science.sciencemag.org/content/366/6464/447/suppl/DC1

Supplementary Materials for

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

Ziad Obermeyer*, Brian Powers, Christine Vogeli, Sendhil Mullainathan*†

*These authors contributed equally to this work.
†Corresponding author. Email: sendhil.mullainathan@chicagobooth.edu

Published 25 October 2019, *Science* **366**, 447 (2019)
DOI: 10.1126/science.aax2342

This PDF file includes:

Materials and Methods

The algorithm in context

The Affordable Care Act (Obamacare) created substantial pressure on hospitals and health systems to reduce health care costs. To do so, many systems entered into new contracts with health insurers in which they assume financial risk for the quality and cost of the care they provide. Concretely, these hospitals are either paid a flat annual fee for each patient (in contrast to the traditional "fee-for-service" model where the more care hospitals provided, the more they were paid), or receive end-of-year monetary adjustments relative to negotiated per-patient cost targets. The goal of these models is to align the incentives of hospitals with the incentives of society, around reducing costs.

These new financial incentives led to operational changes at many hospitals and health systems. Chief among them was the development and implementation of high-risk care management programs (*14*). These programs aim to provide additional resources to medically complex patients, before their health deteriorates. The theory is that better, earlier access to care for these high-risk patients will prevent burdensome and costly complications like emergency visits and hospitalizations, thereby achieving both higher quality and lower cost.

High-risk care management programs, which are staffed by skilled health care providers caring for small groups of patients, are costly to operate. This necessitates precise targeting of patients likely to benefit from the program. Hospitals thus frequently turn to commercial algorithms to predict which patients will have the most significant and complex health care needs, in order to select them for entry into care coordination programs (*18*, *19*).

We reproduce in the table below data from the Society of Actuaries, which conducted a comprehensive evaluation of the ten most widely used algorithms, developed and used by non-profit hospitals, academic groups, and governmental agencies, including the particular algorithm we study (6). The enthusiasm for cost prediction is not restricted to our particular algorithm: the object of interest was cost prediction or resource utilization (*21*). This approach is likewise described in academic literature on targeting population health interventions (*18*, *19*).

Manufacturer	*Model*	*Notes on algorithm label used for prediction*
Johns Hopkins University	ACG System	"concurrent and prospective cost models measure the morbidity burden of patient populations"
University of California at San Diego	Chronic Illness & Disability Payment System, MedicaidRx	"classification system for Medicaid programs to use to make health-based capitated payments"
3M Health Information Systems	Clinical Risk Groups	"relates the historical clinical and demographic characteristics of the enrollee… to the amount and type of healthcare resource that enrollee will consume"
Verisk Health	DxCG Intelligence	"Scores correlate with the cost of the underlying illness

		burden that individuals carry”
Centers for Medicare and Medicaid Services	HHS-HCC Model	“One unique aspect of the HHS-HCC model is that the model does not predict allowed costs, but rather predicts plan liability at each of the five ACA metal levels”
Optum	ImpactPro	“Uses a member's clinical episodes of care, prior use of health care services, prescription drugs, and lab results as markers of their future health care use”
Milliman	Advanced Risk Adjusters	“uses demographic and claim data in conjunction with its library of risk adjusters to estimate morbidity and healthcare resource use”
SCIO Health Analytics	Prospective Cost of Care Model	“The model aims at predicting the total costs and financial risk per member”
Truven Health, an IBM Company	Cost of Care Model	“estimates both retrospective and future expected healthcare payments”
Wakely Consulting Group	Risk Assessment Model	“anticipate what the HHS-HCC model may look like” (refers to the Centers for Medicare and Medicaid Services model above)

Algorithm implementation in the health system we study

At the health system we study, the algorithm is given a data frame with two elements.

1. C_{it} (label): Total medical expenditures (which for simplicity we denote “costs”) in year *t*
2. $X_{i,t-1}$(features): For the commercially insured sample, we observe the raw insurance claims data that form the totality of inputs to the predictive algorithm (though we do not observe how these raw data are combined to form the specific variables used for prediction). These data are records of care utilized and billed to the patient’s insurer over the year *t-1*:
    a. Demographics (e.g., age and sex, but specifically excluding race),
    b. Insurance type,
    c. ICD-9 diagnosis and procedure codes,
    d. Prescribed medications,
    e. Encounters, categorized by type of service (e.g., surgical, radiology, etc.),
    f. Billed amounts, categorized by type (e.g., outpatient specialists, dialysis, etc.).

A programmer collects these data for all eligible patients (those enrolled in risk-based insurance contracts) for a given year *t-1*, and feeds them into the commercial software, which delivers back a risk score for year *t*. The algorithm’s stated goal (from promotional materials) is to predict *which individuals are in need of specialized intervention programs and which intervention programs have the most impact on the quality of individuals’ health*. These scores, which are meant to *flag individuals for intervention before their health becomes catastrophic*, are a key part of the decision to enroll a patient in the care management program (which is described in more detail below). The algorithm is run three times per year, during the enrollment period for the program. Patients whose scores exceed a critical threshold, approximately the 97th percentile in our data, are auto-identified for enrollment in the program at the health system we study (though this does not guarantee that they will be enrolled: they may not qualify based on other criteria, which are not available in administrative data but become clear to program staff during attempted enrollment). Those whose scores exceed a lower threshold, the 55th percentile, are referred to their primary care physicians, who are provided with additional metrics about the patients and prompted to consider whether they would benefit from enrollment.

Study outcomes

As we described in a previous version of the current analysis, submitted to the non-archival track of a computer science conference,(*46*) we study several measures of health H with respect to the algorithmic risk score. To construct H, we draw on a rich dataset of EHRs linked to algorithmic predictions. We first construct a summary measure of health status, the total number of chronic illnesses for which the patient had a medical encounter over year t. This approach is used extensively in medical research (*24*) to provide a comprehensive view of a patient's health (*25*). In addition, the number of active chronic illnesses is thought to be a measure of medical complexity that correlates with the treatment effect of care management programs (*18*).

While a tally of chronic illnesses is a reasonable summary measure, more biological measures derived from EHRs can provide a more granular sense of health status. To generate these, we first identify the individual chronic illnesses that contribute to comorbidity score. As shown in Table 1, hypertension and diabetes are the most common illnesses in our sample, at 30% and 14% overall prevalence respectively.

Our goal was to construct biomarker-based measures of severity for as many of these illnesses as possible. This was meant to measure not just the presence or absence of these illnesses, but the degree to which they are well managed: with good adherence to medication regimens, and timely access to primary care for adjustment of therapy, these biomarkers should theoretically be optimized and in the normal range. However, some patients—because of medical complexity, or because of non-adherence to medication regimen—do not achieve adequate control of these illnesses, leading to catastrophic complications: atherosclerotic cardiovascular disease (ASCVD, e.g., heart attack and stroke, which for acute events are also quite common in our sample), limb amputations, and need for life-long dialysis. These are exactly the patients who are thought to benefit most from care coordination to prevent these events: one of the major goals of care management programs is to control these illnesses, which are some of the largest drivers of health needs in primary care populations (*18*, *26*).

For several of these common illnesses, biomarker-based measures of control are available in EHR data: for chronic hypertension we use the fraction of all outpatient (i.e., clinic) measurements in year t where systolic blood pressure is elevated (above 139 mmHg). For diabetes, we use a laboratory study, mean hemoglobin A1C (HbA1c), which is elevated in the setting of uncontrolled high blood sugar. For renal failure, we use mean creatinine clearance rate, which measures the ability of the kidneys to filter blood. For ASCVD, we use the mean low-density lipoprotein (LDL) or "bad" cholesterol. Finally, we use another biomarker as an integrative measure of the burden of chronic illnesses: anemia, a deficiency of hemoglobin that results in a decreased ability to carry oxygen in the blood, has many causes, but in a population of older, sicker patients it is often used as a measure of the physiological burden of chronic illnesses. This anemia, known as the "anemia of chronic disease," (*47*) is a well-known phenomenon in the setting of chronic illness, irrespective of its exact nature. We measure this by mean hematocrit, the fraction of blood volume made up by red blood cells.

Of note, these EHR data are not routinely analyzable, as they must be pulled and cleaned extensively from hospital data warehouses. As a result, algorithm developers typically do not have access to them to fit or validate their predictions, making this exercise particularly useful to assess algorithm performance in general.

In these biomarkers, we can then compute the differences between Black and White patients, conditional on risk score. Considering outcomes in the highest-risk patients (at the 97th percentile of risk score), Blacks have:

- More severe hypertension (systolic blood pressure: 134.3 mmHg vs. 128.6 mmHg for Whites, $P<0.001$). To scale this difference, a 5.7 mm Hg increase translates into a 11.9% higher risk of heart attack and a 7.6% lower all-cause mortality rate (using estimates from a meta-analysis of clinical trials of blood pressure reduction) (*27*).
- More severe diabetes (HbA1c: 7.0% vs. 6.4% for Whites, $P<0.001$). For reference, every 1% absolute increase in HbA1c correlates with a 30% increase in all-cause mortality and a 40% increase in cardiovascular mortality, among individuals with diabetes (*28*).
- More severe renal failure (creatinine: 1.38 mg/dL vs. 1.04 mg/dL, $P<0.001$). This means that marginal Black patients already met the definition of reduced kidney function, at which cardiovascular and all-cause mortality begin to increase rapidly (*48*).
- Worse anemia (hematocrit: 36.5% vs. 38.7% for Whites, $P<0.001$). Hematocrit declines with age by 0.155% per year, (*49*) meaning that this difference is equivalent to Black patients being nearly 14 years older than Whites in terms of this biomarker.
- Worse cholesterol, in the higher parts of the risk distribution, although the difference did not reach statistical significance (LDL: 94.9 vs. 90.2 mg/dL for Whites, $P=0.26$). This level of difference would translate into a 3.3% higher risk of major cardiovascular events (using effect size derived from trials of cholesterol-lowering therapy with statins) (*50*).

Additional robustness checks related to program effect

We use a range of measures H to assess bias in realized health in a given year t: are Black patients sicker than Whites at a given level of risk. Of course, the care management program D is allocated as a function of algorithm score, and could affect these measures of health. This could pose a problem: if enrollment affected health similarly for Blacks and Whites, we could still estimate the extent of bias consistently, but differential program effects by race would induce bias. As a result, we perform several experiments to determine whether it is reasonable to abstract from any program effect of this kind for the analysis described in the main text.

We test for this in several ways. First, we use health scores $H_{(t-1)}$ instead of H_t, before the program is allocated. These measures are correlated with scores in year t but of course cannot be affected by the program. We find similar patterns of calibration in number of chronic conditions in year *t-1* (Figure S2) as we saw in *t*.

Second, we compare health for enrolled vs. unenrolled patients in *t*, i.e., (H_t | D=1) vs. (H_t | D=0), and likewise find the same patterns, analogous to an "as-treated" analysis (Figure S3). We find no evidence that the program affects biomarkers for Whites differently from Blacks (e.g., it would have been problematic if we found that biomarkers were more improved for Whites than Blacks among those ultimately enrolled in the program).

Third, we test for kinks around the thresholds of screening and auto-identification for the program, and find no evidence of any difference in calibration by program effect, analogous to regression discontinuity. There too we find no evidence of a program effect on biomarkers that differs by race, either at the auto-identification threshold (Figure S4), or at the screening threshold (Figure S5).

Training the experimental algorithms

Our three new predictive algorithms are trained to predict the following outcomes:

1. *Total cost* in year *t*. This functions to tailor cost predictions to our own dataset rather than the national training set used by the algorithm manufacturer.
2. *Avoidable cost* in year *t*, due to emergency visits and hospitalizations.
3. *Health* in year *t*, as measured by the number of active chronic conditions. Of note, this is a measure of how many chronic conditions are flaring up and driving utilization, not simply an indicator of previously diagnosed chronic conditions (for which predictions are not necessarily required).

We train all models as follows. We begin by randomly dividing all patient-years into a ⅔ training set and a ⅓ holdout set (at the patient level, i.e., no patient can appear in both sets). For each observation, we generate 149 features using electronic health record data and insurance claims from year *t-1*. These include demographics, indicators for active chronic conditions, costs including total costs and subcategory breakdowns, and biomarkers (related to chronic diseases, as described above, as indicators: normal, low, high). More detail on the features can be found in the synthetic dataset; all summary statistics match those in the original dataset. As with the original algorithm, we exclude race from the feature set; we show in the appendix that models perform similarly when race is included (Table S3), and that predictions are correlated with race but do not substantially reconstruct it (Table S4). Using these features, we train an L1-regularized regression (lasso), with the regularization penalty tuned via tenfold cross validation in the training set, and show results from the holdout set only.

Ethical approval and synthetic dataset

The Institutional Review Board of Partners HealthCare approved this study, and judged that patient consent was not required because the use of routinely collected data posed limited risk. Given that the study data are difficult to deidentify, we are unable to provide the data used in this study. We do make available all code to reproduce the results, and provide a simulated dataset (using synthpop (*51*)) along with detailed data descriptions at https://gitlab.com/labsysmed/dissecting-bias, so that others can replicate our analyses.

Figures and Tables

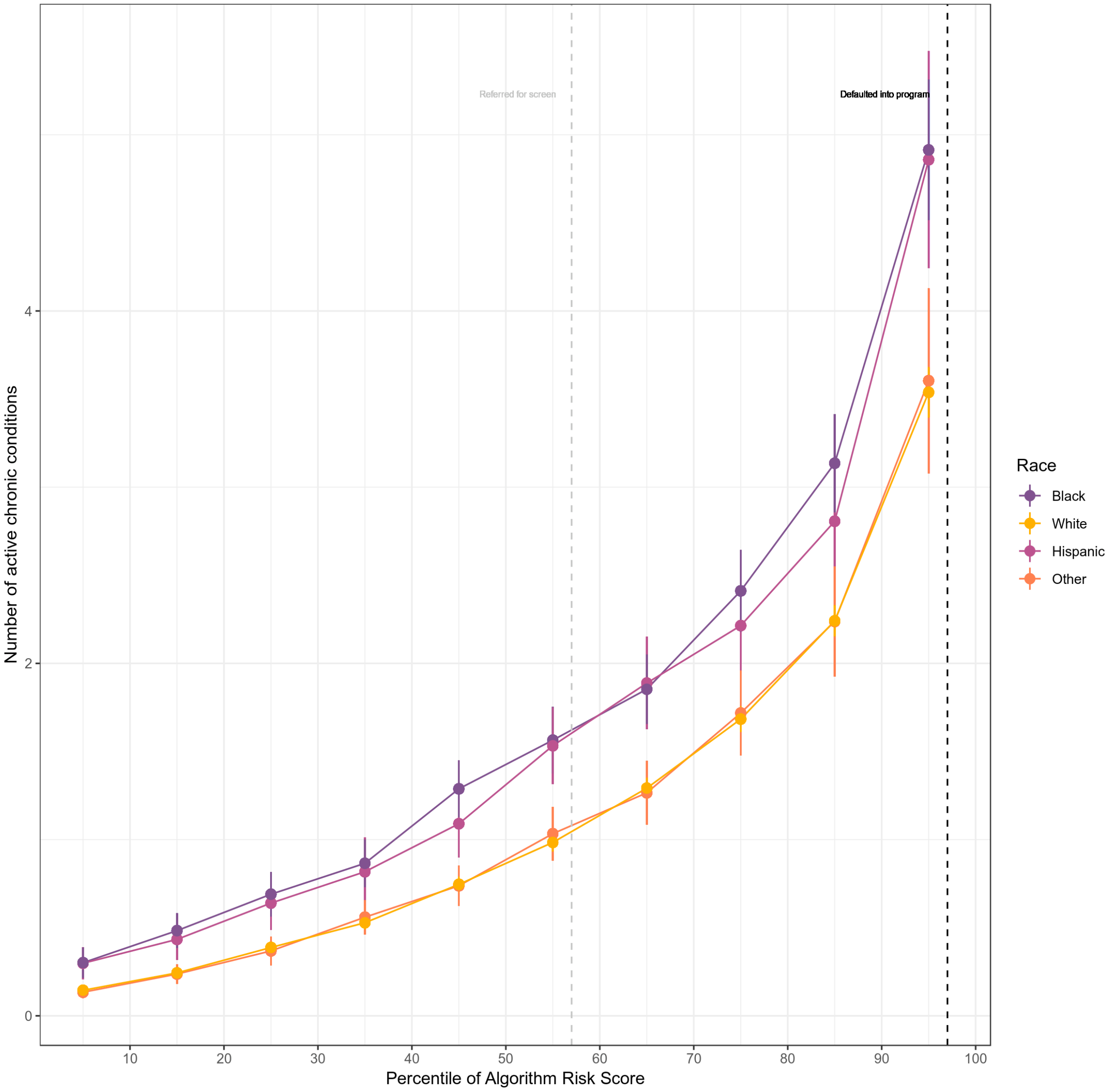


Fig. S1. Number of chronic illnesses vs. algorithm-predicted risk, including non-Black, non-White patients. Mean number of chronic conditions by race conditional on algorithm risk score.

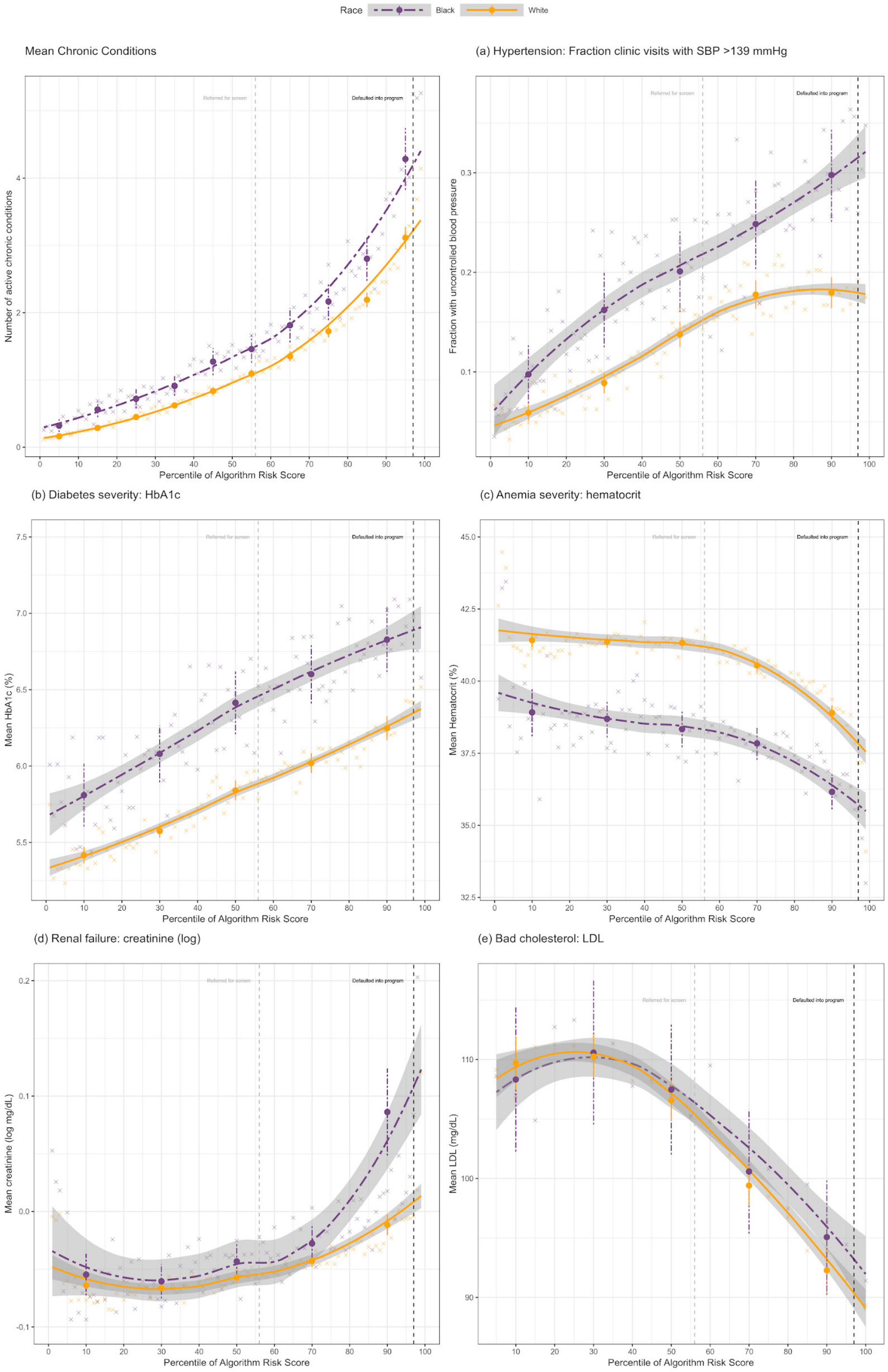


Fig. S2. Health measures in year t-1 vs. algorithm-predicted risk. Racial differences in a range of health measures, conditional on algorithm risk score, for number of chronic illnesses and biomarkers measuring severity of the most common diseases in the population studied. The x's show risk percentiles; points show risk quintiles with 95% confidence intervals clustered by patient.

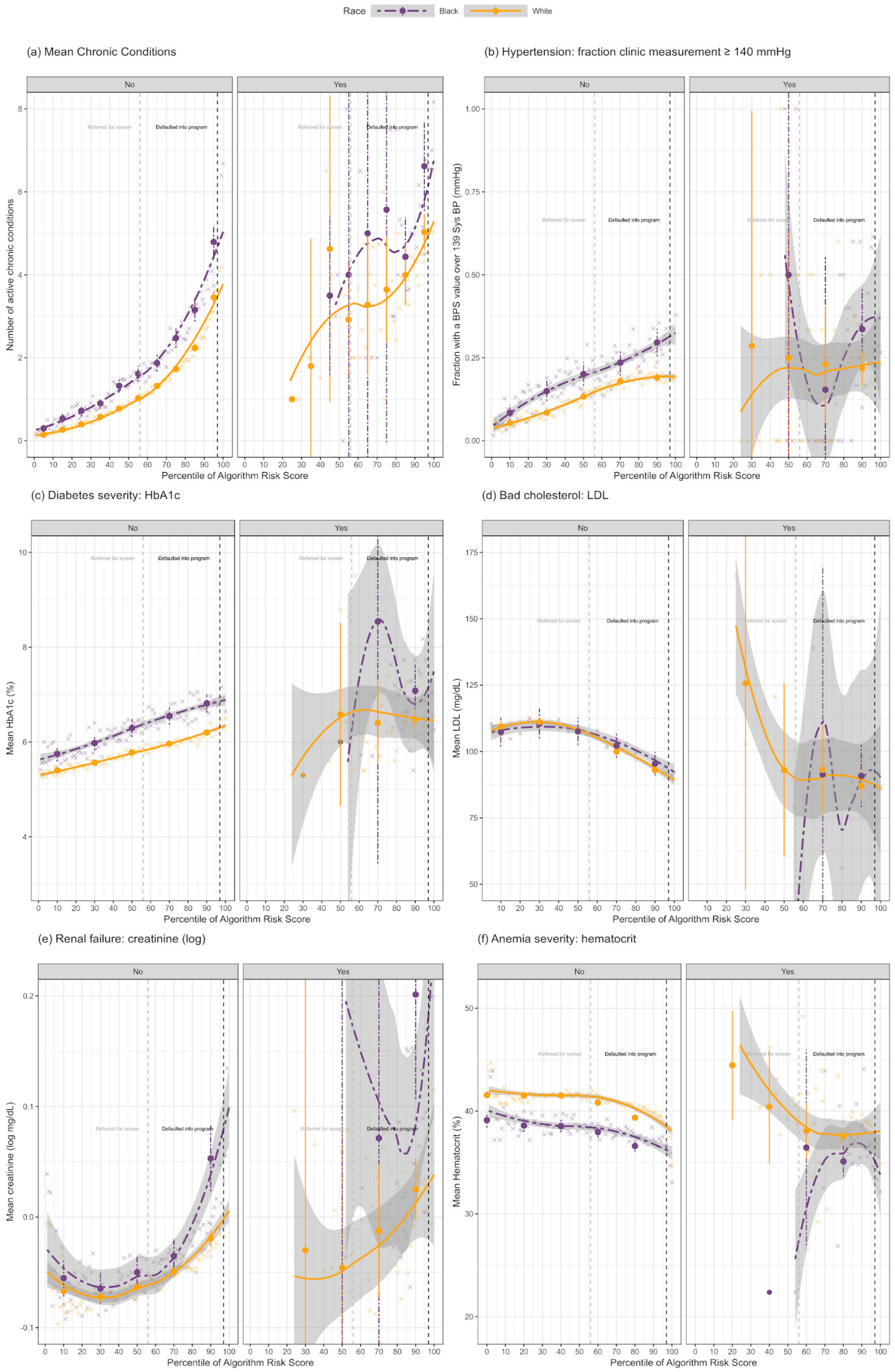


Fig. S3. Health measures vs. algorithm-predicted risk, by program enrollment. Racial differences in a range of health measures, comparing those enrolled vs. not enrolled in the care management program, conditional on algorithm risk score, for total number of chronic illnesses and biomarkers measuring severity of the most common diseases in the population studied. The x's show risk percentiles; points show risk quintiles with 95% confidence intervals clustered by patient.

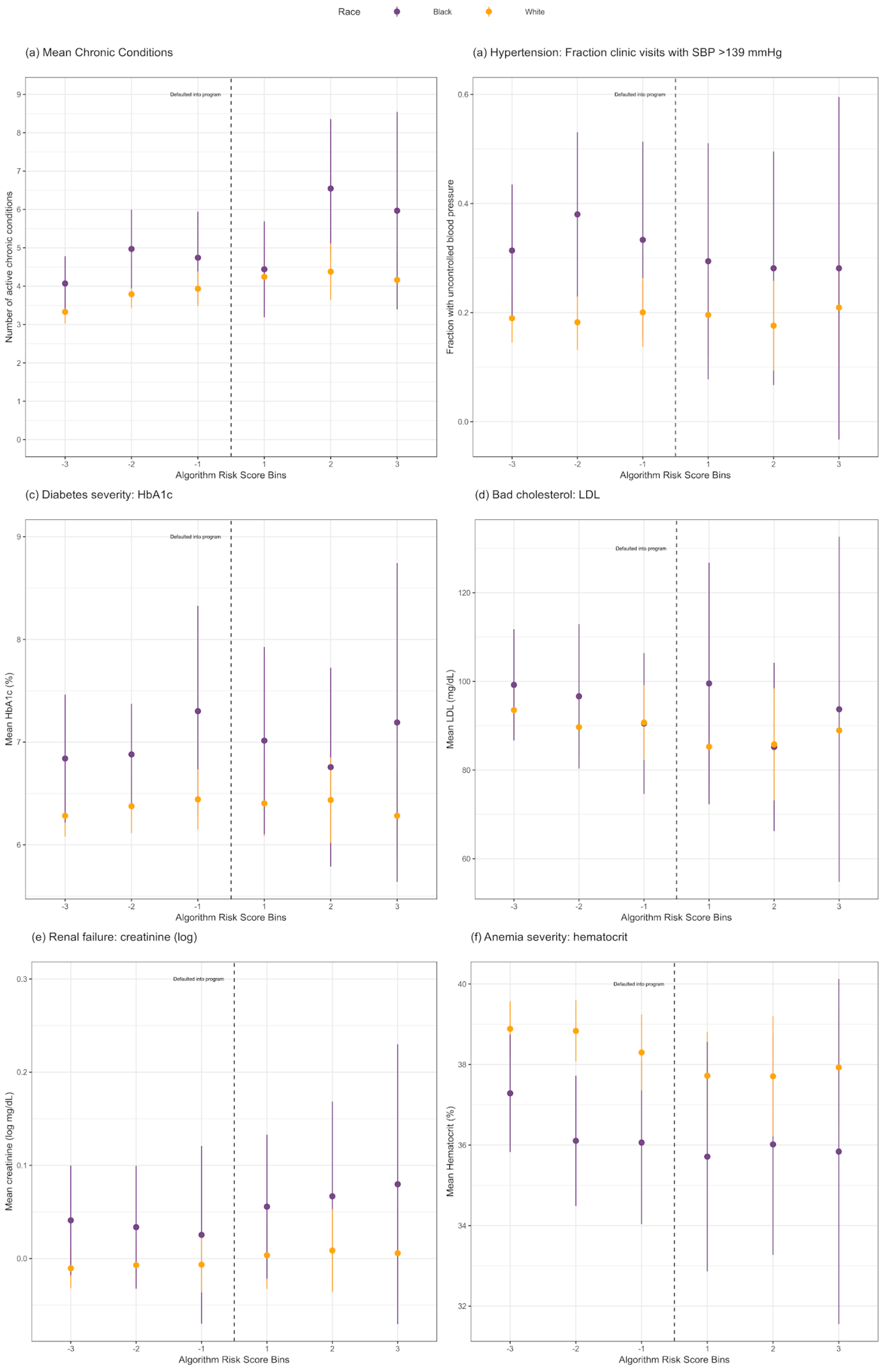
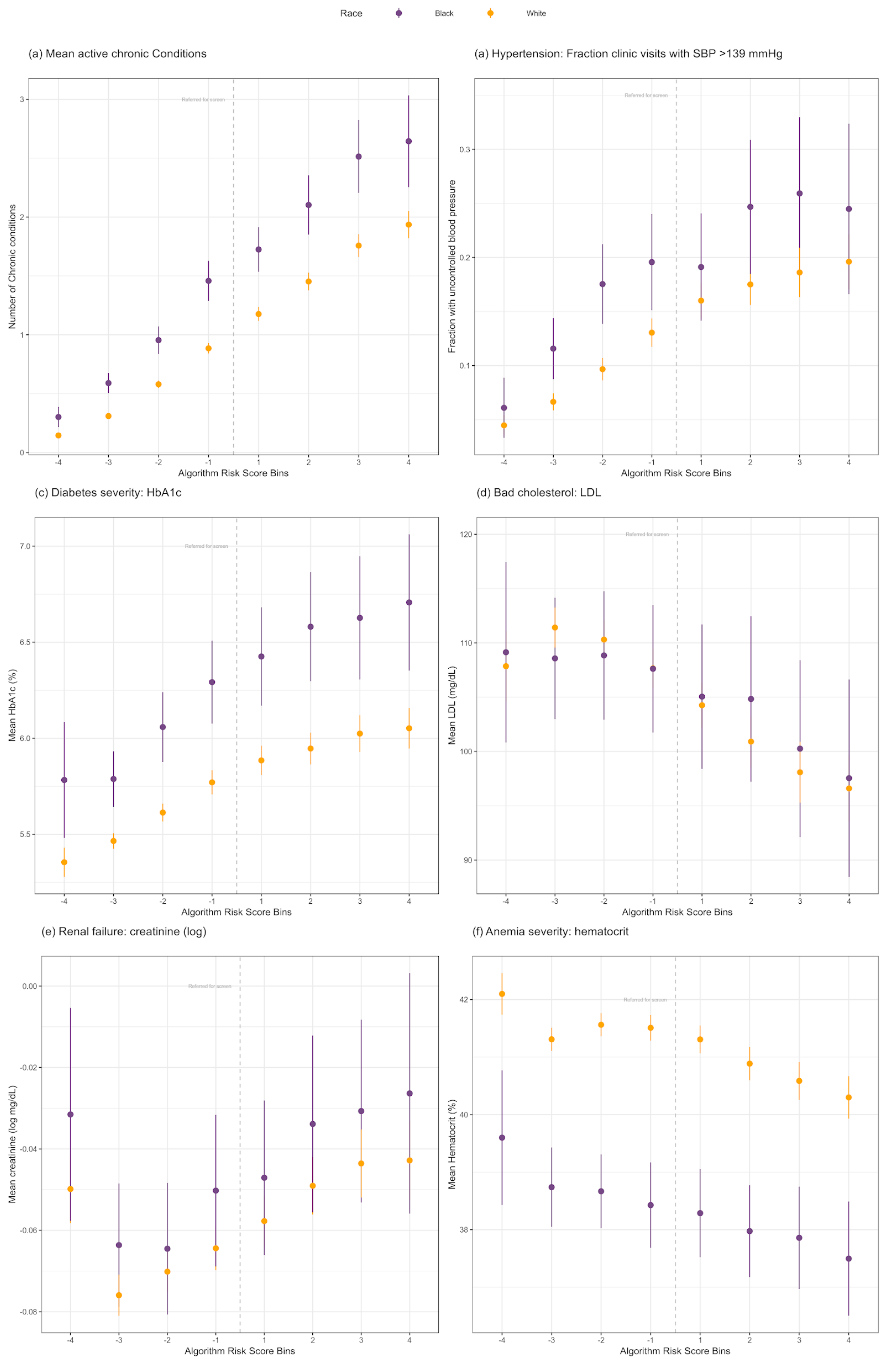


Fig. S4. Health measures vs. absolute bins of algorithm-predicted risk around auto-identification threshold. Racial differences in a range of health measures, comparing those above and below algorithm risk score thresholds used in auto-identification for enrollment in the care management program, for total number of chronic illnesses and biomarkers measuring severity of the most common diseases in the population studied. The x axis shows bins of risk immediately below and above the auto-identification threshold, which is indicated by the dotted black line.



(a) Mean active chronic Conditions

(a) Hypertension: Fraction clinic visits with SBP >139 mmHg

(c) Diabetes severity: HbA1c

(d) Bad cholesterol: LDL

(e) Renal failure: creatinine (log)

(f) Anemia severity: hematocrit

Fig. S5. Health measures vs. absolute bins of algorithm-predicted risk around screening threshold. Racial differences in a range of health measures, comparing those above and below algorithm risk score thresholds used in screening for enrollment in the care management program, for total number of chronic illnesses and biomarkers measuring severity of the most common diseases in the population studied. The x axis shows bins of risk immediately below and above the auto-identification threshold, which is indicated by the dotted grey line.

Table S1. Sample means: non-Black, non-White patients. Note: chronic illnesses are presented in the order used in Table 1.

n (patient-years)	19,070
n (patients)	10,440
Demographics	
Age	45.2
Female	0.60
Care management program	
Algorithm score	2.21
Fraction enrolled in program	0.07
Care utilization	
Actual cost	$7,316
Hospitalizations	0.09
Hospital days	0.50
Emergency visits	0.21
Outpatient visits	4.61
Mean biomarker values	
HbA1c	6.2
Systolic BP	124.6
Diastolic BP	73.3
Creatinine	0.87
Hematocrit	39.8
LDL	104.5
Prevalence of chronic illnesses	
Total number of illnesses	1.17
Hypertension	0.27

Diabetes, uncomplicated	0.13
Arrhythmia	0.05
Hypothyroid	0.05
Obesity	0.10
Pulmonary disease	0.08
Cancer	0.04
Depression	0.08
Anemia	0.06
Arthritis	0.03
Renal failure	0.03
Electrolyte disorder	0.03
Heart failure	0.02
Psychosis	0.04
Valvular disease	0.02
Stroke	0.02
Peripheral vascular disease	0.01
Diabetes, complicated	0.02
Heart attack	0.01
Liver disease	0.02

Table S2. Black-White differences by category of cost, and program effect. This table shows specific categories of medical expenditures by race, and indicates the difference in cost between Black and White patients, from a regression adjusting for total number of active chronic conditions (1) and the particular individual chronic conditions a patient has (2). It also shows whether high-risk care management programs have been found to be effective in reducing a particular category of cost (3).

	(1)	*(2)*	*(3)*	
Difference	*Sum of illnesses*	*Individual illnesses*	*Program effect*	*References*
Overall cost	-1807.84	-1148.11		
Overall cost (log)	-0.429	-0.367		
IP Surgical	-639.71	-430.804		
OP Specialists	-422.69	-369.215		
OP Surgery	-291.096	-236.993		
IP Medical	-285.991	-123.854	↓	(*40–46*)
Other	-207.637	-29.419		
Pharmacy	-192.561	-48.149		
Physical therapy	-71.94	-66.8		
Radiology	-46.974	-25.297		
Home health	-29.869	9.783	↑	(*44*)
OP Primary care	-26.665	-34.736	↑	(*44*)
Laboratory	-20.276	6.105		
Dialysis	108.919	91.638		
Emergency	110.999	120.854	↓	(*40*, *41*, *46*)

Table S3. Performance of experimental algorithms trained on a feature set including race.

	Fraction total outcome in highest-risk group (SE)						**Fraction black in highest-risk group (SE)**	
Algorithm training label	*Total costs*		*Avoidable costs*		*Active chronic conditions*			
Total costs	0.163	(0.003)	0.184	(0.003)	0.106	(0.002)	0.112	(0.002)
Avoidable costs	0.141	(0.003)	0.213	(0.003)	0.130	(0.003)	0.241	(0.003)
Active chronic conditions	0.122	(0.003)	0.181	(0.003)	0.148	(0.003)	0.285	(0.004)
Best-worst difference	*0.041*		*0.032*		*0.042*		*0.173*	

Table S4. Relationship of experimental algorithms' predictions to race. We regress an indicator for black race on our three experimental algorithms (using OLS and logit as alternative specifications). Predictions are correlated with race (as measured by the coefficient on the predictions) but do not substantially reconstruct it. Results are shown for predictors that include the race variable in the feature set for maximum potential correlation with race; R^2 and coefficients are lower when race is not included.

		(1)		*(2)*	
Algorithm: by training label and whether or not it includes race in the feature set	*R^2*	*OLS Estimate (SE)*	*P*	*Logit Estimate (SE)*	*P*
Total costs					
with race variable	0.002	-0.030 (0.005)	<0.001	-0.313 (0.052)	<0.001
without race variable	0.001	0.022 (0.007)	<0.001	0.203 (0.060)	<0.001
Avoidable costs					
with race variable	0.028	0.048 (0.002)	<0.001	0.360 (0.018)	<0.001
without race variable	0.010	0.056 (0.0024	<0.001	0.436 (0.018)	<0.001
Active chronic conditions					
with race variable	0.031	0.040 (0.002)	<0.001	0.305 (0.015)	<0.001
without race variable	0.018	0.030 (0.002)	<0.001	0.238 (0.015)	<0.001

References and Notes

1. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine Bias," *ProPublica* (23 May 2016); www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

2. S. Barocas, A. D. Selbst, Big data's disparate impact. *Calif. Law Rev.* **104**, 671 (2016).

3. A. Chouldechova, A. Roth, The frontiers of fairness in machine learning. arXiv:1810.08810 [cs.LG] (20 October 2018).

4. A. Datta, M. C. Tschantz, A. Datta, Automated experiments on ad privacy settings. *Proc. Privacy Enhancing Technol.* **2015**, 92–112 (2015). doi:10.1515/popets-2015-0007

5. L. Sweeney, Discrimination in online ad delivery. *Queue* **11**, 1–19 (2013). doi:10.1145/2460276.2460278

6. M. Kay, C. Matuszek, S. A. Munson, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM, 2015), pp. 3819–3828.

7. B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, A. K. Jain, Face Recognition Performance: Role of Demographic Information. *IEEE Trans. Inf. Forensics Security* **7**, 1789–1801 (2012). doi:10.1109/TIFS.2012.2214212

8. J. Buolamwini, T. Gebru, in *Proceedings of the Conference on Fairness, Accountability and Transparency* (PMLR, 2018), pp. 77–91.

9. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017). doi:10.1126/science.aal4230 Medline

10. S. Corbett-Davies, S. Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023 [cs.CY] (31 July 2018).

11. M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. T. Kalai, Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. arXiv:1901.09451 [cs.IR] (27 January 2019).

12. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 259–268.

13. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018). Medline

14. C. S. Hong, A. L. Siegel, T. G. Ferris, Caring for high-need, high-cost patients: What makes for a successful care management program? *Issue Brief (Commonwealth Fund)* **19**, 1–19 (2014). Medline

15. N. McCall, J. Cromwell, C. Urato, "Evaluation of Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Massachusetts General Hospital and Massachusetts General Physicians Organization (MGH)" (RTI International, 2010).

16. J. Hsu, M. Price, C. Vogeli, R. Brand, M. E. Chernew, S. K. Chaguturu, E. Weil, T. G. Ferris, Bending The Spending Curve By Altering Care Delivery Patterns: The Role Of

Care Management Within A Pioneer ACO. *Health Aff.* **36**, 876–884 (2017). doi:10.1377/hlthaff.2016.0922 Medline

17. L. Nelson, “Lessons from Medicare’s demonstration projects on disease management and care coordination” (Working Paper 2012-01, Congressional Budget Office, 2012).

18. C. Vogeli, A. E. Shields, T. A. Lee, T. B. Gibson, W. D. Marder, K. B. Weiss, D. Blumenthal, Multiple chronic conditions: Prevalence, health consequences, and implications for quality, care management, and costs. *J. Gen. Intern. Med.* **22** (suppl. 3), 391–395 (2007). doi:10.1007/s11606-007-0322-1 Medline

19. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* **33**, 1123–1131 (2014). doi:10.1377/hlthaff.2014.0041 Medline

20. J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, Prediction Policy Problems. *Am. Econ. Rev.* **105**, 491–495 (2015). doi:10.1257/aer.p20151023 Medline

21. G. Hileman, S. Steele, “Accuracy of claims-based risk scoring models” (Society of Actuaries, 2016).

22. J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807 [cs.LG] (19 September 2016).

23. A. Chouldechova, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* **5**, 153–163 (2017). doi:10.1089/big.2016.0047 Medline

24. V. de Groot, H. Beckerman, G. J. Lankhorst, L. M. Bouter, How to measure comorbidity. a critical review of available methods. *J. Clin. Epidemiol.* **56**, 221–229 (2003). doi:10.1016/S0895-4356(02)00585-1 Medline

25. J. J. Gagne, R. J. Glynn, J. Avorn, R. Levin, S. Schneeweiss, A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J. Clin. Epidemiol.* **64**, 749–759 (2011). doi:10.1016/j.jclinepi.2010.10.004 Medline

26. A. K. Parekh, M. B. Barton, The challenge of multiple comorbidity for the US health care system. *JAMA* **303**, 1303–1304 (2010). doi:10.1001/jama.2010.381 Medline

27. D. Ettehad, C. A. Emdin, A. Kiran, S. G. Anderson, T. Callender, J. Emberson, J. Chalmers, A. Rodgers, K. Rahimi, Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *Lancet* **387**, 957–967 (2016).

28. K.-T. Khaw, N. Wareham, R. Luben, S. Bingham, S. Oakes, A. Welch, N. Day, Glycated haemoglobin, diabetes, and mortality in men in Norfolk cohort of European Prospective Investigation of Cancer and Nutrition (EPIC-Norfolk). *BMJ* **322**, 15 (2001).

29. K. Fiscella, P. Franks, M. R. Gold, C. M. Clancy, Inequality in quality: Addressing socioeconomic, racial, and ethnic disparities in health care. *JAMA* **283**, 2579–2584 (2000). doi:10.1001/jama.283.19.2579 Medline

30. N. E. Adler, K. Newman, Socioeconomic disparities in health: Pathways and policies. *Health Aff.* **21**, 60–76 (2002). doi:10.1377/hlthaff.21.2.60 Medline

31. N. E. Adler, W. T. Boyce, M. A. Chesney, S. Folkman, S. L. Syme, Socioeconomic inequalities in health. No easy solution. *JAMA* **269**, 3140–3145 (1993). doi:10.1001/jama.1993.03500240084031 Medline

32. M. Alsan, O. Garrick, G. C. Graziani, "Does diversity matter for health? Experimental evidence from Oakland" (National Bureau of Economic Research, 2018).

33. K. Armstrong, K. L. Ravenell, S. McMurphy, M. Putt, Racial/ethnic differences in physician distrust in the United States. *Am. J. Public Health* **97**, 1283–1289 (2007). doi:10.2105/AJPH.2005.080762 Medline

34. M. Alsan, M. Wanamaker, Tuskegee and the health of black men. *Q. J. Econ.* **133**, 407–455 (2018). doi:10.1093/qje/qjx029 Medline

35. M. van Ryn, J. Burke, The effect of patient race and socio-economic status on physicians' perceptions of patients. *Soc. Sci. Med.* **50**, 813–828 (2000). doi:10.1016/S0277-9536(99)00338-X Medline

36. K. M. Hoffman, S. Trawalter, J. R. Axt, M. N. Oliver, Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4296–4301 (2016). doi:10.1073/pnas.1516047113 Medline

37. J. J. Escarce, F. W. Puffer, "Black-White Differences in the Use of Medical Care by the Elderly: A Contemporary Analysis" in *Racial and Ethnic Differences in the Health of Older Americans* (National Academies Press, 1997), chap. 6; www.ncbi.nlm.nih.gov/books/NBK109841/.

38. S. Passi, S. Barocas, Problem Formulation and Fairness. arXiv:1901.02547 [cs.CY] (8 January 2019).

39. S. Mullainathan, Z. Obermeyer, Does Machine Learning Automate Moral Hazard and Error? *Am. Econ. Rev.* **107**, 476–480 (2017). doi:10.1257/aer.p20171084 Medline

40. K. E. Joynt Maddox, M. Reidhead, J. Hu, A. J. H. Kind, A. M. Zaslavsky, E. M. Nagasako, D. R. Nerenz, Adjusting for social risk factors impacts performance and penalties in the hospital readmissions reduction program. *Health Serv. Res.* **54**, 327–336 (2019). doi:10.1111/1475-6773.13133 Medline

41. K. E. Joynt Maddox, M. Reidhead, A. C. Qi, D. R. Nerenz, Association of Stratification by Dual Enrollment Status With Financial Penalties in the Hospital Readmissions Reduction Program. *JAMA Intern. Med.* **179**, 769–776 (2019). doi:10.1001/jamainternmed.2019.0117 Medline

42. K. Lum, W. Isaac, To predict and serve? *Significance* **13**, 14–19 (2016). doi:10.1111/j.1740-9713.2016.00960.x

43. I. Ajunwa, "The Paradox of Automation as Anti-Bias Intervention," available at SSRN (2016); https://ssrn.com/abstract=2746078.

44. S. DellaVigna, M. Gentzkow, "Uniform pricing in US retail chains" (National Bureau of Economic Research, 2017).

45. C. A. Gomez-Uribe, N. Hunt, The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manag. Inf. Syst.* **6**, 13 (2016). doi:10.1145/2843948

46. Z. Obermeyer, S. Mullainathan, in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2019), p. 89; extended abstract.

47. G. Weiss, L. T. Goodnough, Anemia of chronic disease. *N. Engl. J. Med.* **352**, 1011–1023 (2005). doi:10.1056/NEJMra041809 Medline

48. M. Tonelli, N. Wiebe, B. Culleton, A. House, C. Rabbat, M. Fok, F. McAlister, A. X. Garg, Chronic kidney disease and mortality risk: A systematic review. *J. Am. Soc. Nephrol.* **17**, 2034–2047 (2006). doi:10.1681/ASN.2005101085 Medline

49. H. Ujiie, M. Kawasaki, Y. Suzuki, M. Kaibara, Influence of age and hematocrit on the coagulation of blood. *J. Biorheol.* **23**, 111–114 (2009). doi:10.1007/s12573-010-0015-y

50. M. G. Silverman, B. A. Ference, K. Im, S. D. Wiviott, R. P. Giugliano, S. M. Grundy, E. Braunwald, M. S. Sabatine, Association Between Lowering LDL-C and Cardiovascular Risk Reduction Among Different Therapeutic Interventions: A Systematic Review and Meta-analysis. *JAMA* **316**, 1289–1297 (2016). doi:10.1001/jama.2016.13985 Medline

51. B. Nowok, G. M. Raab, C. Dibben, synthpop: Bespoke creation of synthetic data in R. *J. Stat. Softw.* **74**, 1–26 (2016). doi:10.18637/jss.v074.i11