Diagnosing Physician Error with Machine Learning

Ziad Obermeyer
UC Berkeley

Joint work with Sendhil Mullainathan
University of Chicago

Today's agenda

- Our health care system is broken
  - $4.3T/year in spending; worsening and unfair outcomes
- A microcosm of this: Testing for ACS in the ED
  - Wasted tests: up to 90%)
  - Missed MI: still top malpractice claim
- Can AI provide a way out?
  - Cut testing in predictably low-risk patients
  - Reallocate some of those to untested high-risk patients
  - Lower cost AND better quality

Important question: What is ACS?

- Not a physiology question
  - Blockage in coronary arteries causing infarction
- A data question
  - AI is just data—which variable is it predicting?
  - Troponin? ST-elevation?
- How would we get the data if money were no object?
  - How do they do it in pharmaceutical RCTs?

Common solution: substitute human judgment

- ML has adopted this 'human labels' playbook wholesale
  - Diabetic retinopathy (Gulshan et al., JAMA 2016)
  - Many studies of ECGs, digital pathology, …
- What is the algorithm learning?
  - How to automate human judgment, bias, and error
- This will not solve problems of our health care system
  - It will replicate and even scale them up
- How to get AI to learn from nature, not humans?

What we do

1. Train AI to predict test outcomes
   - Back to basics: Blockage in coronary arteries on cath
   - A good (but not perfect) proxy for ground truth

2. Compare predictions to patient outcomes
   - In the tested: Easy
   - In the untested (98-99%): hmmm
     - Detective work to find proxies for missed MI
     - As-good-as-random variation in testing

3. Diagnose human errors and cognitive biases
   - By comparing human decision to AI ‘decision’

Prediction setup

Features

Outcomes

t_0: ER visit

Over 2 years before visits, construct candidate features

k = 16,381

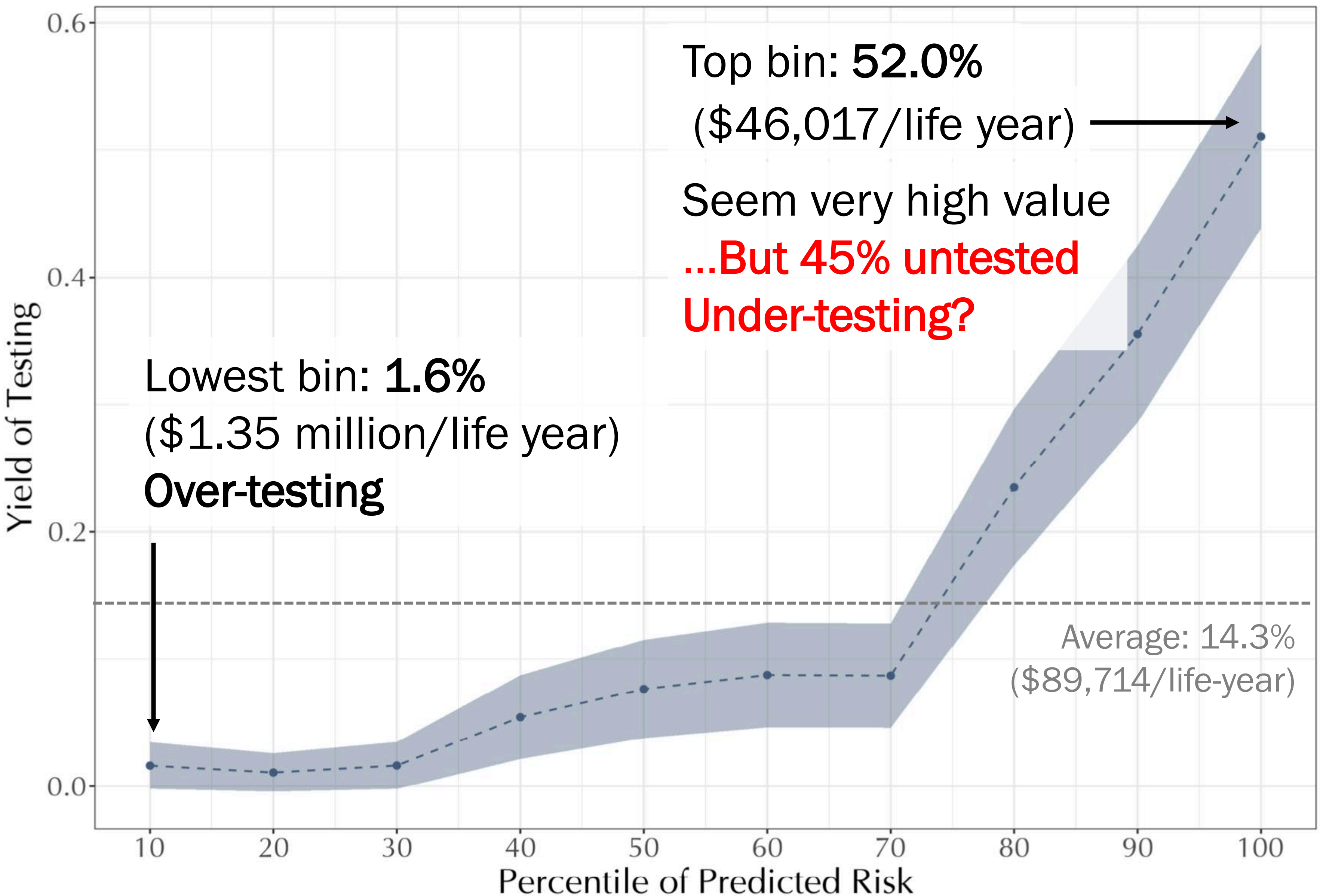
Over 10 days after visits, observe

- Tests, Treatment

- n = 246,265 ER visits (129,859 patients), 2012-15
  - Remove: ≥80yo, serious illness, nursing home, etc.

- Train ensemble to predict blockage in 3/4 random sample
  - Show results from 1/4 hold-out set only

Tested patients: Predictable variation in yield



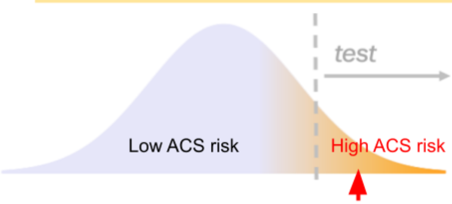
Untested patients: Selection bias makes this much harder

- Yes, physicians **fail to test** apparently high-risk patients

- But physicians may fail to test for **good reasons**
  - Symptoms, exam, ECG, labs, ...

Example: Algorithm sees everything ***up until triage...***

Recommendation: Strongly consider testing
Risk 4x accepted thresholds*



Mr Wright -------- 93rd pctile

45% chance of ACS on cath
12% 30-day adverse event rate if untested

Traditional risk factors
- Age over 50 (Age 64)
- Prior MI
- High recent LDL (203)

Other risk factors
- Low income (<$50k)
-

*Algorithmic predictions exceed HEART, TIMI, GRACE risk thresholds

...but not ***physical exam***

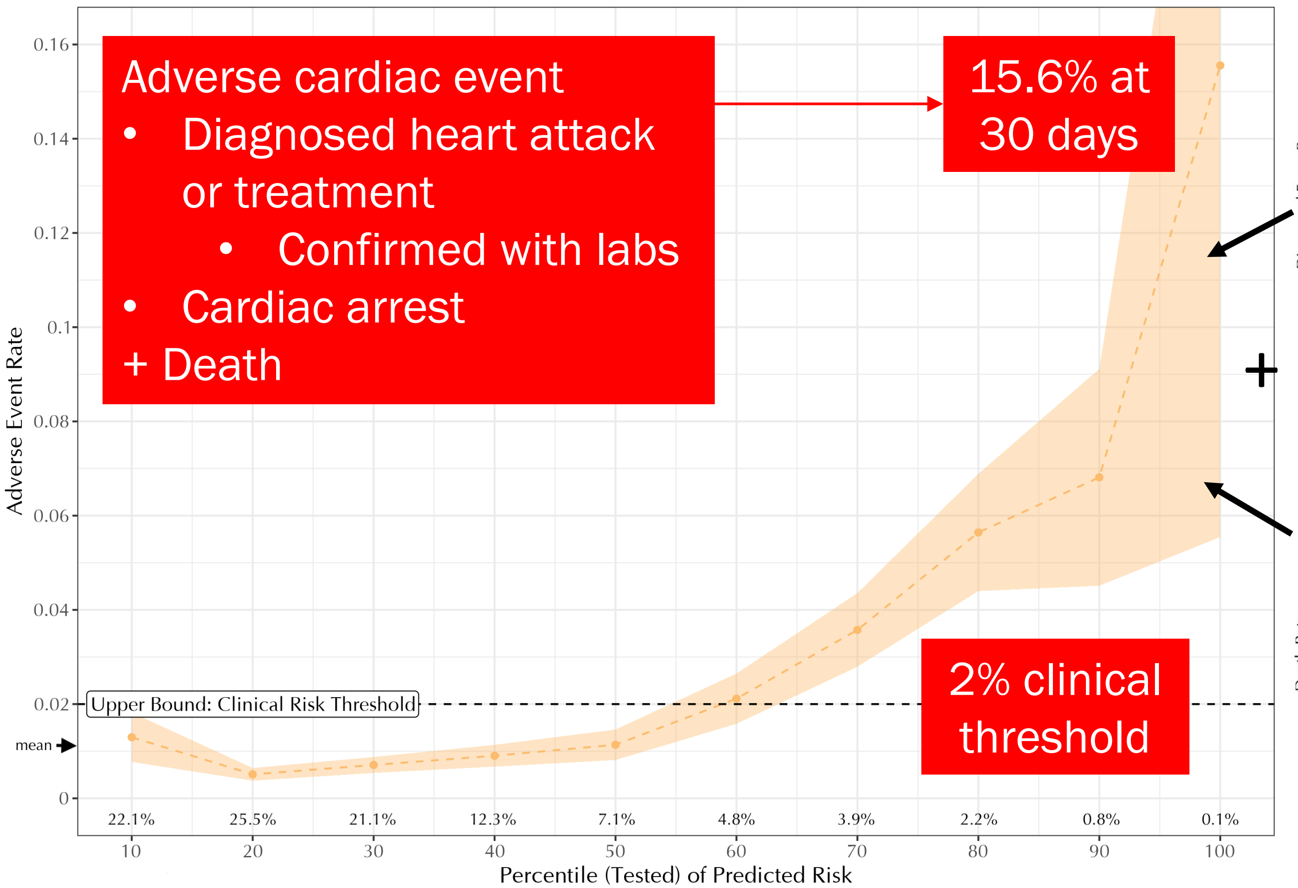
Untested patients: Selection bias makes this much harder

- Yes, physicians fail to test apparently high-risk patients
- But physicians may fail to test for good reasons
  - Symptoms, exam, ECG, labs, …
- In the tested: We looked at test result to see who's right
  - In the untested: No test results!
- Detective work
  - Solution 1: Adverse events in untested
  - Solution 2: Quasi-experiment that shifts testing rate

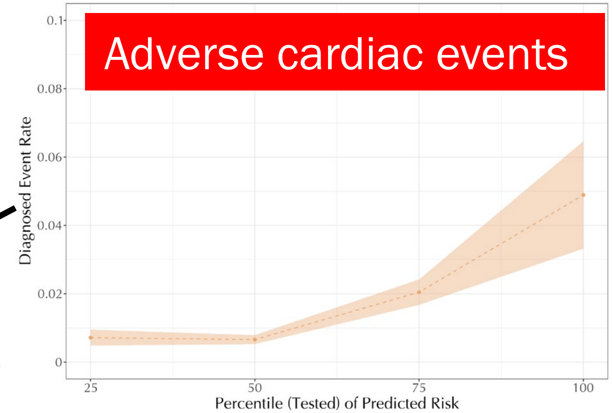
1a. Untested patients: Short-term adverse events

*excluded: usual suspects (frail), those with diagnosed heart problem in ER

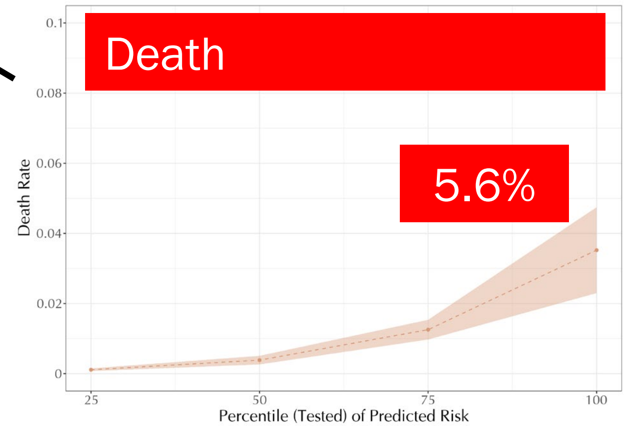
Total Adverse Event Rate



Components



+



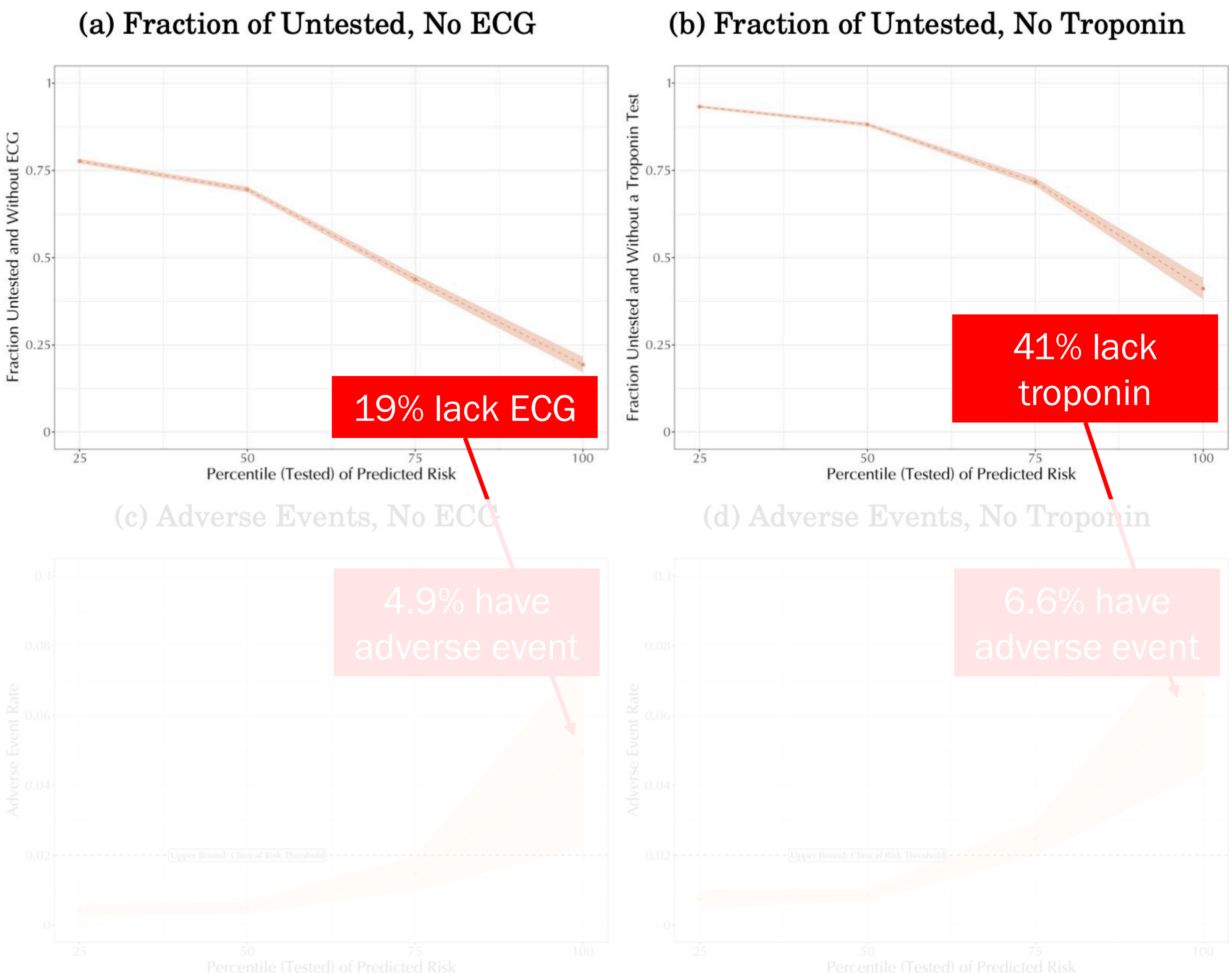
Would these patients benefit from treatment?

- Adverse events show high-risk people are truly high risk
  - But physicians may be aware of this risk
  - And decide not to test because of limited benefit
    - e.g., in the frail we haven't managed to exclude

- Insight: Low-cost screening tests proxy for suspicion
  - ECG, troponin done on everyone—even very low risk
  - And even those with low treatment benefit

- Adverse event rate in unsuspected patients: Lower bound
  - Here, physicians are unaware of heart attack risk
  - So failure to test can't reflect private information

1b. Untested, unsuspected patients: Short-term adverse events



2. Quasi-experiment that moves testing rate

Does testing improve health on average?

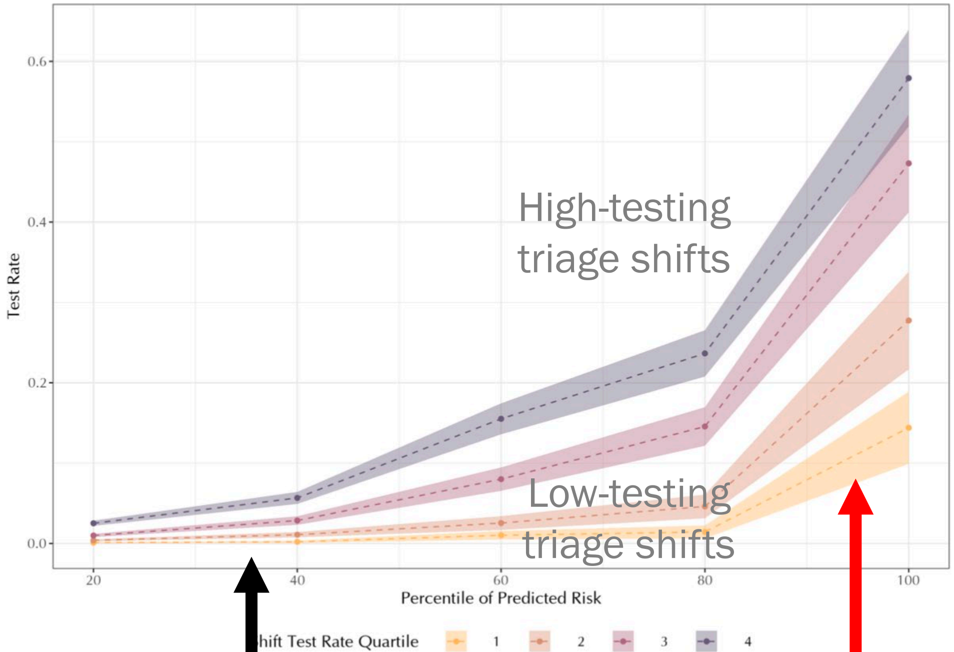
- Compare all patients on high-testing shifts
  - Vs. low-testing shifts
- No difference in heart attack rates, death rates
- Looks like "flat of the curve", wasteful testing

But the average patient isn't having a heart attack!

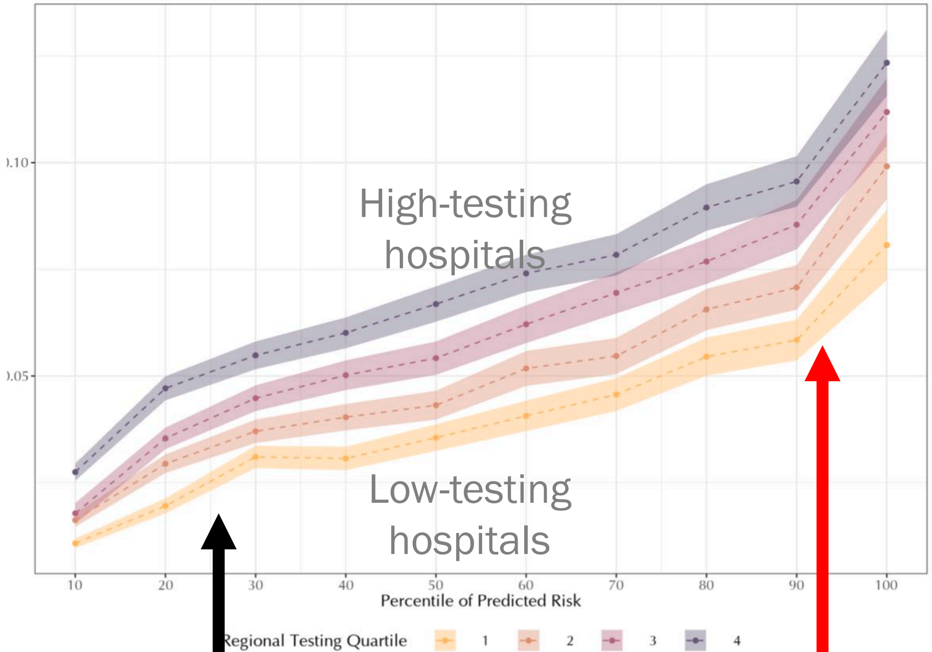
- Zoom in: highest-risk 1-2%
- When these patients walk in on high-testing shifts
  - They die 32% less over the next year
- Testing is wasteful on average—but not for those with heart attack!

Policy implication: Incentives can backfire

(a) Hospital Sample

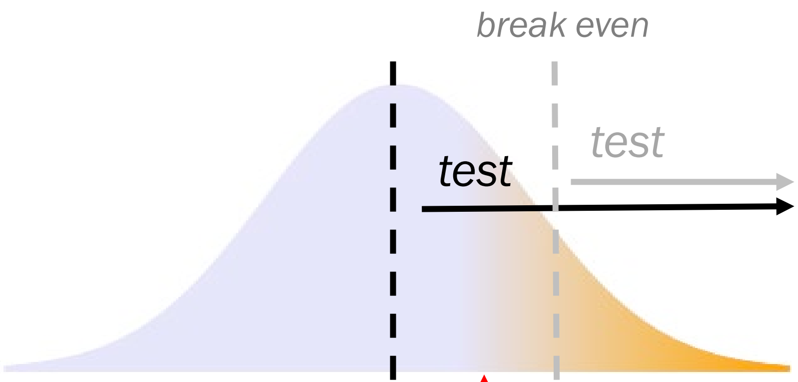


(b) National Medicare Sample



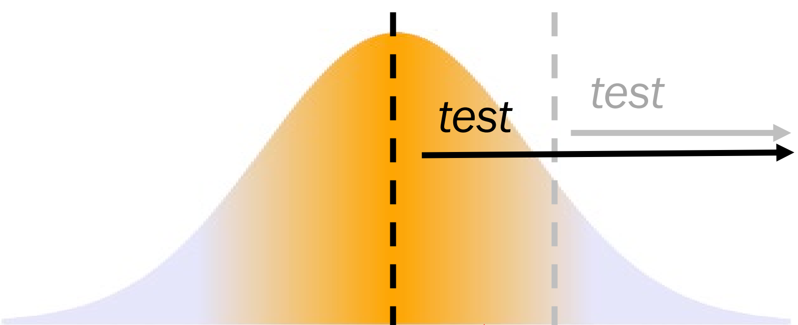
- Low-testing **physicians** cut wasteful tests
  - And also valuable tests
- Low-testing **hospitals** cut wasteful tests
  - And also valuable tests

Why do physicians go wrong? Two behavioral models



Incentives

- Test over a threshold
  - Threshold too low
- Low average yield



Errors

- Test high and low risk
  - At any threshold
- Low average yield

Mis-prediction: Untested high-risk patients

The nature of physician mis-prediction

- We examine how testing decisions deviate from risk
  - Clinical judgment vs. statistical models

- Specific tests of two hypotheses

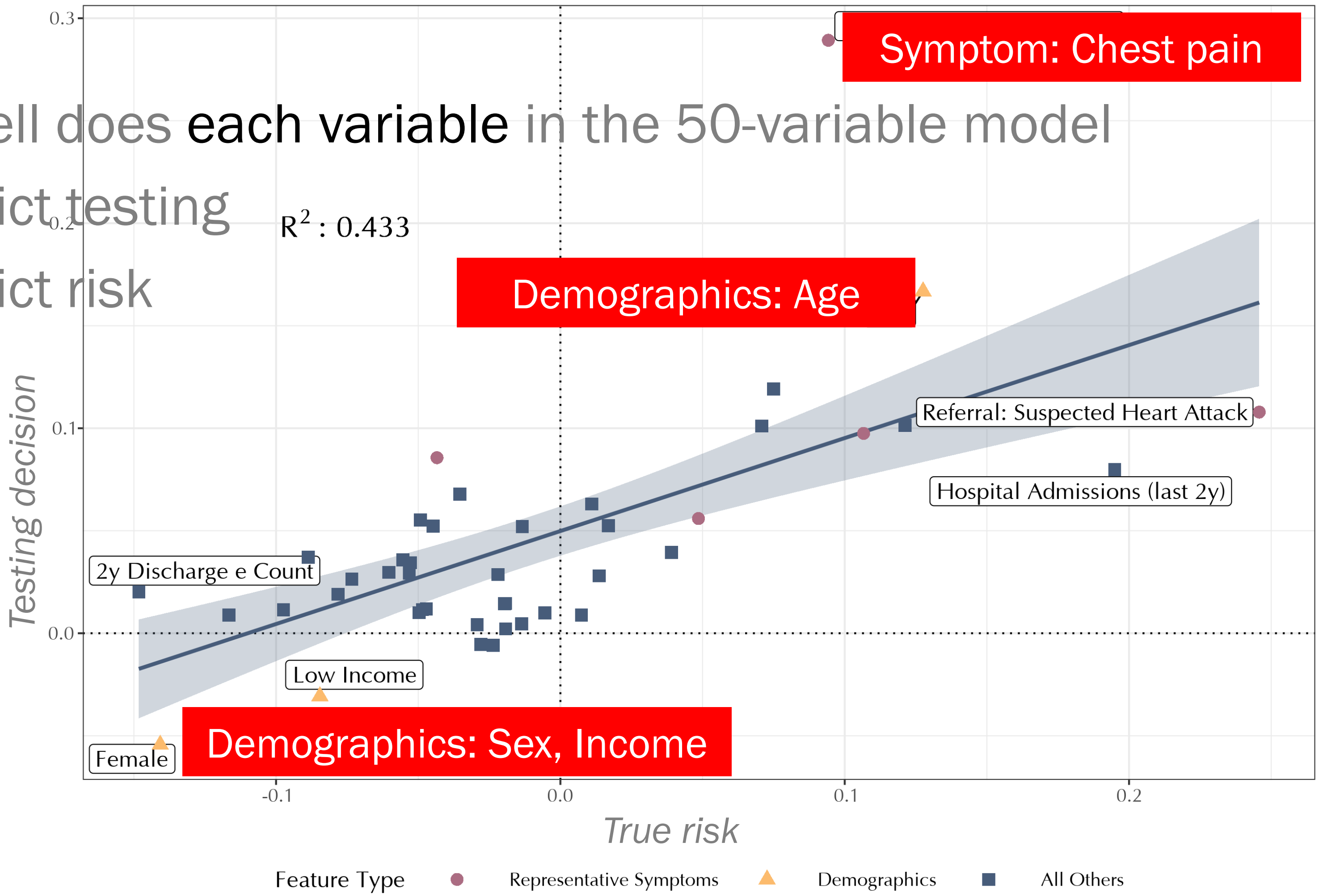
1. Bounded rationality
   - Physicians use too simple a model of risk

2. Systematic errors and biases
   - Physicians mis-weight specific variables

Physicians are ‘boundedly’ rational **and** systematically biased

1. Predict coronary blockage with **16,381 vs. 50 variables**
   – Which one looks more like the physician?

2. How well does **each variable** in the 50-variable model
   – Predict testing
   – Predict risk



Some variables are more salient than others

- Symptoms, demographics
  - The **first thing** we see about patients
  - A key part of vignettes, medical **education**
  - Very over-weighted: ACS symptoms
- Quantitative labs, vitals
  - Under-weighted

The NEW ENGLAND JOURNAL of MEDICINE

CASE RECORDS of the MASSACHUSETTS GENERAL HOSPITAL

CASE RECORDS OF THE MASSACHUSETTS GENERAL HOSPITAL
JUN 24, 2021

Case 19-2021: A 54-Year-Old Man with Irritability, Confusion, and Odd Behaviors

Kontos N., Parsons M.W., Biffi A., and González R.G. | N Engl J Med 2021; 384:2438-2445

A 54-year-old man was evaluated in the neuropsychology clinic because of irritability, confusion, and odd behaviors. Nine months earlier, he had been treated for cancer, after which chronic pain had developed. Six weeks before the current evaluation, he had been found unresponsive with medication bottles nearby. A diagnostic test was performed.

CME

PDF

CASE RECORDS OF THE MASSACHUSETTS GENERAL HOSPITAL
JUN 17, 2021

Case 18-2021: An 81-Year-Old Man with Cough, Fever, and Shortness of Breath

Hibbert K.A., Goiffon R.J., and Fogerty A.E. | N Engl J Med 2021; 384:2332-2340

An 81-year-old man presented with fever, cough, and shortness of breath. Within a few hours after presentation, chest pain and respiratory distress developed. A chest radiograph showed bilateral patchy airspace opacities, with predominance in the peripheral lower lung zone and with relative sparing of the perihilar region. A diagnostic test was performed.

FREE

PDF

CASE RECORDS OF THE MASSACHUSETTS GENERAL HOSPITAL
JUN 10, 2021

Case 17-2021: An 82-Year-Old Woman with Pain, Swelling, and Ecchymosis of the Left Arm

Finn K.M., Sutphin P.D., Carlson J.C.T., Raskin K.A., and Van Cott E.M. | N Engl J Med 2021; 384:2242-2250

An 82-year-old woman was admitted with pain, swelling, and discoloration of the left arm. CT revealed hematoma involving the brachioradialis muscle. The prothrombin time was 13.3 seconds (normal range, 11.5 to 14.5) and the activated partial-thromboplastin time 72.4 seconds (normal range, 22 to 36). A diagnostic test was performed.

CME

PDF

Summary

- Mis-prediction is a driver of both over- and under-use
  - Preferred estimate: keep 38% old tests… add 16% new
  - Not so much how much testing, but who is tested
- Many believe ML will transform health care
  - Most focus on ML as a product
  - e.g., hospital buys software to replace radiologists
- ML is also a powerful new tool for understanding
  - New inefficiencies, new models of physician behavior
- Paper at `ziadobermeyer.com/research`