Revised December 10, 2025

14.03/003 Microeconomic Theory & Public Policy Fall 2025

Lecture slides 23. Machine Predictions and Human Decisions

David Autor (Prof), MIT Economics and NBER

Prediction and decision — Two applications

Prediction and decision — Two applications

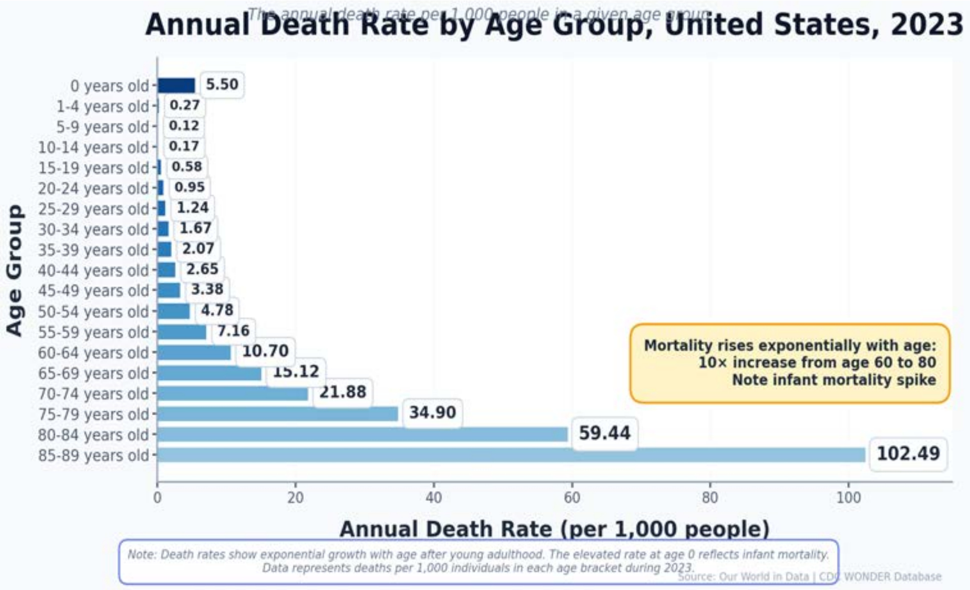
1. **Predicting the inevitable: It's harder than you think**

2. **Predicting medical need: What goes wrong, and how to fix it**

“Predictive modeling of U.S. health care spending in late life”

Einav, Finkelstein, Mullainathan, Obermeyer, *Science*, 2016

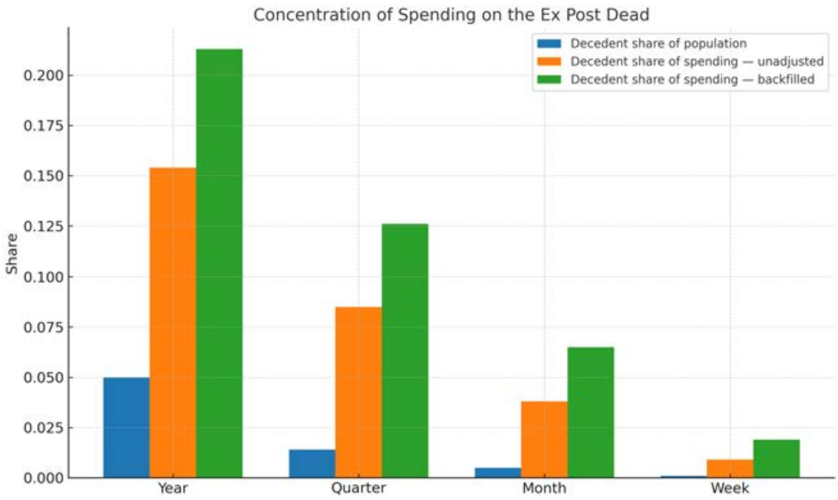
Annual mortality rates in the US in 2023 — per 1,000 people



Measuring medical costs of the deceased (in 2008)

1. Unadjusted – Starting from January 1, 2008, expenditure on those who die in that year
2. Backfilled – Starting from the date of death, measure spending over the prior year (*what is commonly reported in popular discussion*)
3. In both cases, the denominator is total Medicare beneficiary spending in the year (2008)

Medical spending on the 'ex post deceased'



Einav, Finkelstein, Mullainathan, Obermeyer, *Science*, 2016

Interpreting the medical costs of the those who have died

What factors should determine the ‘right’ level of spending on the dying?

Many considerations

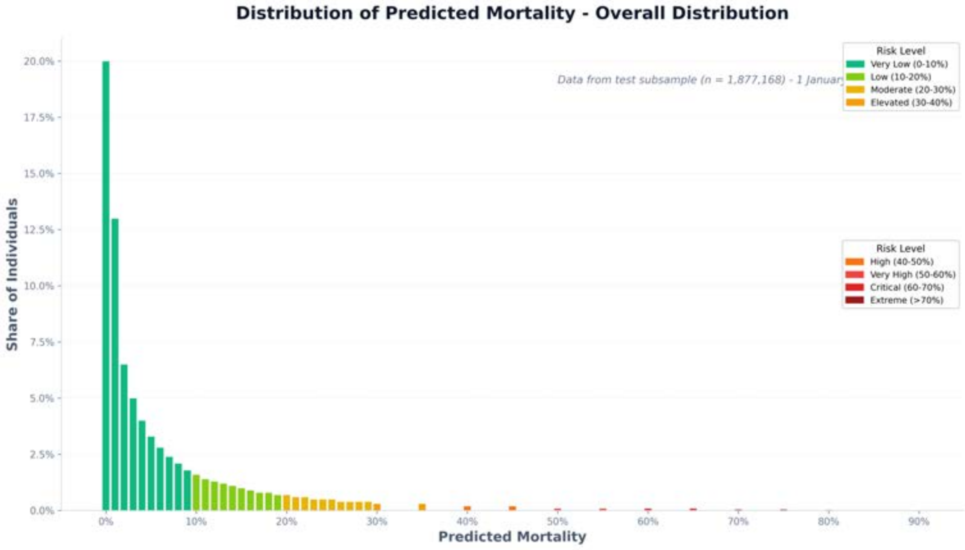
- How much does spending comfort the dying?
- How much does spending comfort the living?
- What is the marginal effectiveness of medical care at preventing death?
- How accurately can we predict who will live and who will die?

Interpreting the medical costs of those who have died

Those who end up dying are not the same as those who are *relatively likely* to die

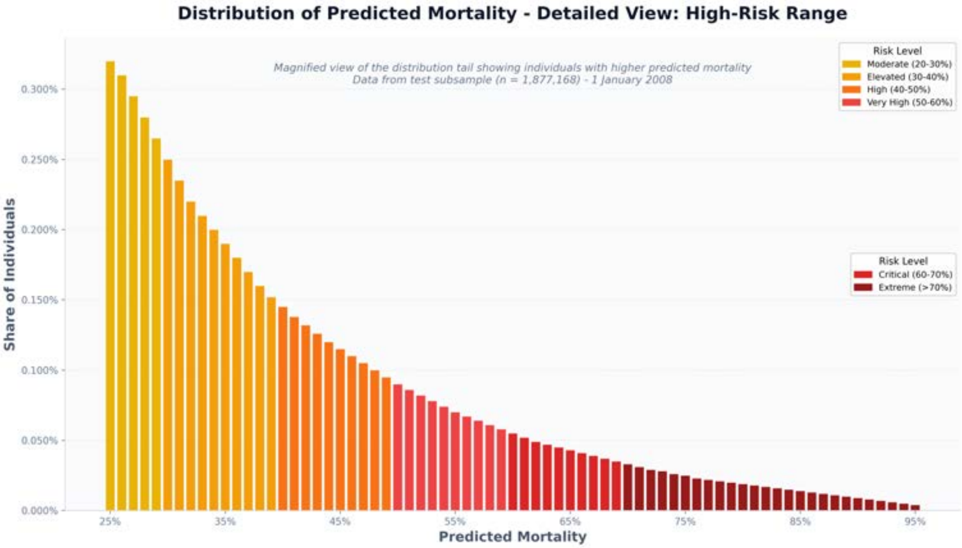
- We spend more on sicker patients
- We could end up spending a large fraction of $$ on patients in their final year of life, *even if we could not predict who their deaths at the start of the year*

Distribution of predicted one-year mortality on January 1, 2008



Einav, Finkelstein, Mullainathan, Obermeyer, *Science*, 2016

Distribution of predicted one-year mortality on January 1, 2008



Einav, Finkelstein, Mullainathan, Obermeyer, *Science*, 2016

Death is surprisingly hard to predict

1. There is no sizable mass of people for whom death is certain (or even near certain) within the year
2. The 95th percentile of predicted annual mortality is only about 25%.
3. Less than 10% of those who end up dying within the year have an annual mortality probability above 50%

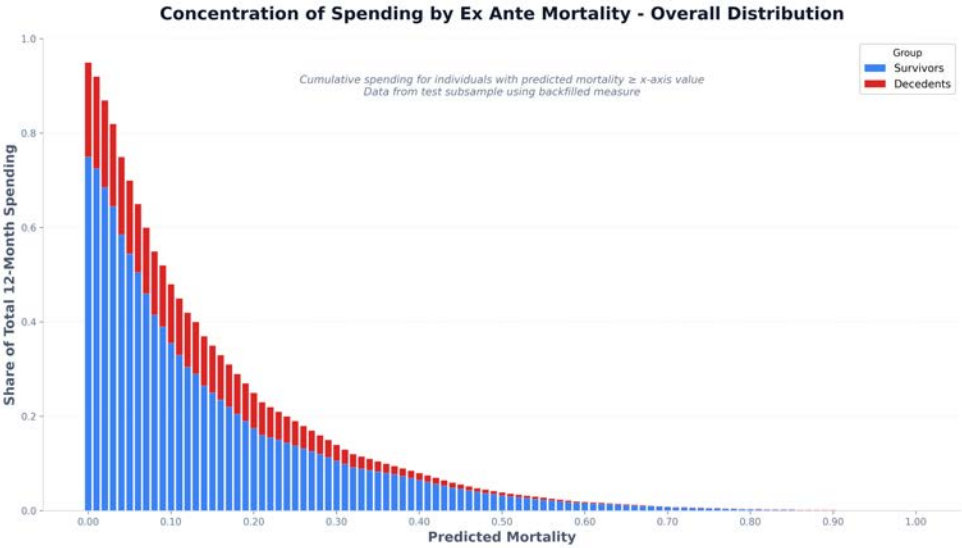
Less than 10% of those who end up dying within the year have an annual mortality probability above 50%. Why is this?

- **Note these *approximate* statistics**
  - Annual mortality among Medicare population is $\Pr[D] = 0.05$
  - $\Pr[D > 0.50] = 0.01$
  - $\Pr[D <= 0.50] = 0.99$
  - $E[D | \Pr(D > .50)] = 0.51$
  - $E[D | \Pr(D \leq .50)] = 0.045$
- **Applying Bayes' Rule**

$$\Pr(A|B) = \frac{\Pr(B|A) \times \Pr(A)}{(\Pr(B|A) \times \Pr(A)) + (\Pr(B|\neg A) \times \Pr(\neg A))}$$

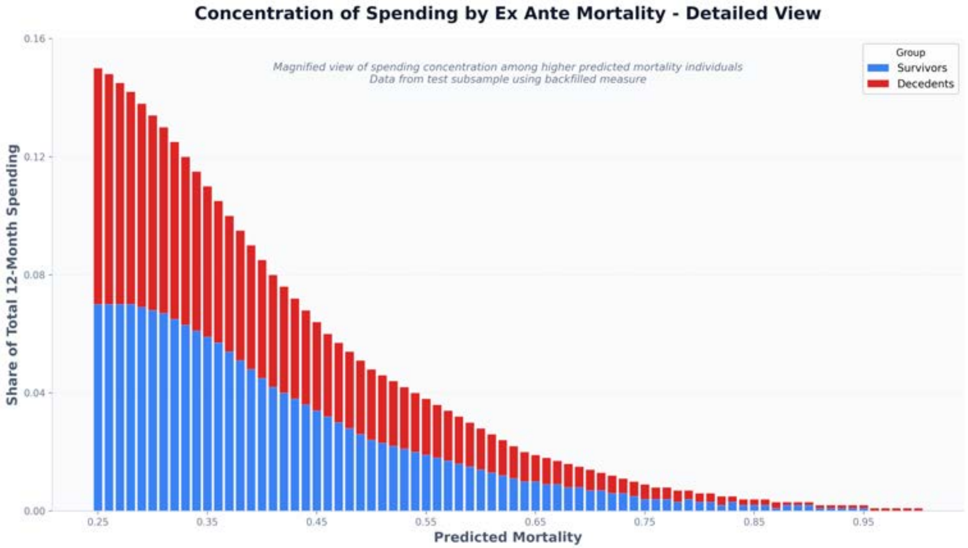
$$\Pr[D \geq 0.50 | D] = \frac{0.51 \times 0.01}{0.51 \times 0.01 + 0.045 \times 0.99} = 0.103 = 10.3\%$$

Spending shares of survivors and decedents by predicted mortality



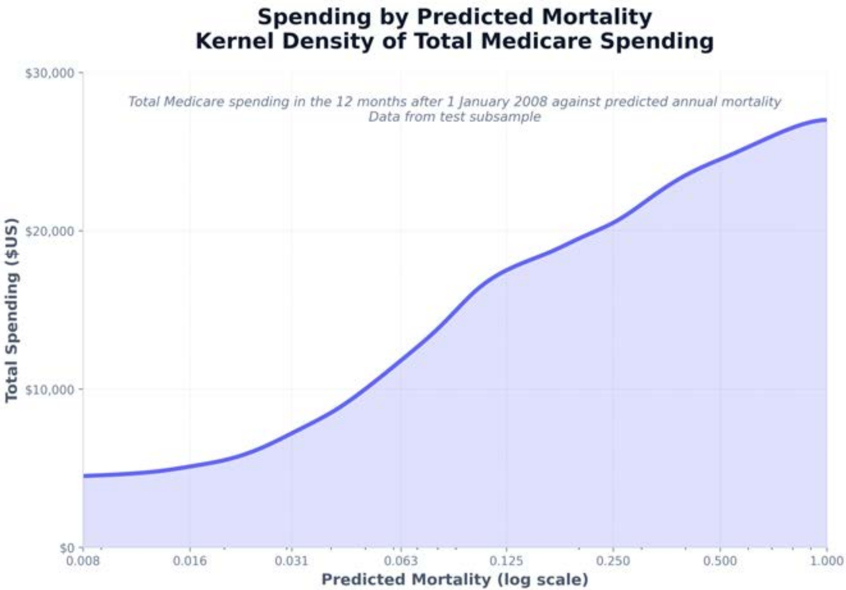
Einav, Finkelstein, Mullainathan, Obermeyer, *Science*, 2016

Spending shares of survivors and decedents by predicted mortality



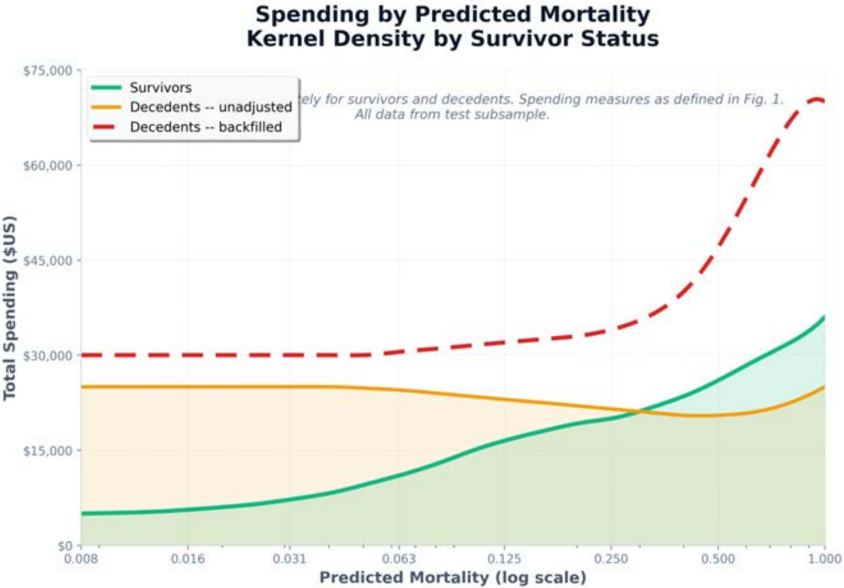
Einav, Finkelstein, Mullainathan, Obermeyer, *Science*, 2016

Spending by predicted mortality – summing survivors + decadents



Einav, Finkelstein, Mullainathan, Obermeyer, *Science* 2016

Spending by predicted mortality – split by survival status



Einav, Finkelstein, Mullainathan, Obermeyer, *Science* 2016

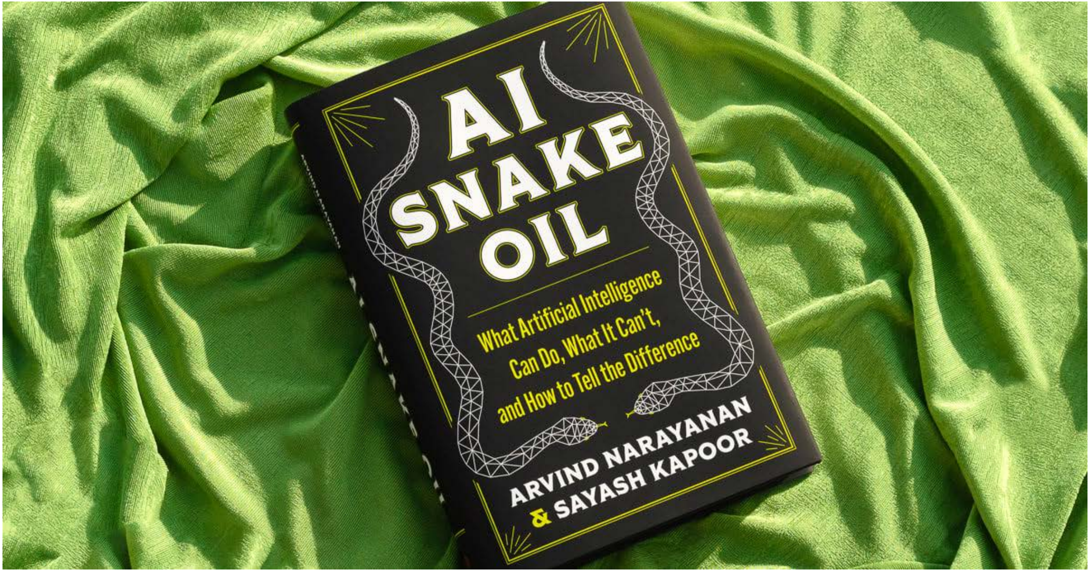
Death is surprisingly hard to predict

Individuals with high predicted mortality account for only a small share of total spending

- The highest-risk percentile – those with predicted mortality above 46% – accounts for under 5% of total spending
- Consistent with prediction, 45% of the highest risk percentile are **survivors**
- To identify a group of decedents who account for at least 5% of total spending, must set a threshold of predicted mortality of 39% or higher
- (Results are based on the backfilled measure of decedent spending. When using the unadjusted measure, spending on decedents is even *lower*)

"...for most people, death comes only after long medical struggle with an incurable condition—advanced cancer, progressive organ failure..., or the multiple debilities of very old age. In all such cases, death is certain, but the timing isn't"

Atul Gwande, *New Yorker*, 2 Aug 2020


AI SNAKE OIL
What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference
ARVIND NARAYANAN & SAYASH KAPOOR

Intention, prediction, and self-fulfilling prophecy

Machines as decision-makers: Will they learn and replicate our biases?

1. Job search ads for highly paid positions are less likely to be presented to women
2. Image searches for professions such as CEO produce fewer images of women
3. Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and minorities
4. Natural language processing algorithms encode language in gendered ways

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

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

Algorithms seek to identify patients who will benefit from treatment

1. Key assumption
   - Those with the *greatest care* needs will benefit the most from the program
2. Under this assumption
   - The targeting problem becomes a *pure prediction problem*
3. Developers build algorithms by training on history
   - Use past data to build a *predictor* of future health care needs

Machines as decision-makers: Will they learn and replicate our biases?”

Empirical investigations of algorithmic bias often hindered by opaqueness

- Algorithms deployed at scale are typically proprietary, making it difficult for independent researchers to dissect them
- Researchers often work “from the outside,” often using workarounds like audit studies
- Can document disparities, but understanding how and why they arise is difficult without greater access to the algorithms themselves

This study

Dataset describes one such health management algorithm

- Contains algorithm's predictions
- Also contains data needed to understand its inner workings—the underlying ingredients used to form the algorithm (data, objective function, etc.)
- Plus links to a rich set of outcome data
- Provide a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise

Descriptive Statistics by Race

Demographics, Care Management & Utilization

	White	Black
n (patient-years)	88,080	11,929
n (patients)	43,539	6,079
Demographics		
Age	51.3	48.6
Female (%)	62	69
Care Management Program		
Algorithm score (percentile)	50	52
Race composition of program (%)	81.8	18.2
Care Utilization		
Actual cost	$7,540	$8,442
Hospitalizations	0.09	0.13
Hospital days	0.50	0.78
Emergency visits	0.19	0.35
Outpatient visits	4.94	4.31

Source: Obermeyer et al., Science 366, 447-453 (2019) | BP = blood pressure; LDL = low-density lipoprotein

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

Descriptive Statistics by Race

Mean Biomarkers & Active Chronic Illnesses

	White	Black
Mean Biomarker Values		
HbA1c (%)	5.9	6.4
Systolic BP (mmHg)	126.6	130.3
Diastolic BP (mmHg)	75.5	75.7
Creatinine (mg/dl)	0.89	0.98
Hematocrit (%)	40.7	37.8
LDL (mg/dl)	103.4	103.0
Active Chronic Illnesses (Comorbidities)		
Total number of active illnesses	1.20	1.90
Hypertension	0.29	0.44
Diabetes, uncomplicated	0.08	0.22
Obesity	0.07	0.18
Pulmonary disease	0.07	0.11
Anemia	0.05	0.10
Renal failure	0.03	0.07
Diabetes, complicated	0.02	0.07

Source: Obermeyer et al., Science 366, 447-453 (2019) | Key comorbidities shown (full list includes 20+ conditions)

Research sample

- Algorithmic risk score generated for each patient-year for $100,009$ patients during the enrollment period for the system's care management program
- Patients above the 97^{th} percentile are automatically identified for enrollment in the program
- Those above the 55^{th} percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment

Research sample

- Algorithm takes raw insurance claims data $X_{i,t1}$ (features) over the year $t-1$
- Demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs
  - 71.2% enrolled in commercial insurance
  - 28.8% enrolled in Medicare
  - Average age 50.9 yrs
  - 63% female
- **The algorithm specifically excludes race**

At each pctile of risk score, Black patients have **more** chronic conditions

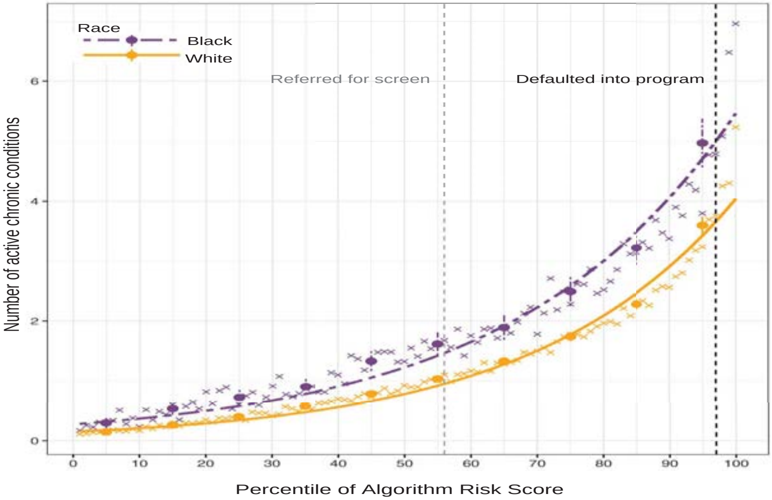


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

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

At each pctile, Black patients have **more** hypertension and diabetes

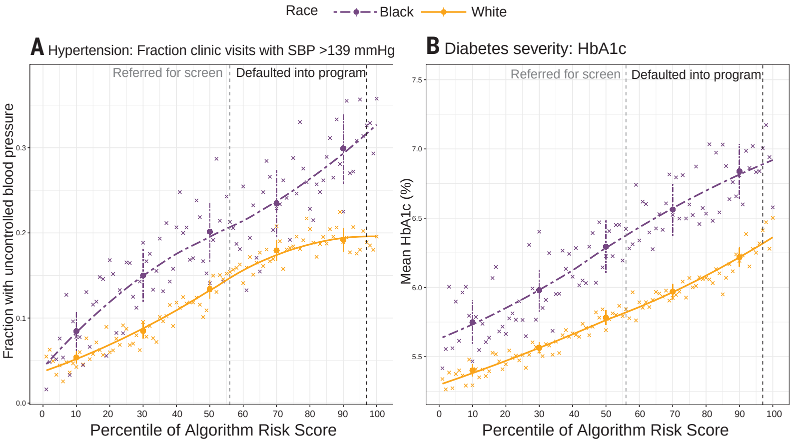


Fig. 2. Biomarkers of health versus algorithm-predicted risk, by race. (**A** to **E**) Racial differences in a range of biological measures of disease severity, conditional on algorithm risk score, for the most common diseases in the population studied. The × symbols show risk percentiles by race, except in (C) where they show risk ventiles; circles show risk quintiles with 95% confidence intervals clustered by patient. The *y* axis in (D) has been trimmed for readability, so the highest percentiles of values for Black patients are not shown. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile).

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

Black and white patients w/same risk scores have ≈ identical med expenditures

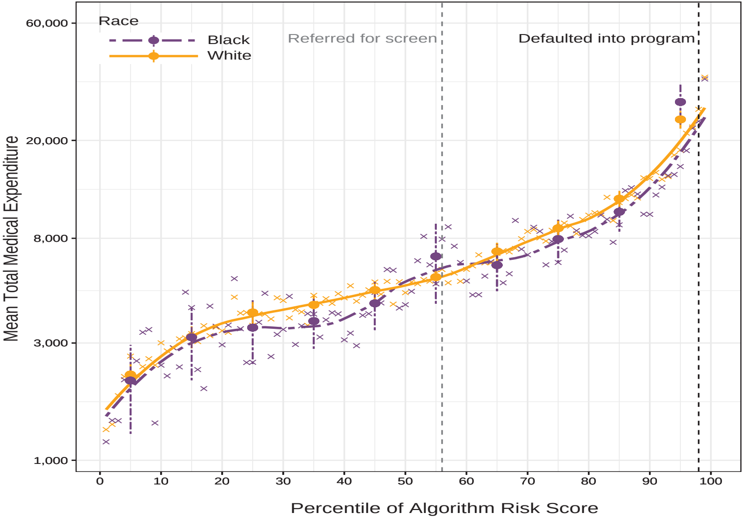


Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. (**A**) Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). (**B**) Total medical expenditures by race, conditional on number of chronic conditions. The × symbols show risk percentiles; circles show risk deciles with 95% confidence intervals

What could be going wrong?

- Race is not included in calculation of risk score
- The system is not *deliberately* race discriminatory
- The risk score Y is ‘calibrated’

$$E[Y|X, R = B] = E[Y|X, R = W].$$

- Black and white patients with identical medical histories receive identical risk scores, Y
- Hint: What do you suppose Y is?

Question — What is going wrong?

With same chronic illness score, Black patients have much lower med expenses

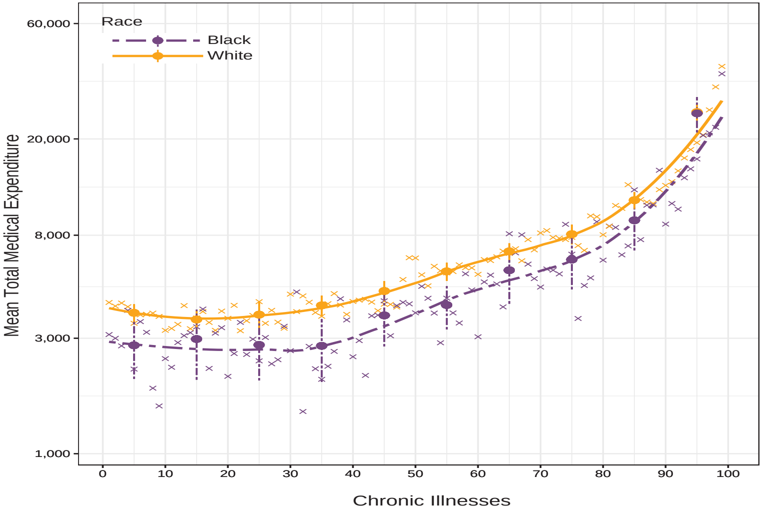


Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. (**A**) Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). (**B**) Total medical expenditures by race, conditional on number of chronic conditions. The × symbols show risk percentiles; circles show risk deciles with 95% confidence intervals

Label bias

- We want a measure of health risk, Y^*
- But this is not an observable variable, so we choose a proxy (AKA, label) $\tilde{Y}$
- This label might be correct on average

$$E[Y^*] = \quad E[\tilde{Y}]$$

- Yet, the label may unintentionally replicate human biases encoded in the data

$$E[\tilde{Y} - Y^* | R = B] \quad < 0 : \text{ Black patients systematically under-treated}$$
$$E[\tilde{Y} - Y^* | R = W] \quad > 0 : \text{ White patients systematically over-treated}$$

- Can happen if race is predictable from X, even if race not in data: $W, B \not\perp X$
- Marzyeh's lecture underscored that this is *commonplace*

Label bias can reinforce existing disparities, create self-fulfilling prophesies

1. High school performance measures (e.g., AP scores, college admissions) mostly reflect affluence of the kids who attend the school, not school's value-added
   - Causes more kids from affluent families to attend
2. Credit-scoring algorithms predict outcomes related to income, thus incorporating disparities in employment and salary
   - Low-income adults have more trouble getting loans, pay higher interest rates, develop worse credit
3. Policing algorithms predict *measured crime*, which also reflects increased scrutiny of some groups
   - Groups that are policed more aggressively will develop records, get further scrutiny
4. Hiring algorithms predict supervisory ratings, which codify existing stereotypes, preferences
   - Prevent qualified people from getting hired, inhibiting career advancement

Fraction Black patients by risk score—risk score and corrected for health

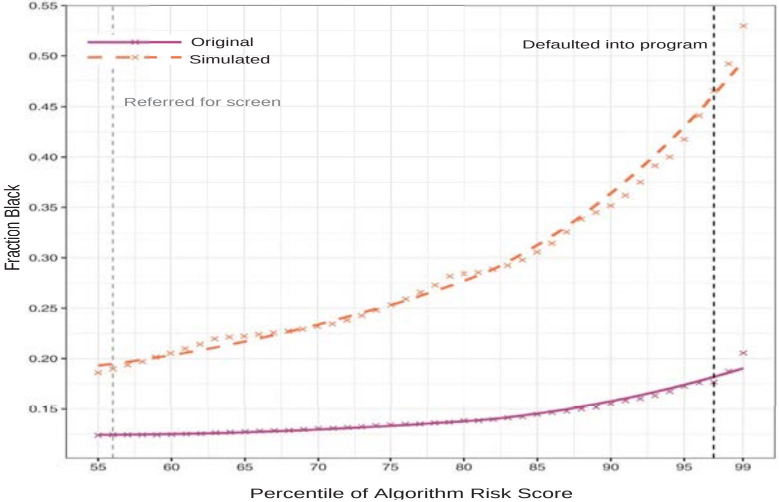


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

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

Epilogue (p1 of 3)

- We contacted the algorithm manufacturer for an initial discussion of our results
- The manufacturer independently replicated our analyses on its national dataset of $3,695,943$ commercially insured patients
- This confirmed our results
- Calculated in their dataset, Black patients had $48,772$ more active chronic conditions than White patients, conditional on risk score

Epilogue (p2 of 3)

- To resolve the issue, we began to experiment with solutions together
- We suggested using the existing model infrastructure— sample, predictors (excluding race, as before), training process, and so forth—but changing the label...
- Rather than future cost, we created an index variable that combined health prediction with cost prediction
- This reduced the number of excess active chronic conditions in Black patients, conditional on risk score, to $7,758$, an 84% reduction in bias.

Epilogue p3 of 3)

- We are establishing an ongoing (unpaid) collaboration...
- To convert the results into a better, scaled predictor of multi-dimensional health measures
- Goal: Rolling these improvements out in a future round of algorithm development