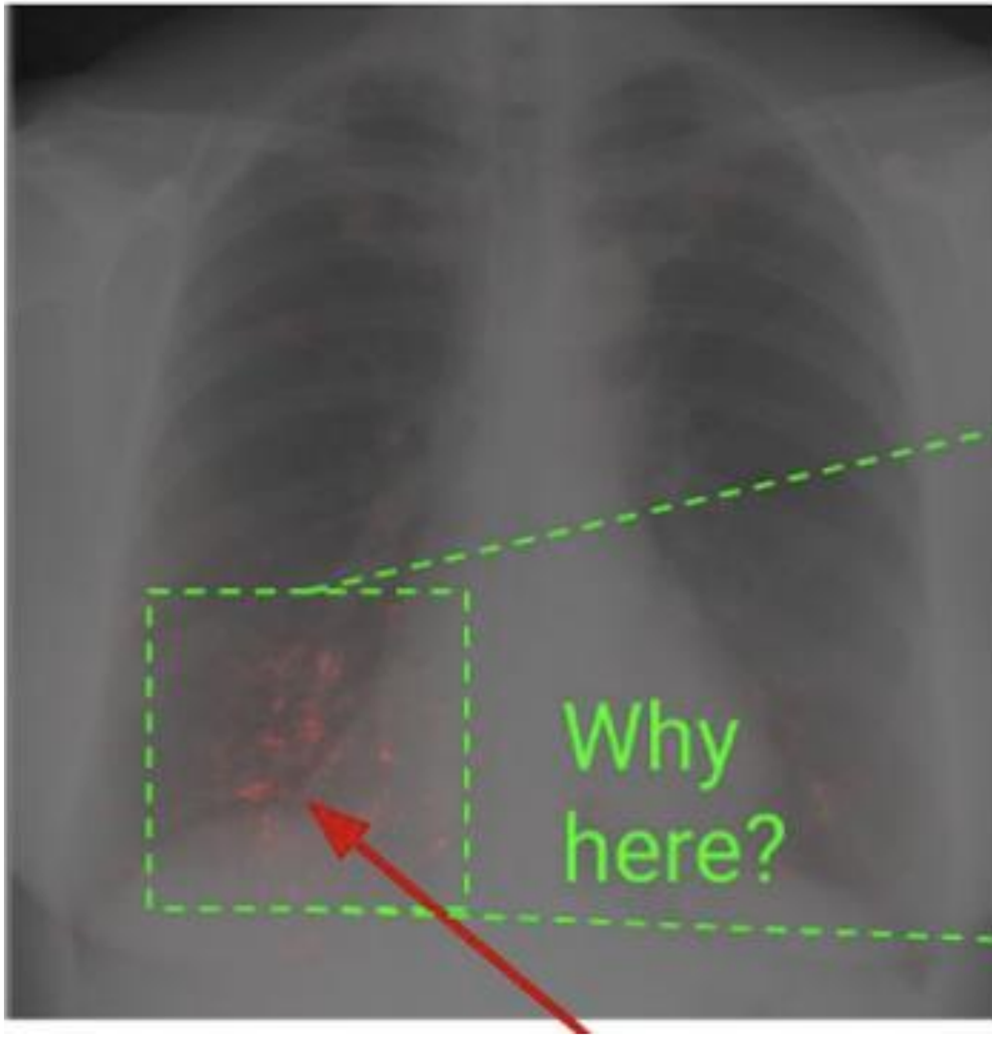
An Example from
Mukund Sundarajan

Algorithm to detect
pathologies in chest xrays

Very successful –

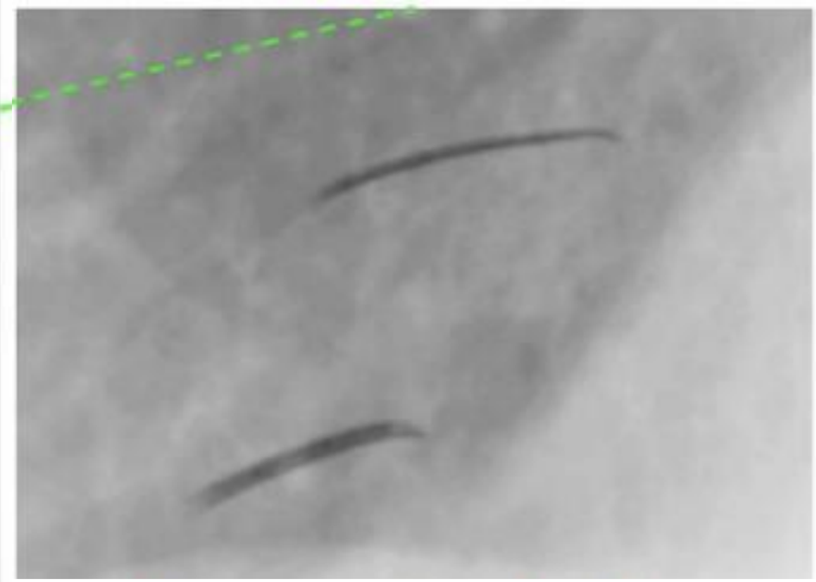
Team got interested in
“What is algorithm
looking at?”

Pathology Detector

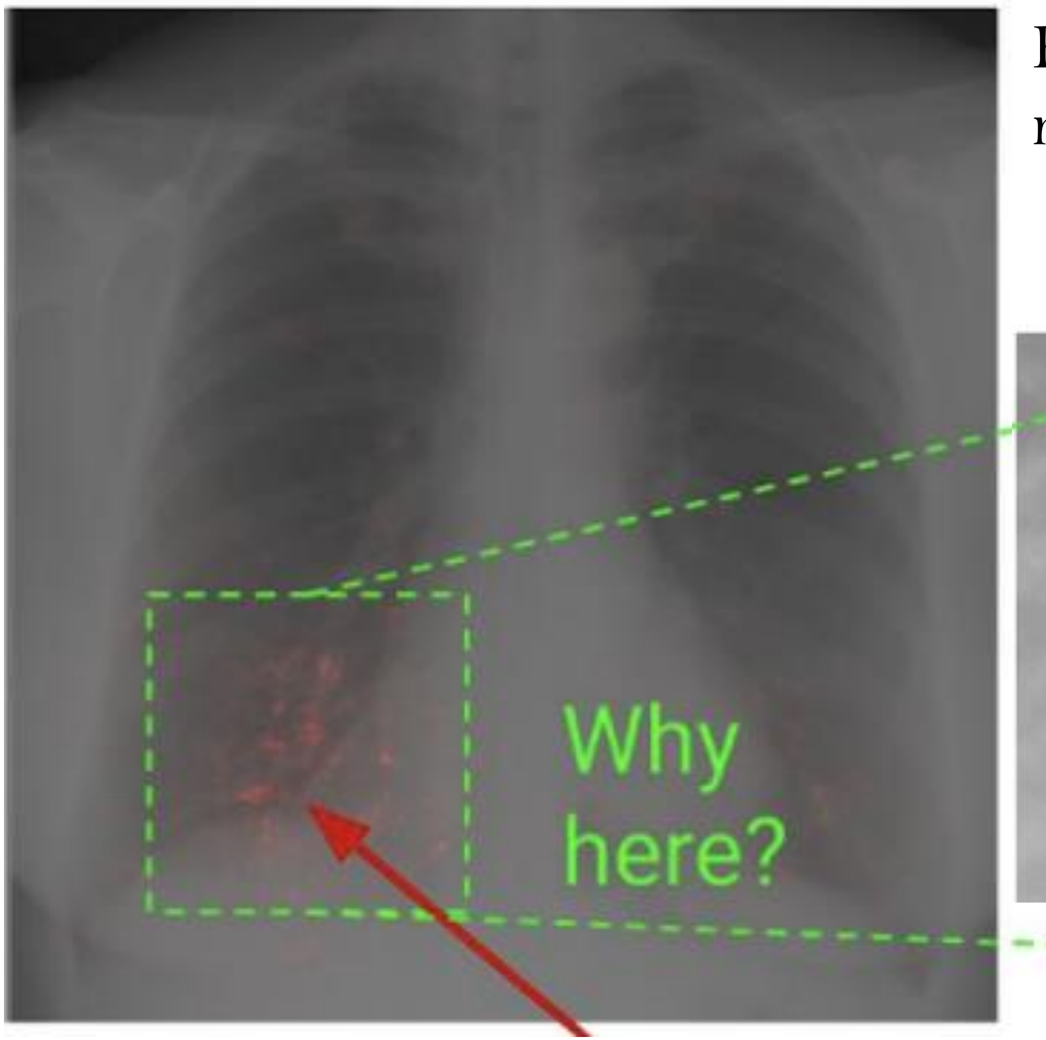


Zoom in and adjust contrast

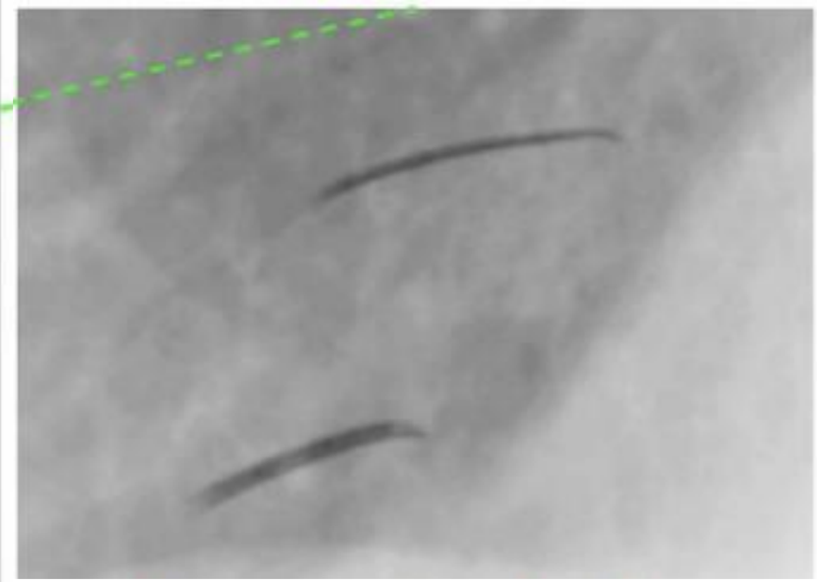
What is this?



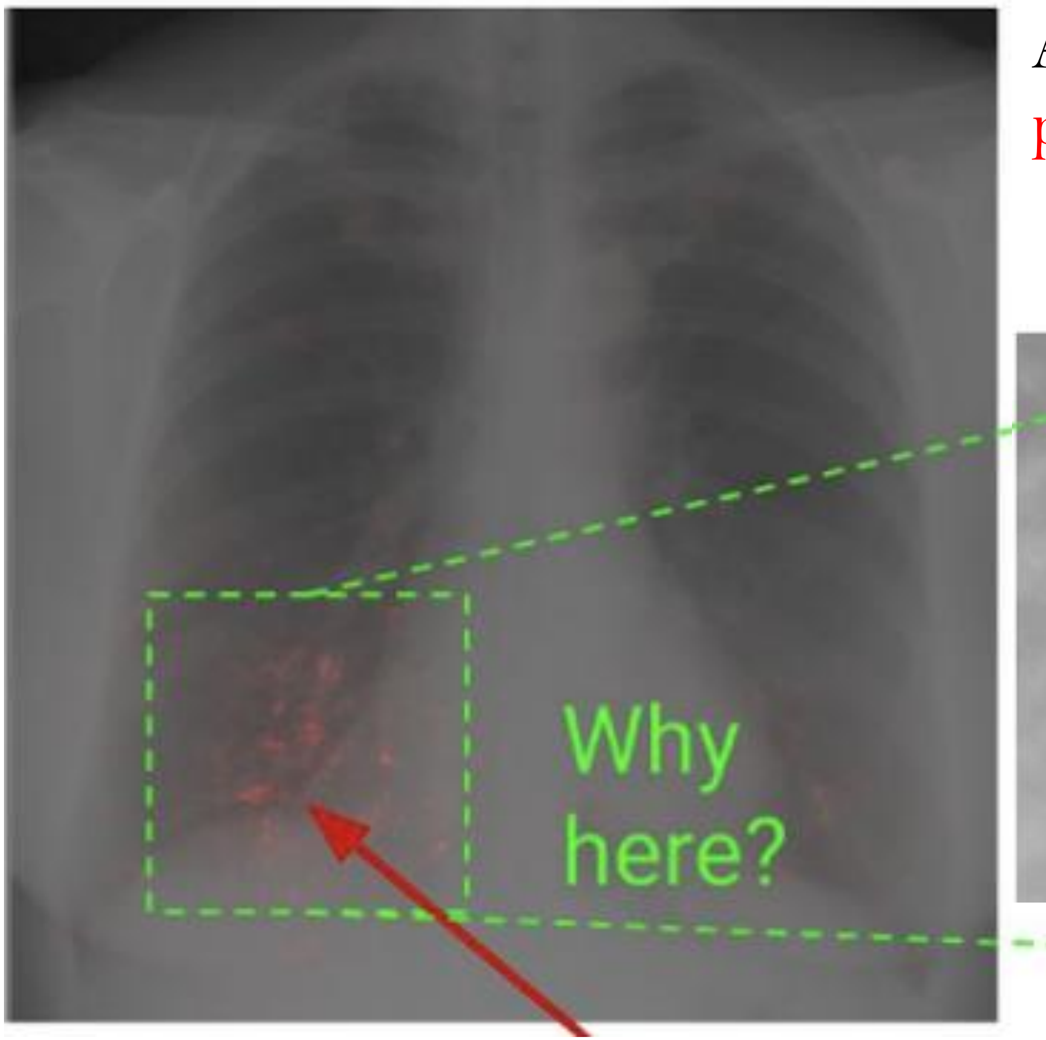
Pathology Detector



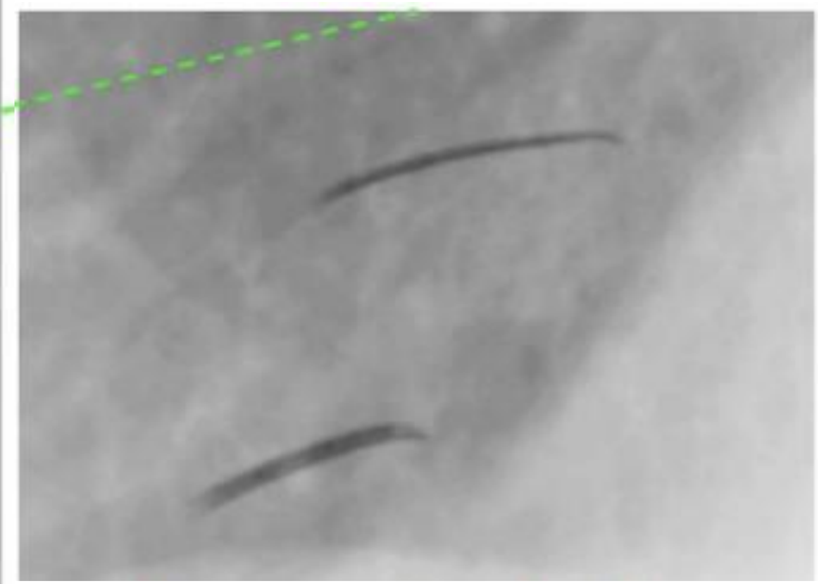
Radiologists penmarks to note problems



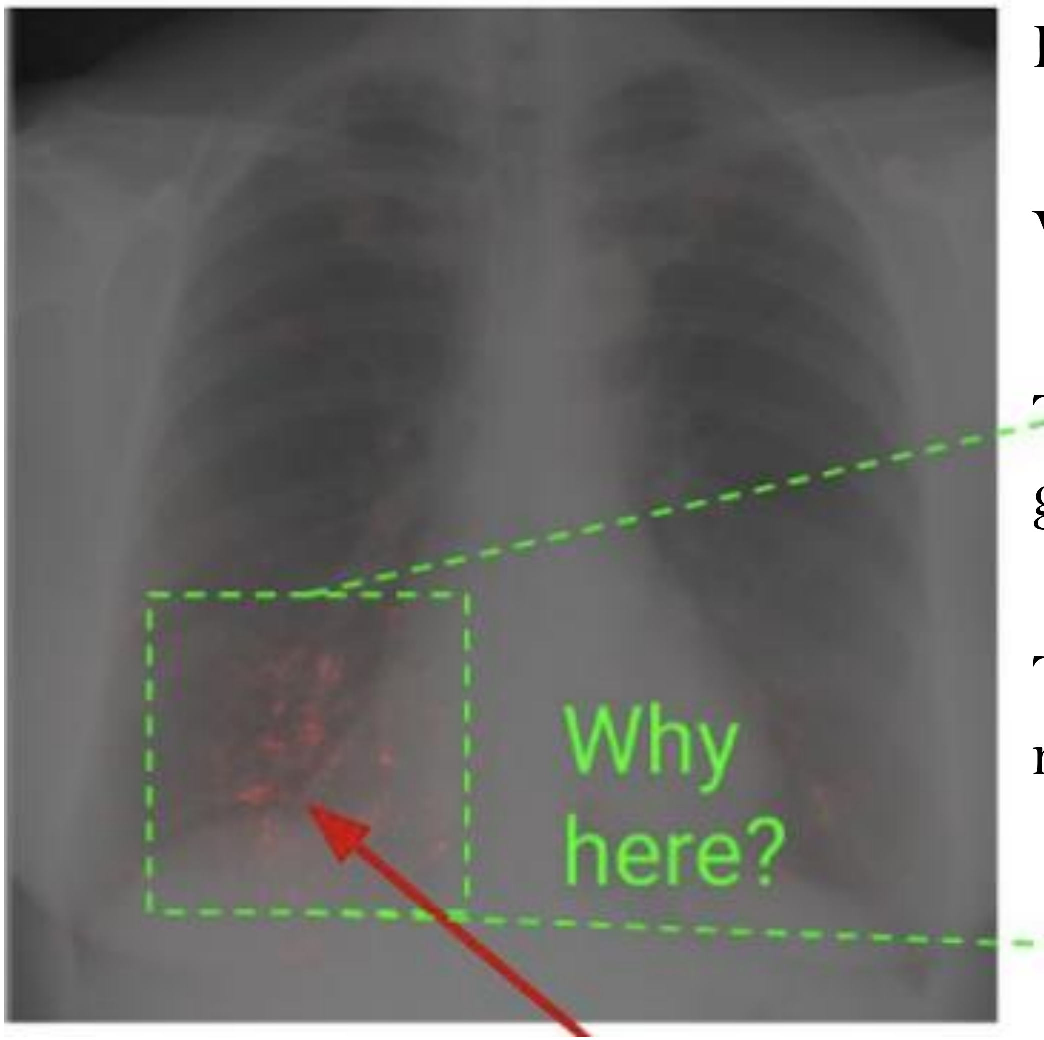
[redacted] Penmark Detector



Algorithm detected
penmarks not pathologies



[redacted] Penmark Detector



Performance was high

Very misleading

The algorithm is only as good as the data

The data itself was misunderstood

Bugs in Data

- We recognize (and fear) bugs in code
  - We think the code is doing on thing
  - But it is actually doing something else
- ML systems have code as well
  - But these don't break
  - All AI failures I know of, the algorithm did exactly what it was asked to do
- Their real code is the training data
  - This is where bugs arise.
  - We think the data is one thing
  - But it is actually something else
- Who is responsible for debugging?
  - Those who know the data! Not the data scientist.

Data breakage

- Is the label the one you wanted?
  - Racial bias in care coordination programs

- Is the data representative of deployment?

Lessons

1. Data not algorithms the scarce resource

2. AI breaks because the data is broken

3. Unrepresentative data

Research Letter

September 22/29, 2020

Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms

Amit Kaushal, MD, PhD[1]; Russ Altman, MD, PhD[1]; Curt Langlotz, MD, PhD[2]

Fifty-six studies (76%) trained algorithms using at least 1 geographically identifiable cohort. Cohorts from California appeared in 22 of the 56 studies (39%), cohorts from Massachusetts in 15 (27%), and cohorts from New York in 14 (25%) (**Table**). Forty of 56 studies (71%) used a patient cohort from at least 1 of these 3 states. Among the remaining 47 states, 34 did not contribute any patient cohorts, and the remainder contributed between 1 and 5 cohorts (Table).

Unrepresentative Data

- Even human knowledge is based on unrepresentative data
- Why do KL scores under-recognize pain in disadvantaged?
- Another kind of unrepresentative:
  - Academic medical centers

Lessons

1. Data not algorithms the scarce resource

2. AI breaks because the data is broken

3. Unrepresentative data

4. Prediction not Emulation

Two Different Approaches

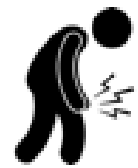
KOOS pain score

Physician Judgment (KL Grade)

Choice of Label

Predict

Pain score



Prediction

Highest return

Predict

KL Grade

Automation

Most of ML in Medicine

Some cost savings
Automate errors

Lessons

1. Data not algorithms the scarce resource

2. AI breaks because the data is broken

3. Unrepresentative data

4. Prediction not Emulation