Dissecting racial bias in an algorithm that guides health decisions for millions

Ziad Obermeyer
UC Berkeley

Joint work with Sendhil Mullainathan

Preview

Preview

- Enormous optimism around machine learning in health
  - Disclaimer: this is most of my work

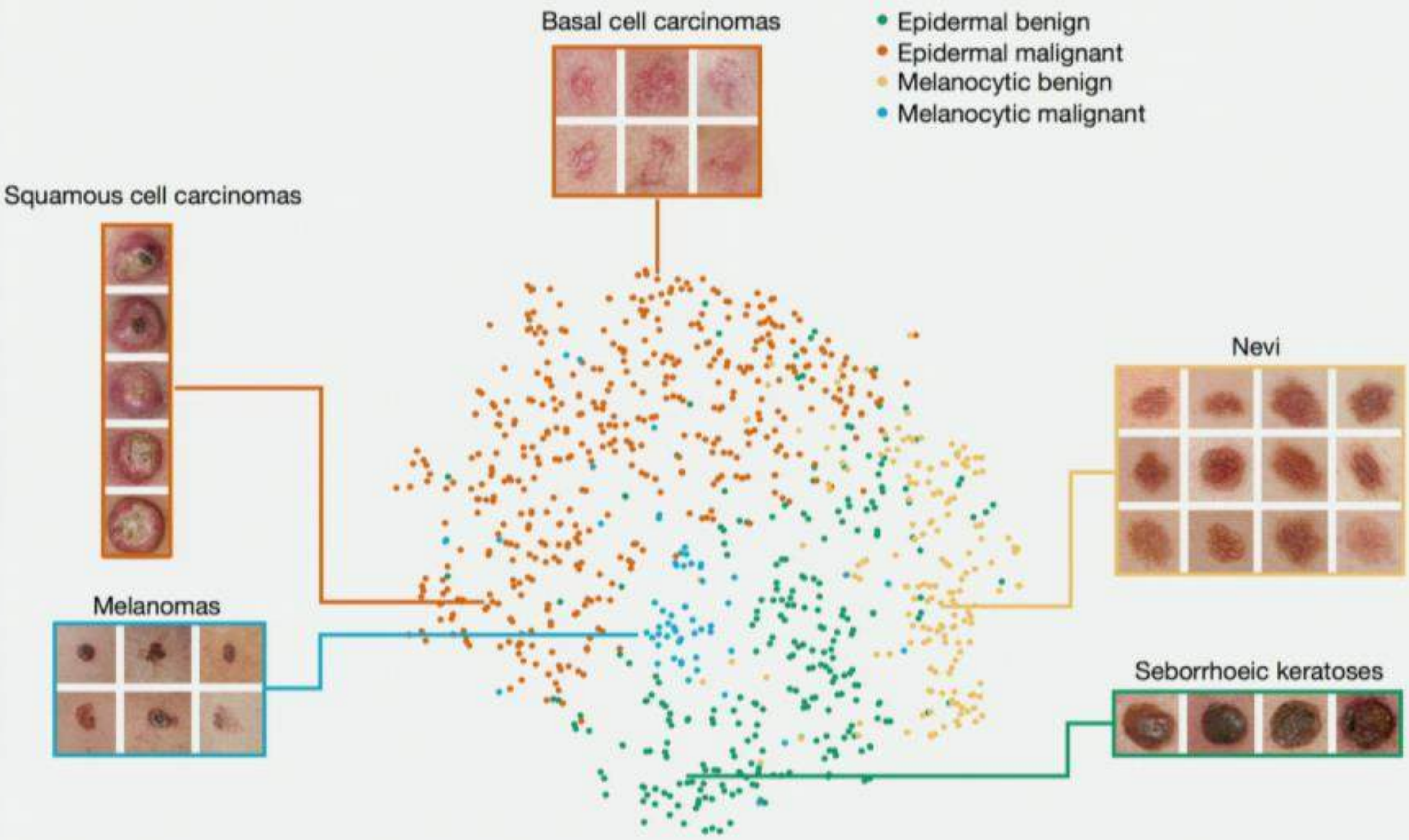
Preview

- Enormous optimism around machine learning in health
  - Disclaimer: this is most of my work
- But these tools can also do harm
  - May be harder to diagnose in health data

Preview

- Enormous optimism around machine learning in health
  - Disclaimer: this is most of my work
- But these tools can also do harm
  - May be harder to diagnose in health data
- We analyze a widely-used health algorithm
  - We find significant racial bias
  - This affects decisions for millions of patients
  - A unique dataset lets us trace the cause
  - And provide a solution

On the horizon: Melanoma detection



Esteva et al, Dermatologist-level classification of skin cancer with deep neural networks *Nature* 2017

Already at scale: Policy algorithms

- As we anticipate deployment, we can imagine problems

Already at scale: Policy algorithms

- As we anticipate deployment, we can imagine problems
- But I'll focus on a different kind of algorithm

Already at scale: Policy algorithms

- As we anticipate deployment, we can imagine problems
- But I'll focus on a different kind of algorithm
- **The kind already deployed in the health system**
  - Decisions for tens of millions affected every year

Already at scale: Policy algorithms

- As we anticipate deployment, we can imagine problems
- But I'll focus on a different kind of algorithm
- **The kind already deployed in the health system**
  - Decisions for tens of millions affected every year
- Response to major policy changes
  - Incentives to manage populations of patients
  - Rather than fee-for-service model

A major problem in our health care system

A major problem in our health care system

- Poor management of complex patients
  - Multiple chronic conditions, medications, etc
  - Multiple doctors: specialists, PCP, etc
  - Burdensome emergency visits, hospitalizations when health deteriorates

A major problem in our health care system

- Poor management of complex patients
  - Multiple chronic conditions, medications, etc
  - Multiple doctors: specialists, PCP, etc
  - Burdensome emergency visits, hospitalizations when health deteriorates

- Can we improve their interface with health care system?

A major problem in our health care system

- Poor management of complex patients
  - Multiple chronic conditions, medications, etc
  - Multiple doctors: specialists, PCP, etc
  - Burdensome emergency visits, hospitalizations when health deteriorates
- Can we improve their interface with health care system?
- Widespread innovation: care coordination programs

What these programs do

What these programs do

- When you have a problem – just call
  - Dedicated phone number
  - Low threshold to raising and solving issues

What these programs do

- When you have a problem – just call
  - Dedicated phone number
  - Low threshold to raising and solving issues
- When you need it – extra help
  - Dedicated staff of trained nurses
  - Extra slots for appointments
  - Help with medications

What these programs do

- When you have a problem – just call
  - Dedicated phone number
  - Low threshold to raising and solving issues
- When you need it – extra help
  - Dedicated staff of trained nurses
  - Extra slots for appointments
  - Help with medications
- Coordinate across doctors to reconcile medications, solve multi-specialty problems, …

Do they work?

Do they work?

- Evidence is generally positive
- Help medically complex patients get the right care
  - Find problems before health deteriorates
  - Avoid emergency visits, hospitalizations

Do they work?

- Evidence is generally positive
- Help medically complex patients get the right care
  - Find problems before health deteriorates
  - Avoid emergency visits, hospitalizations
- Better quality, lower cost
  - Good for both patients and payers

Algorithms play a major role industry-wide

Algorithms play a major role industry-wide

- Programs are expensive – so targeting critical

Algorithms play a major role industry-wide

- Programs are expensive – so targeting critical

- This is where algorithms come in
  - Industry estimates: 150 million lives affected per year
  - The particular algorithm we study: 70 million

Algorithms play a major role industry-wide

- Programs are expensive – so targeting critical
- This is where algorithms come in
  - Industry estimates: 150 million lives affected per year
  - The particular algorithm we study: 70 million
- Flag patients with "complex health needs" for "intervention before their health becomes catastrophic"
  - Predict next year's health care costs
  - Use claims data (and to a lesser degree EHR data)

Algorithms: our example

Algorithms: our example

- Algorithm trained on a national claims dataset
  - Then applied to our data
  - Large primary care practice (n=43,539, 12% black)

Algorithms: our example

- Algorithm trained on a national claims dataset
  - Then applied to our data
  - Large primary care practice (n=43,539, 12% black)

- We have unique access to
  - Algorithm <u>inputs</u> : predictors
  - Algorithm <u>outputs</u> : objective function, predictions
  - Rich contextual data on health

Analyzing this live, scaled algorithm for racial bias

Analyzing this live, scaled algorithm for racial bias

- Many measures of bias
  - Not all can be simultaneously satisfied
  - We'll focus on policy effect: who gets into program?

Analyzing this live, scaled algorithm for racial bias

- Many measures of bias
  - Not all can be simultaneously satisfied
  - We'll focus on policy effect: who gets into program?
- Algorithm score S used to screen for program
  - 97th percentile: Auto-identified for enrollment
  - 55th percentile: Present to doctors for screening

Analyzing this live, scaled algorithm for racial bias

- Many measures of bias
  - Not all can be simultaneously satisfied
  - We'll focus on policy effect: who gets into program?
- Algorithm score S used to screen for program
  - 97th percentile: Auto-identified for enrollment
  - 55th percentile: Present to doctors for screening
- So we'll look at :
  - $E[Y \mid White, S]$ vs. $E[Y \mid Black, S]$
  - For a variety of (future) health outcomes Y

Chronic medical conditions

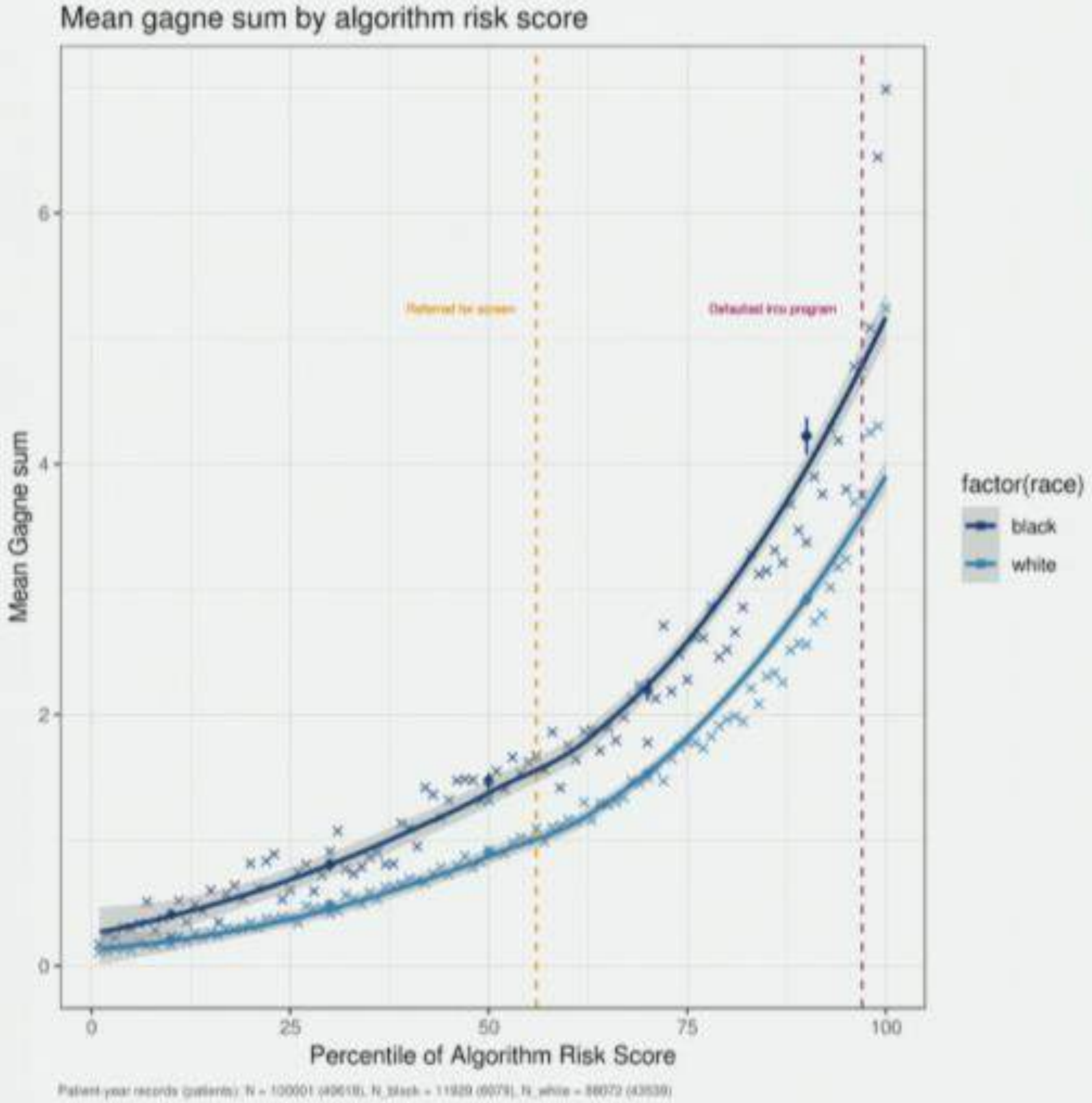
	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: More chronic illnesses

Chronic medical conditions vs risk score

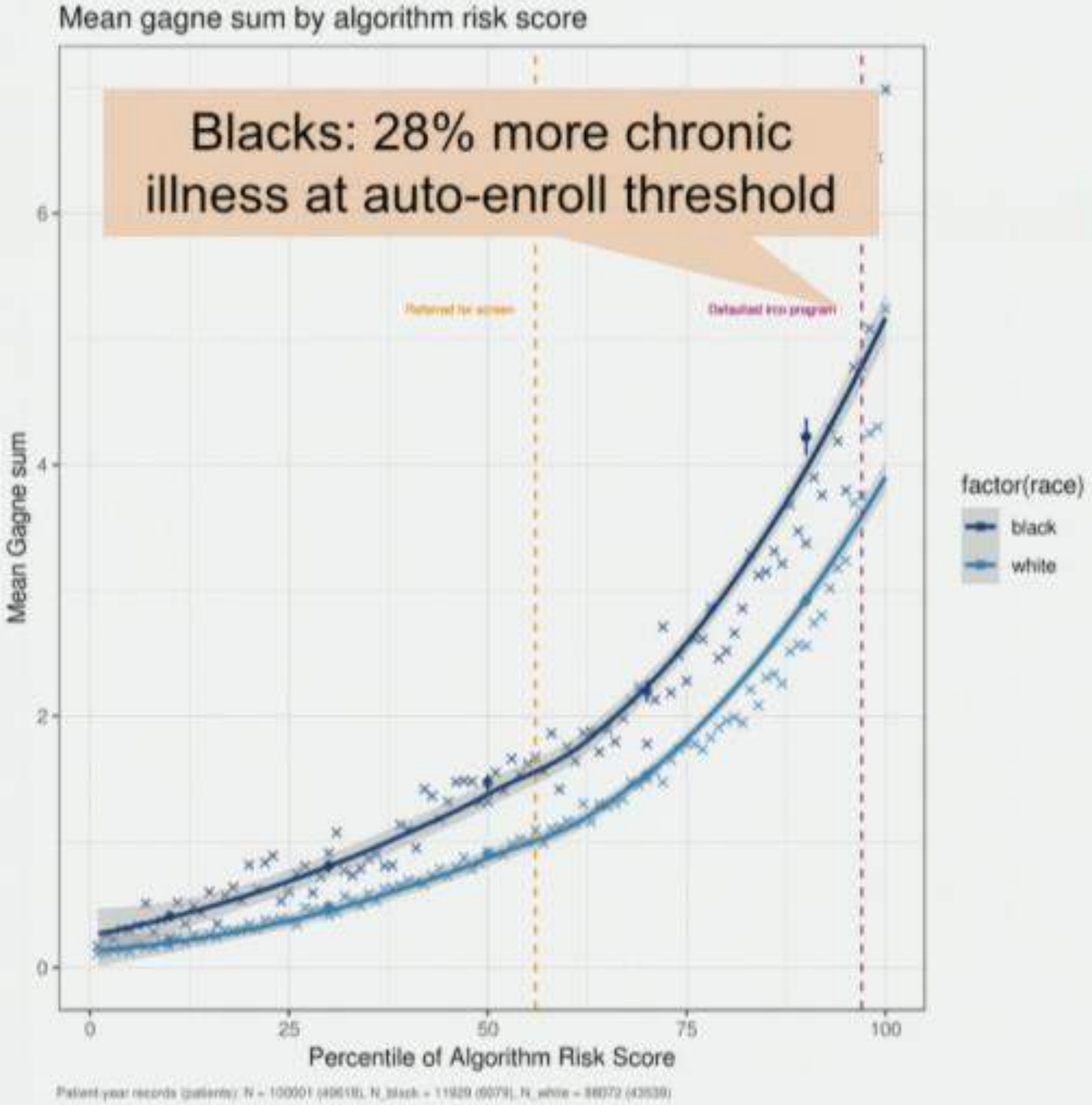


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: More chronic illnesses

Chronic medical conditions vs risk score

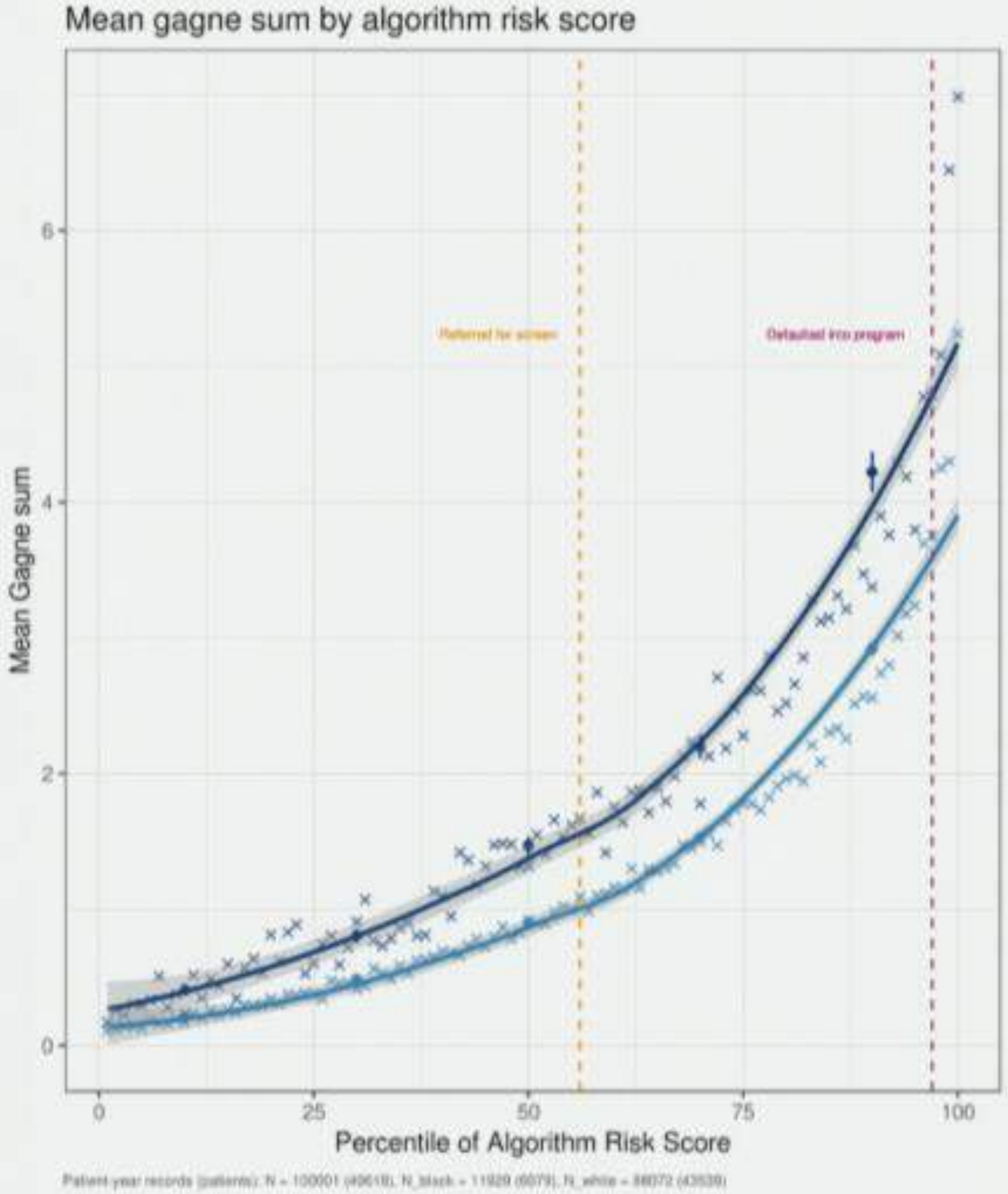


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

How consequential is this difference?

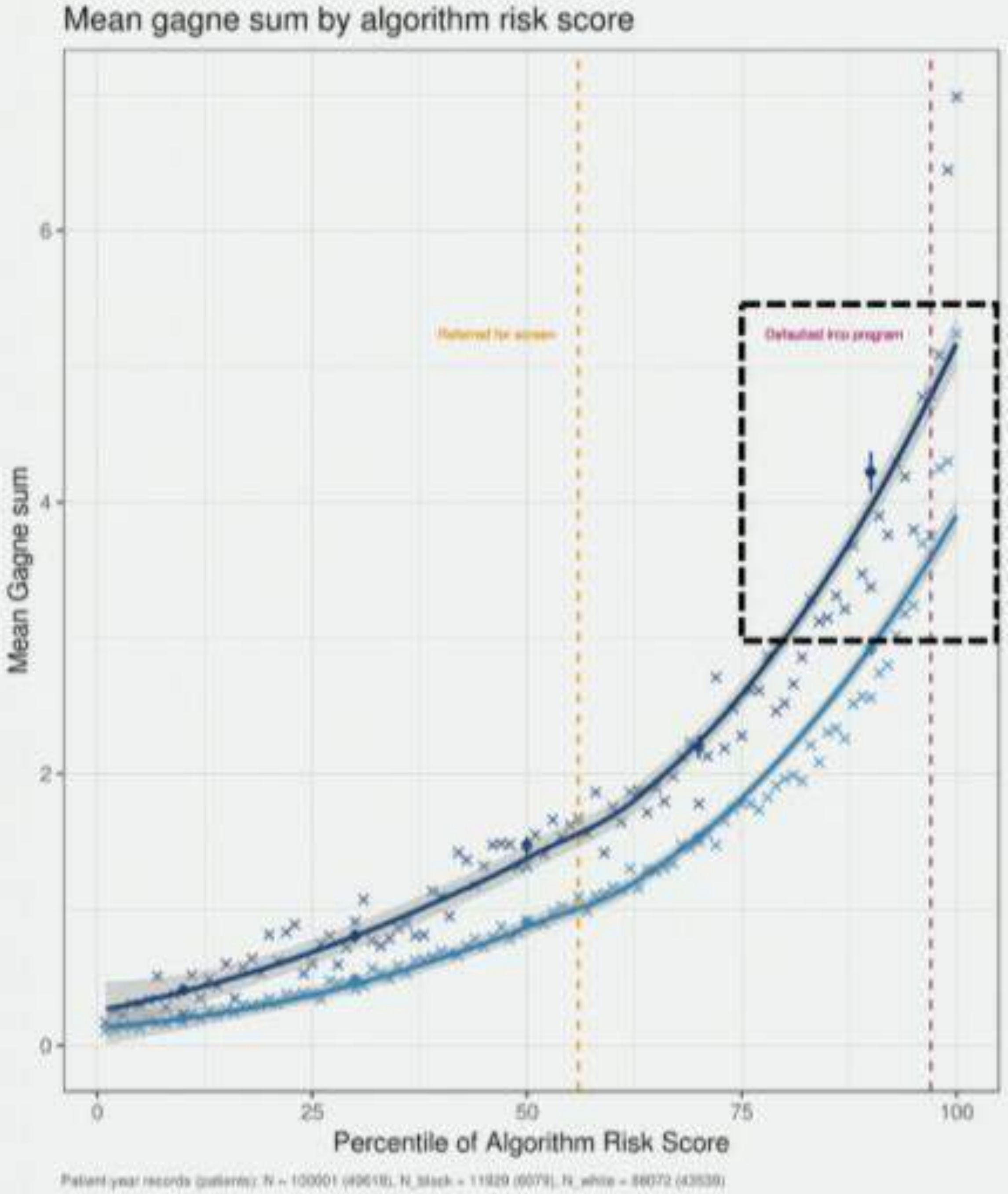
Chronic medical conditions vs risk score



- At that high risk score (auto-enroll: 97th pctile)

How consequential is this difference?

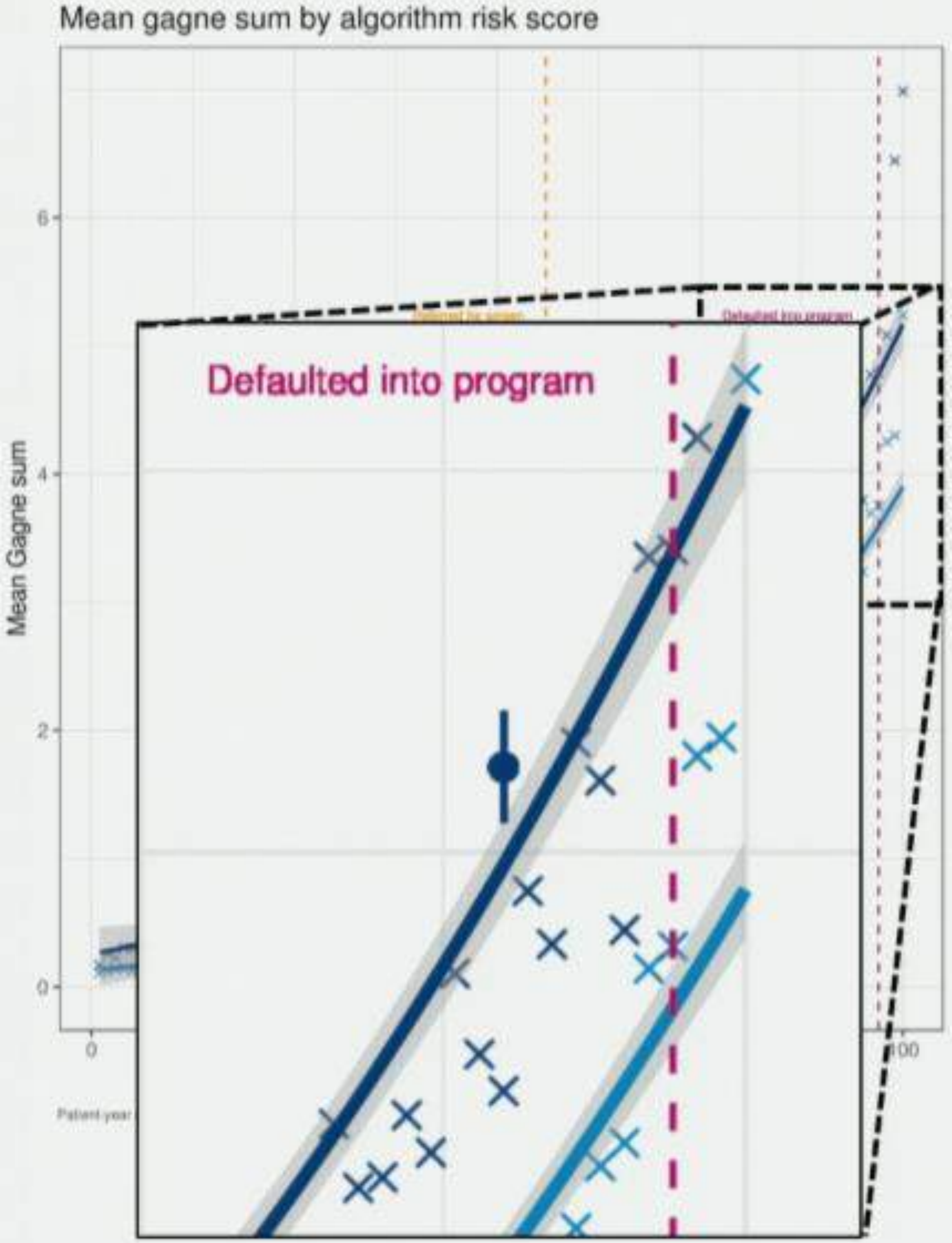
Chronic medical conditions vs risk score



- At that high risk score (auto-enroll: 97th pctile)

How consequential is this difference?

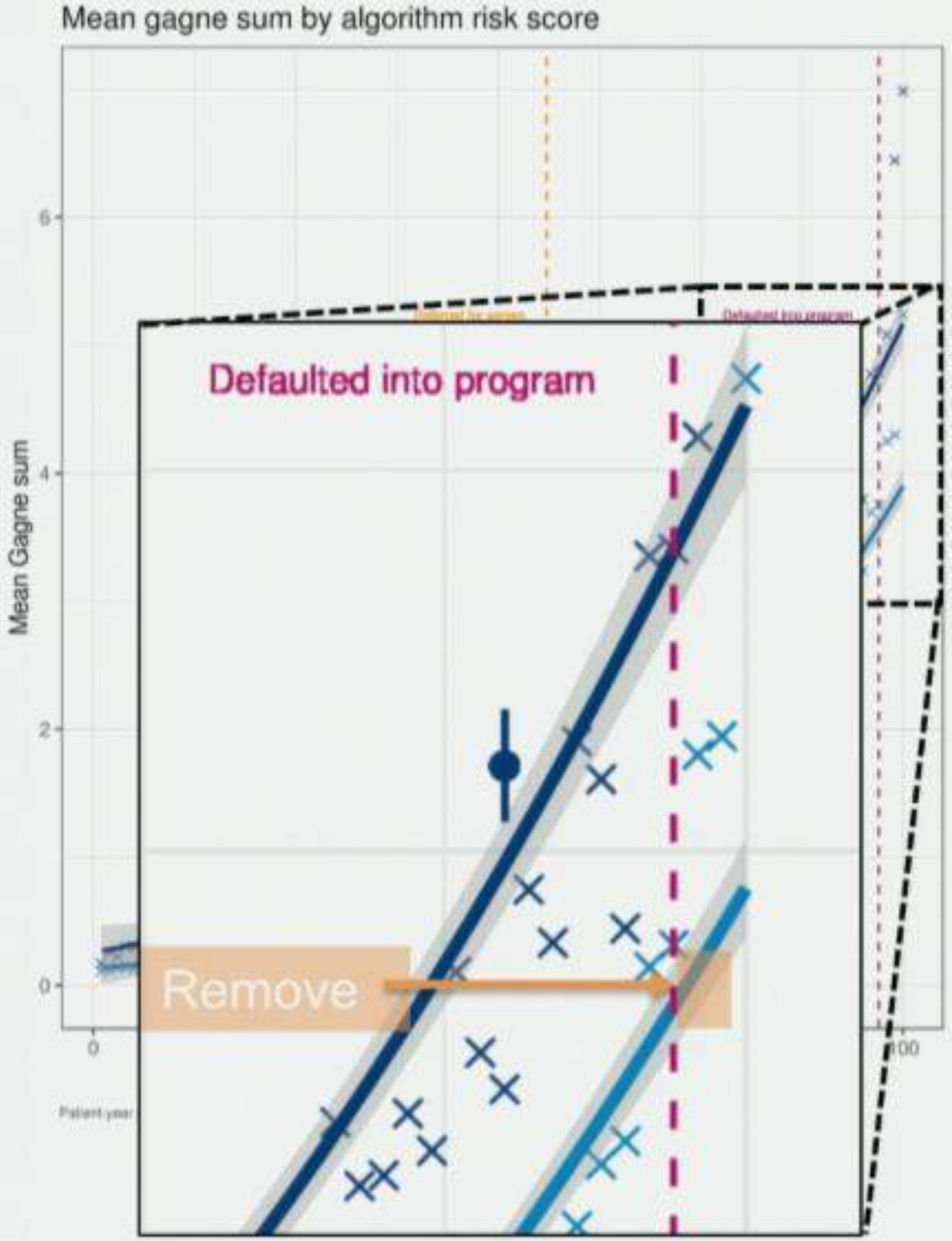
Chronic medical conditions vs risk score



- At that high risk score (auto-enroll: 97th pctile)

How consequential is this difference?

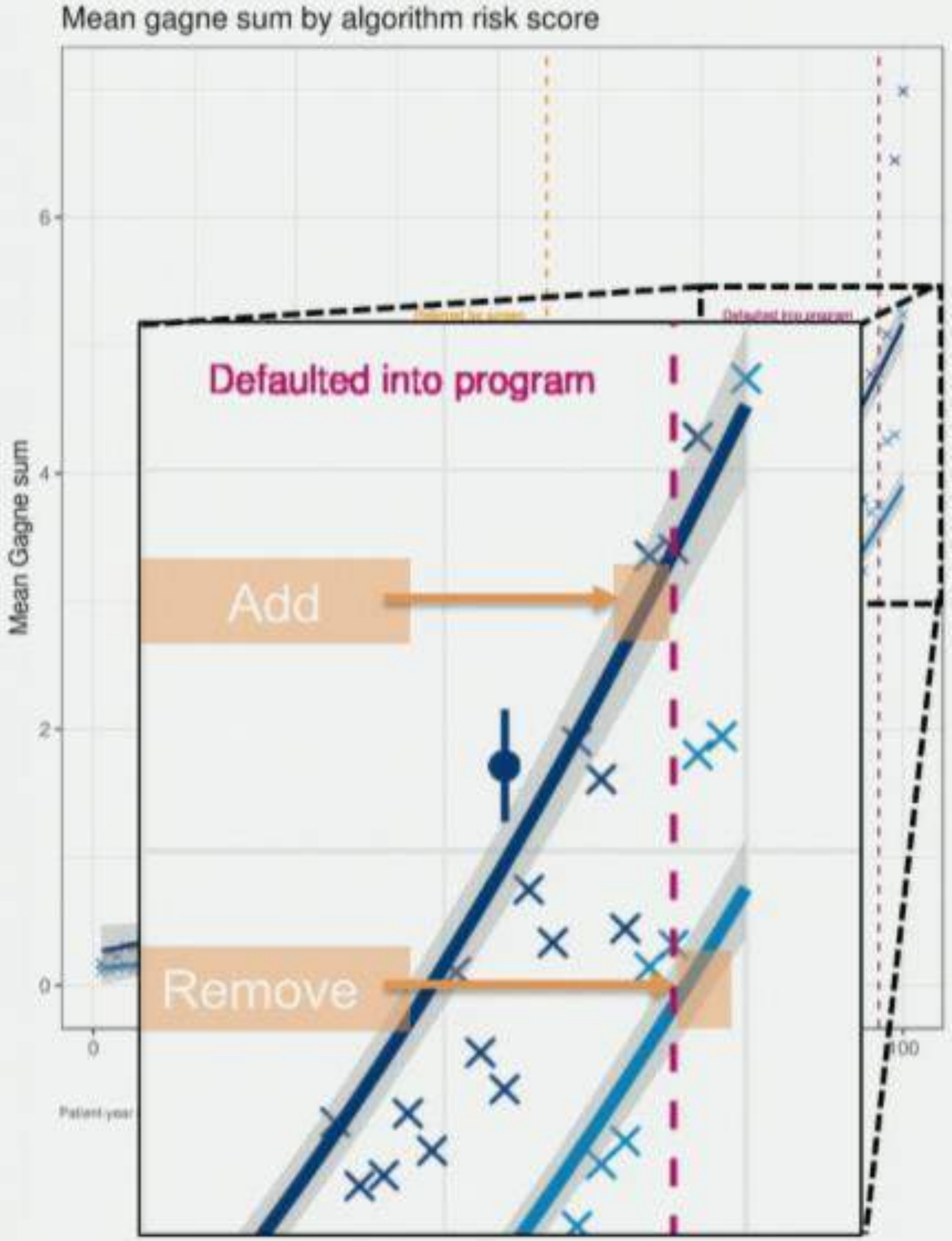
Chronic medical conditions vs risk score



- At that high risk score (auto-enroll: 97th pctile)
- Simulate
  - Remove healthy whites

How consequential is this difference?

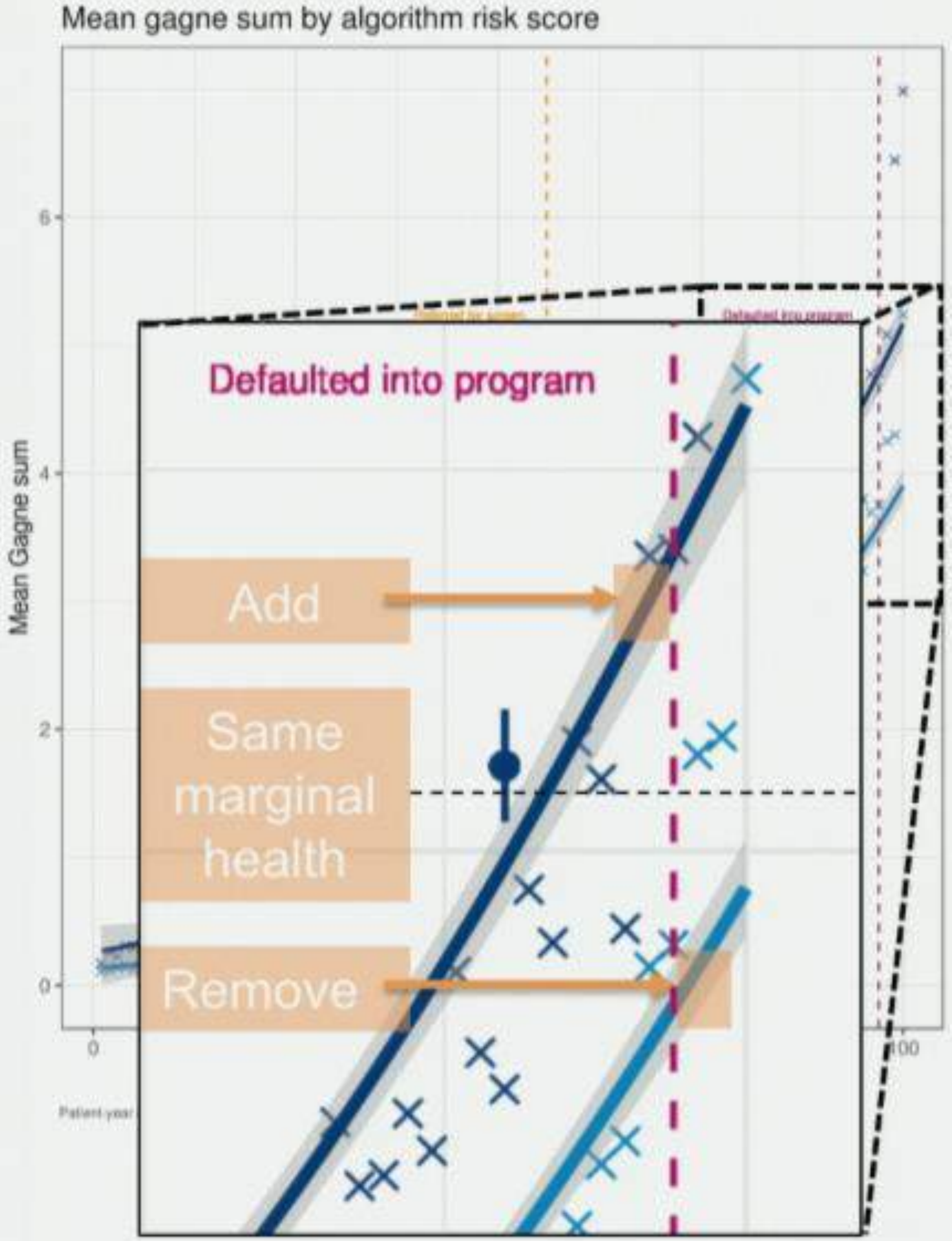
Chronic medical conditions vs risk score



- At that high risk score (auto-enroll: 97th pctile)
- Simulate
  - Remove healthy whites
  - Replace with less healthy blacks

How consequential is this difference?

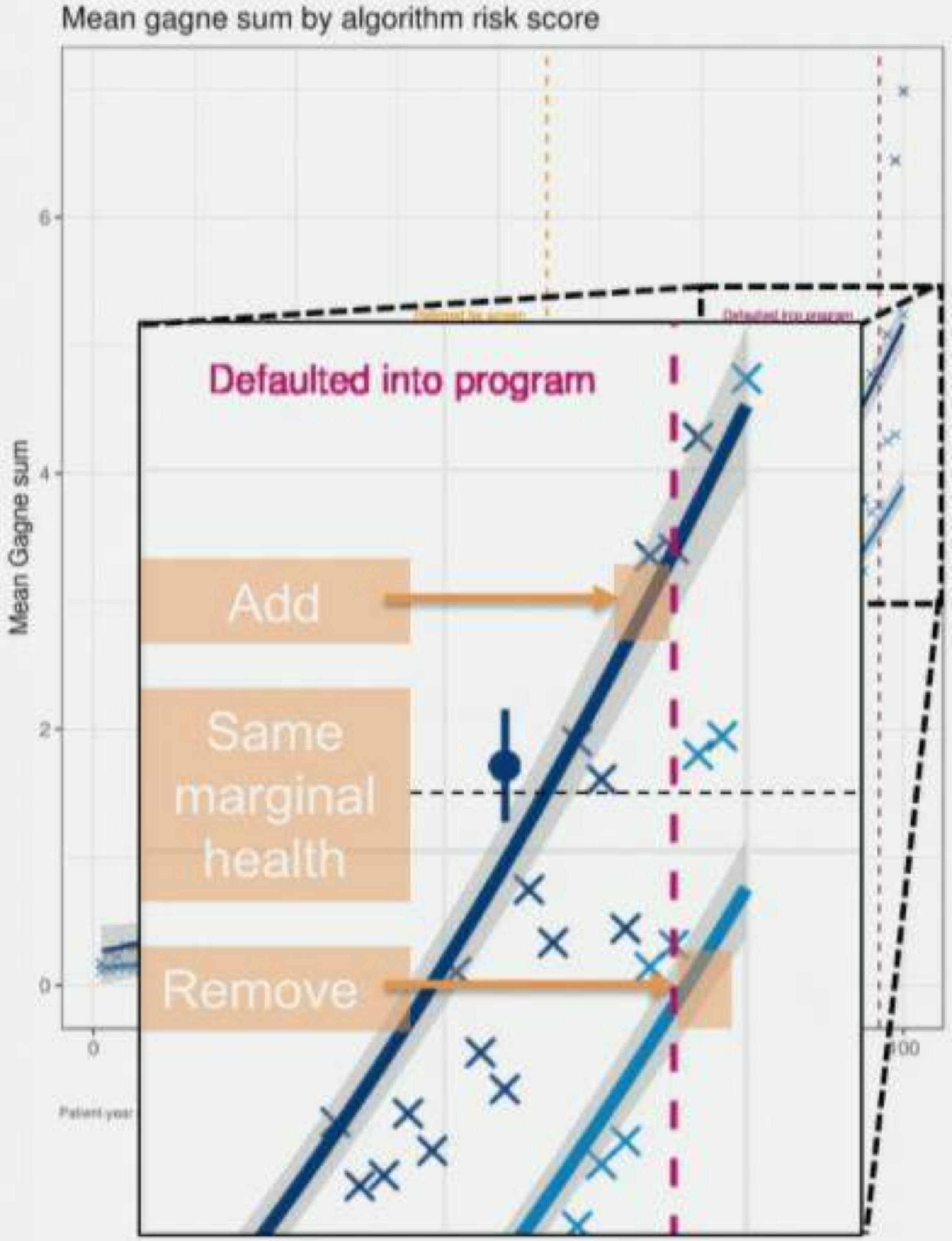
Chronic medical conditions vs risk score



- At that high risk score (auto-enroll: 97th pctile)
- Simulate
  - Remove healthy whites
  - Replace with less healthy blacks
  - Until marginal patient equally healthy

How consequential is this difference?

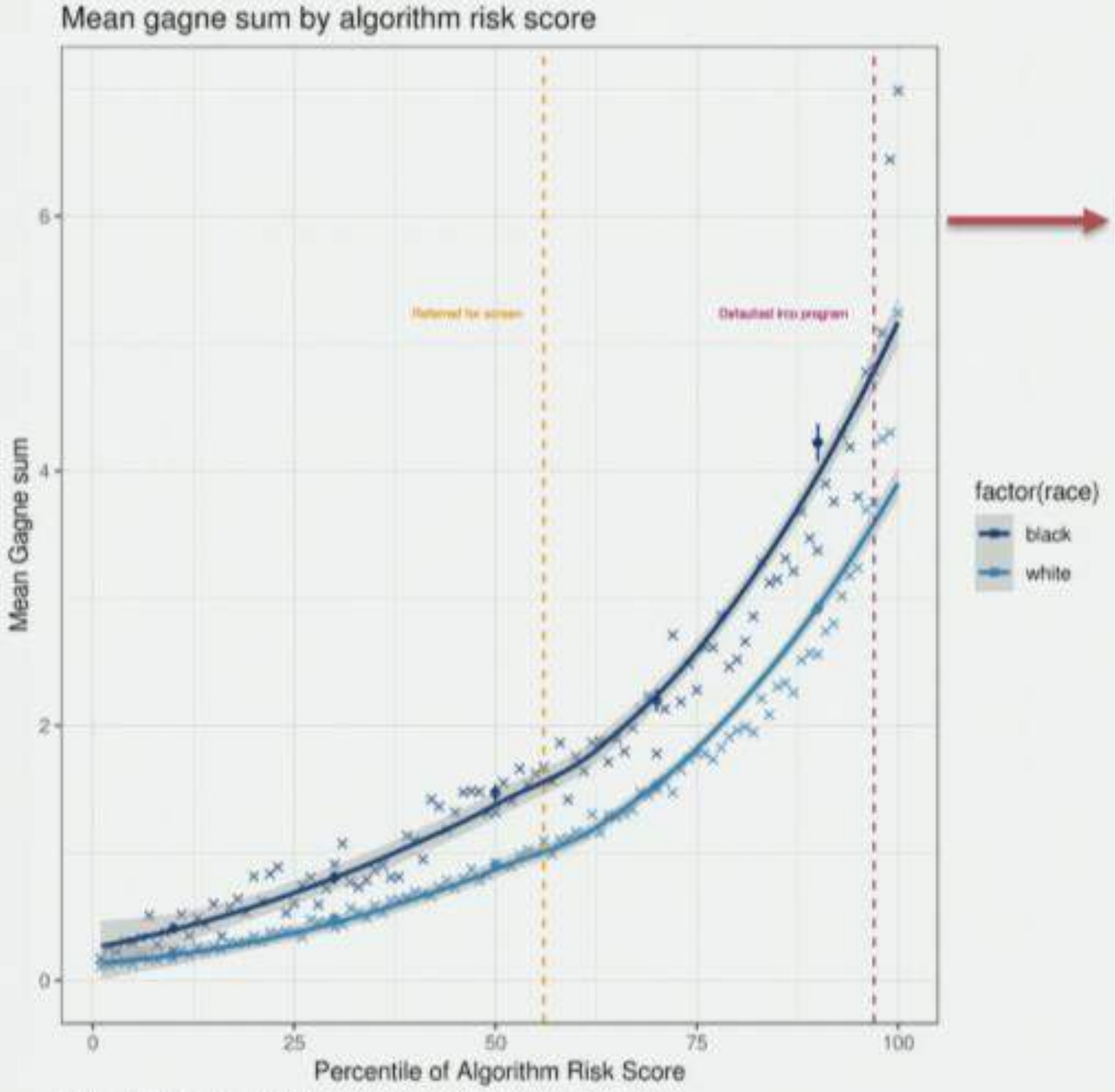
Chronic medical conditions vs risk score



- At that high risk score (auto-enroll: 97th pctile)
- Simulate
  - Remove healthy whites
  - Replace with less healthy blacks
  - Until marginal patient equally healthy
- **Double fraction of auto-enrolled blacks**
  - **17.7% → 46.5%**

Black vs white, same risk score: More chronic illnesses

Chronic medical conditions vs risk score

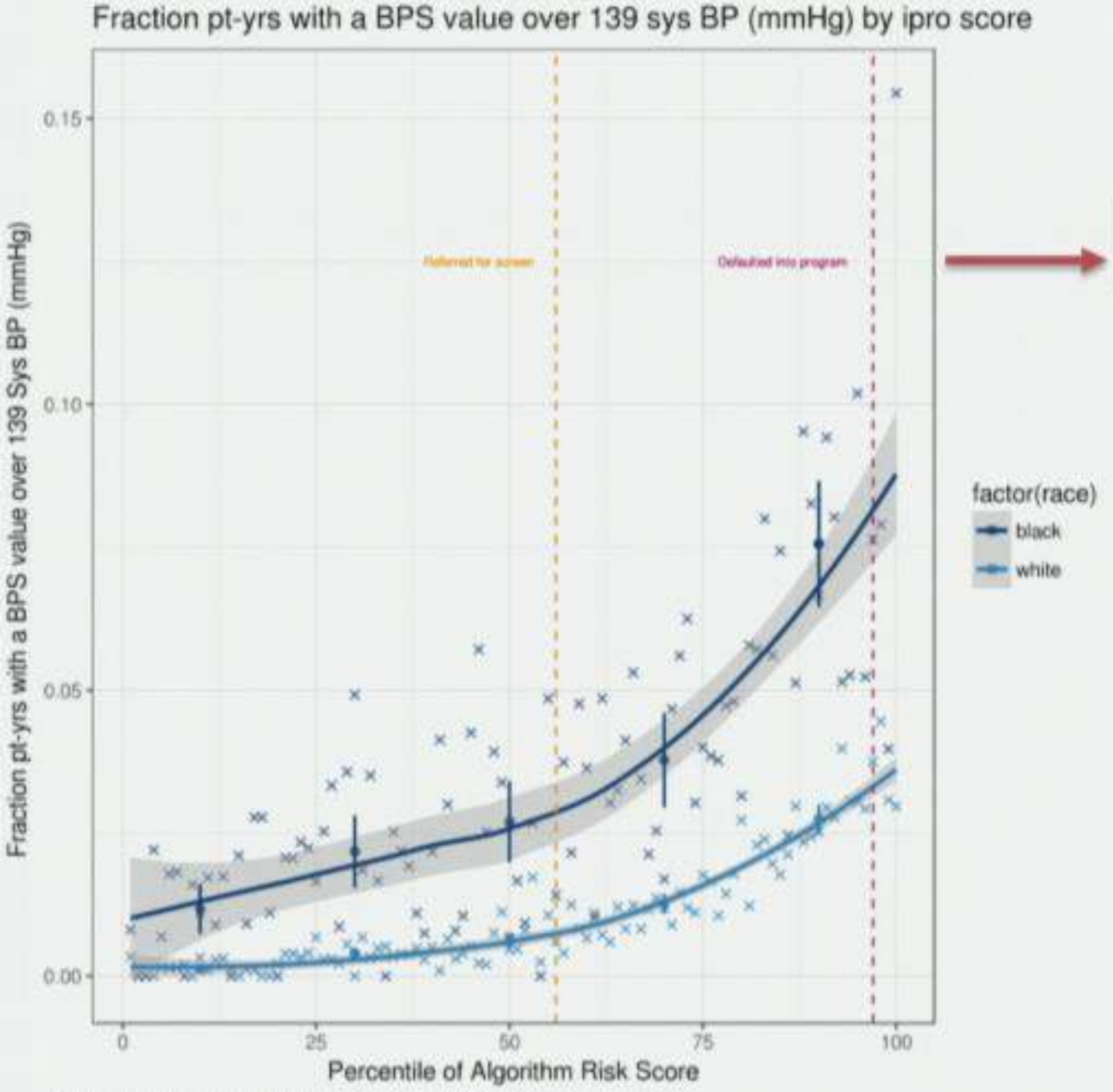


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Higher blood pressure

Uncontrolled SBP vs risk score

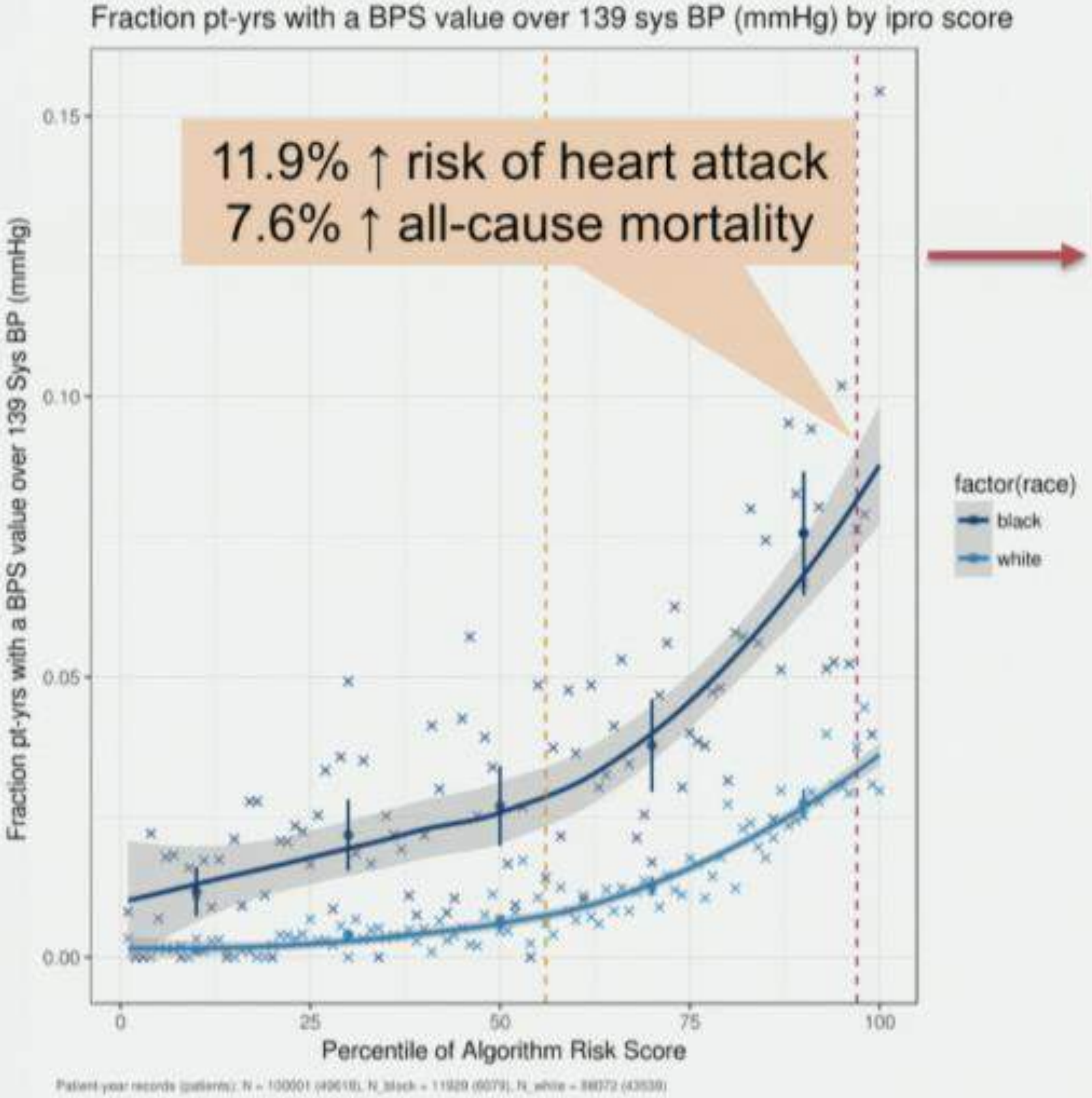


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Higher blood pressure

Uncontrolled SBP vs risk score

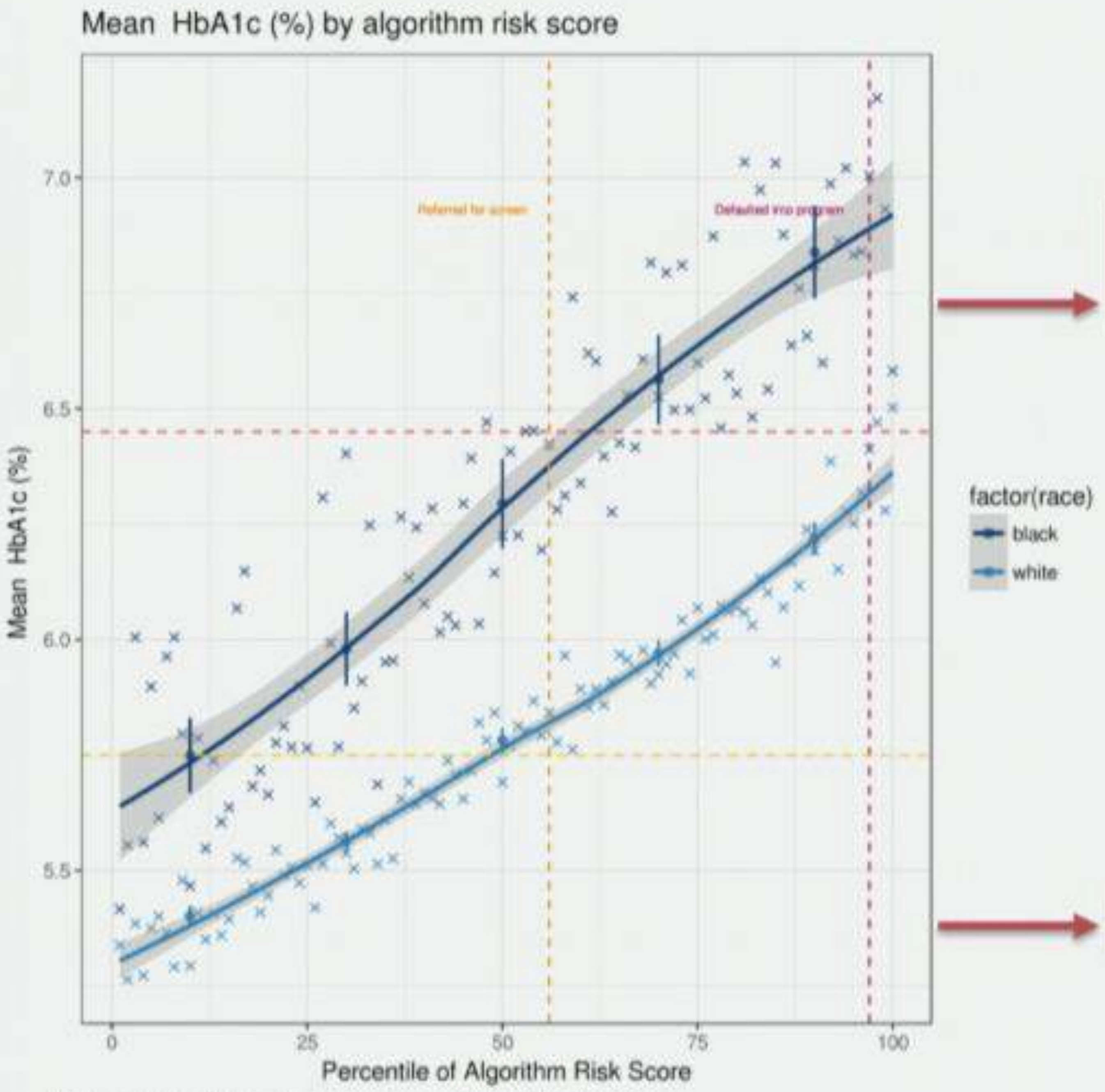


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Worse diabetes

Blood glucose (HbA1c) vs risk score

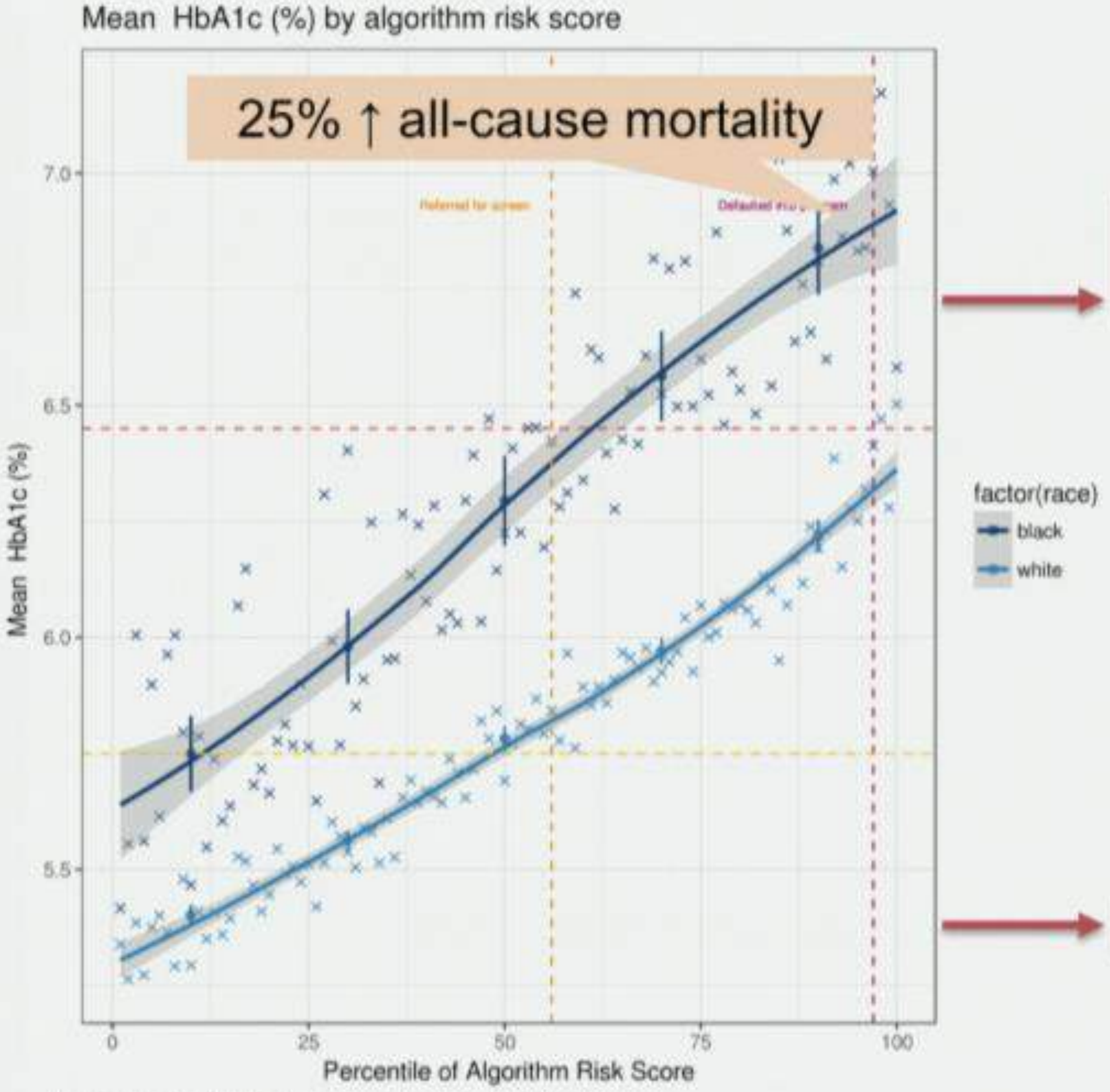


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Worse diabetes

Blood glucose (HbA1c) vs risk score

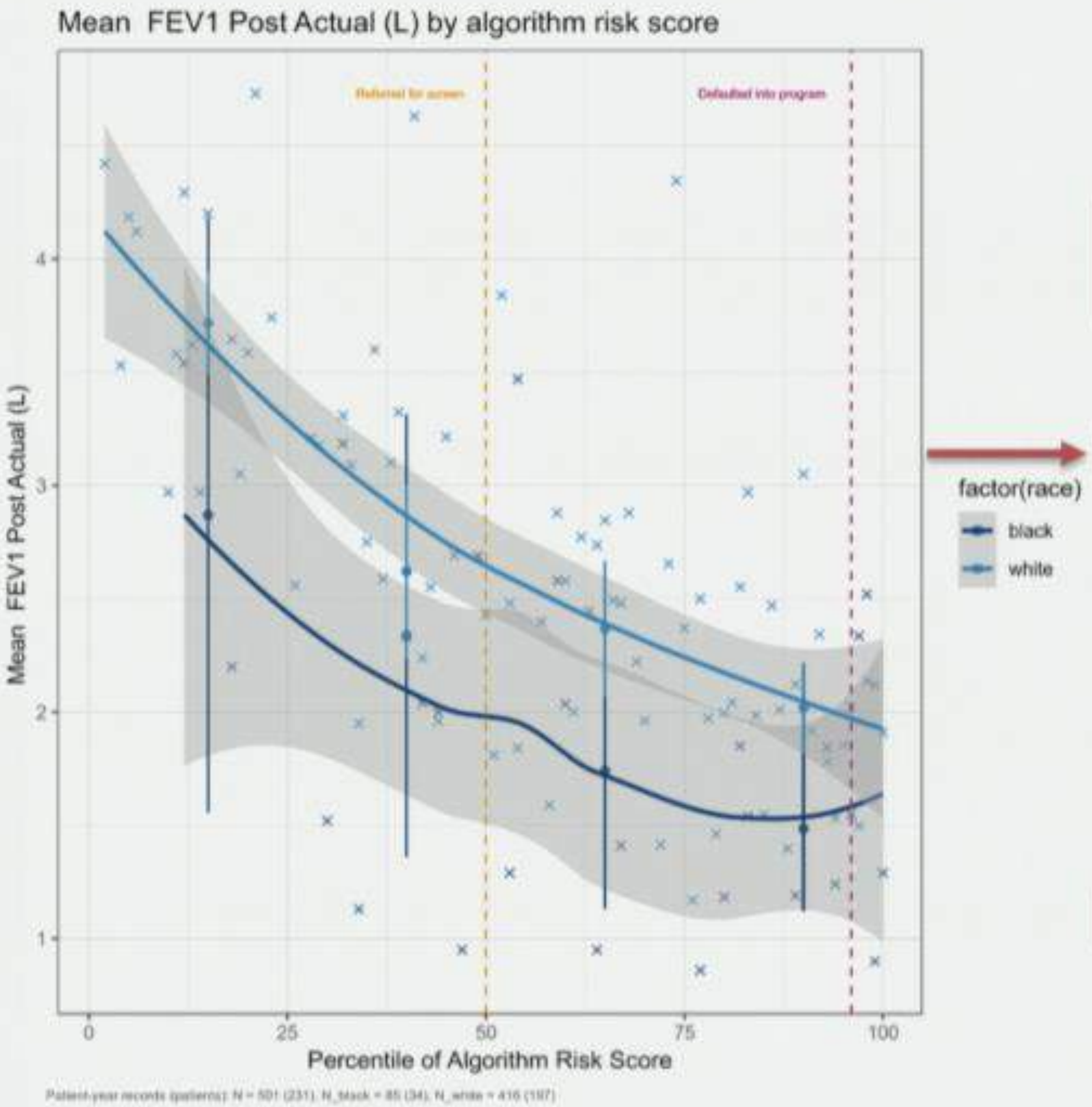


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Worse lung function

Lung function (FEV1) vs risk score

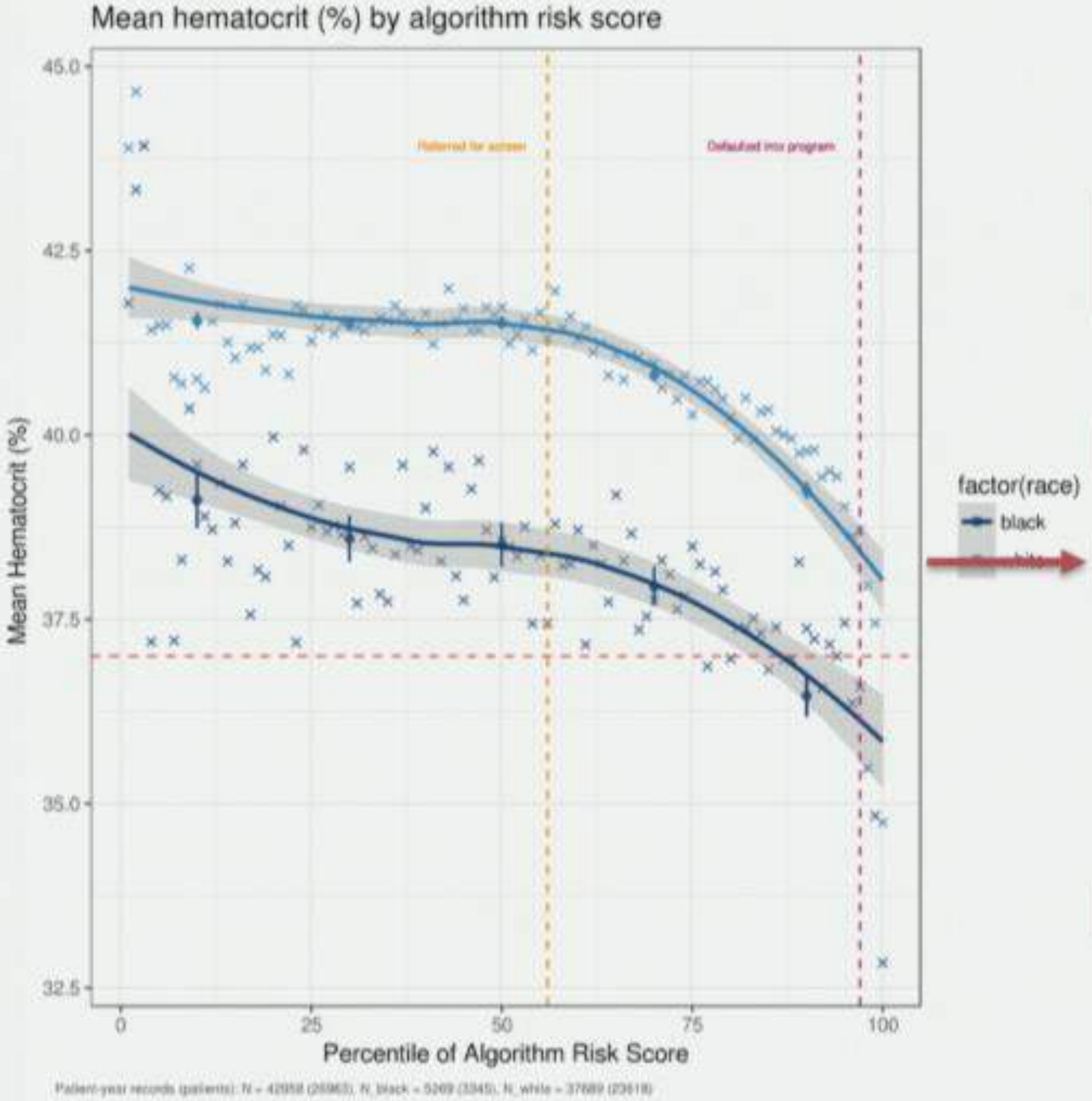


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Worse anemia

Red blood cells (hematocrit) vs risk score

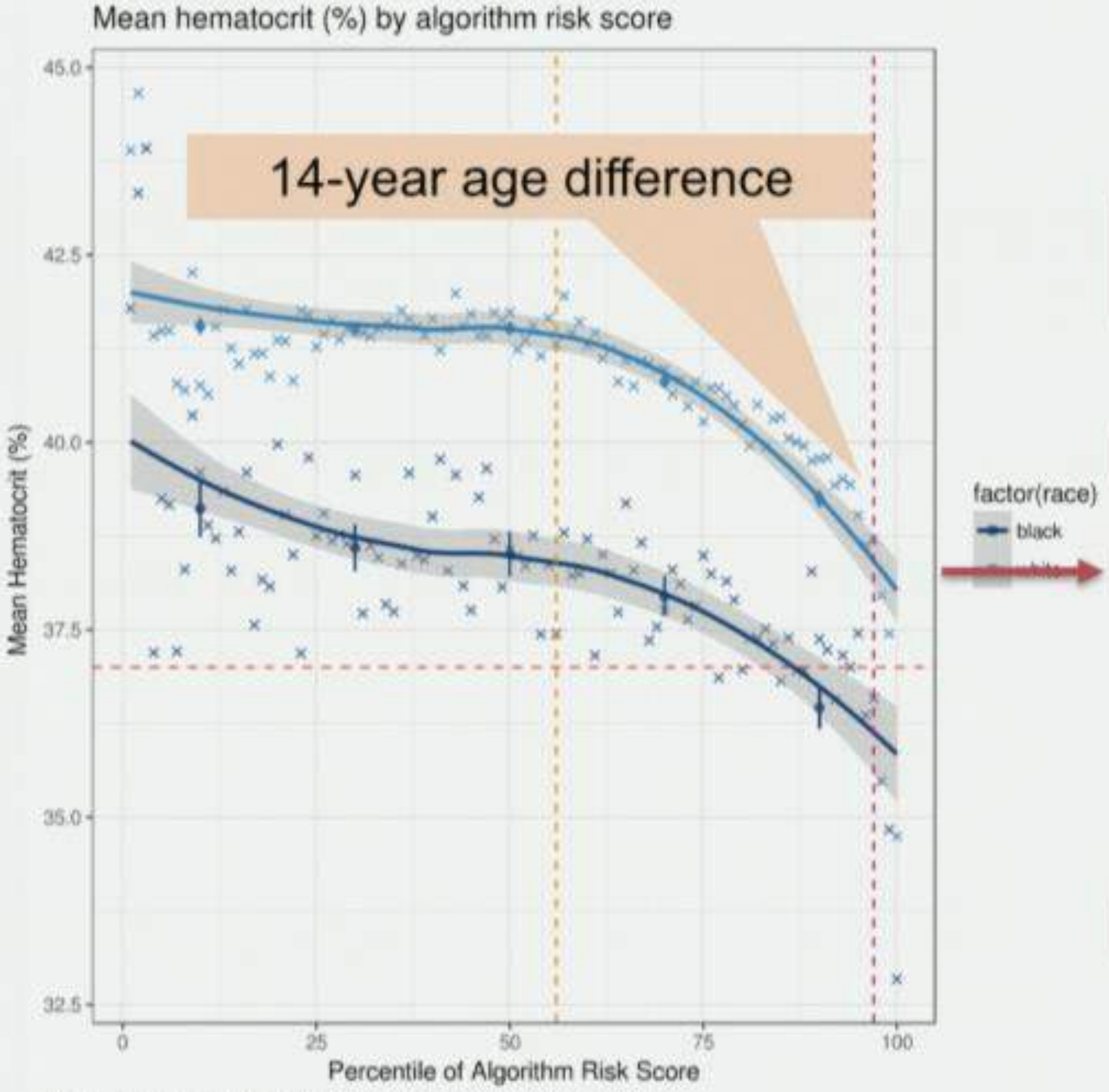


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Worse anemia

Red blood cells (hematocrit) vs risk score

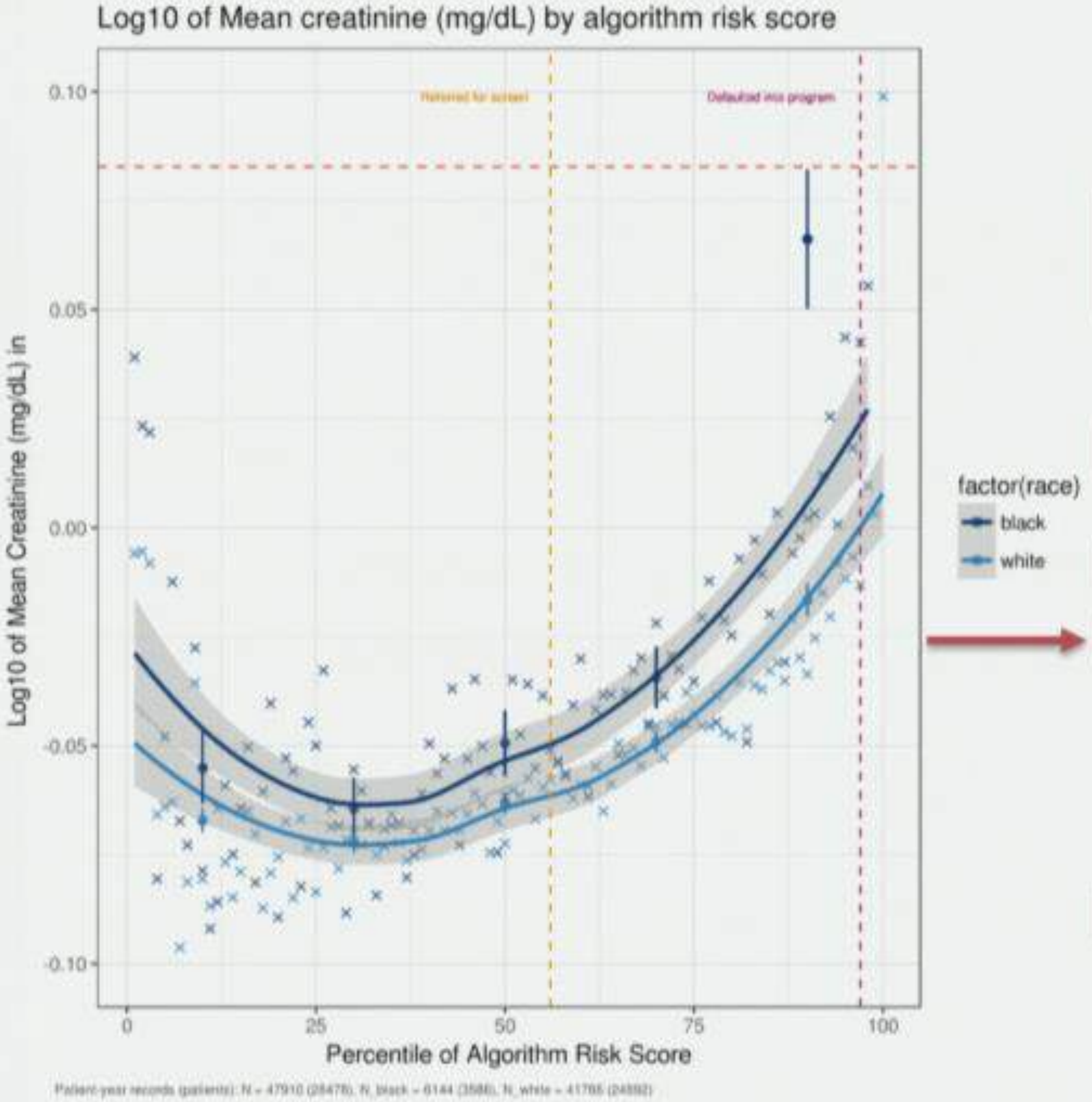


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Worse kidney function

Kidney function (creatinine) vs risk score

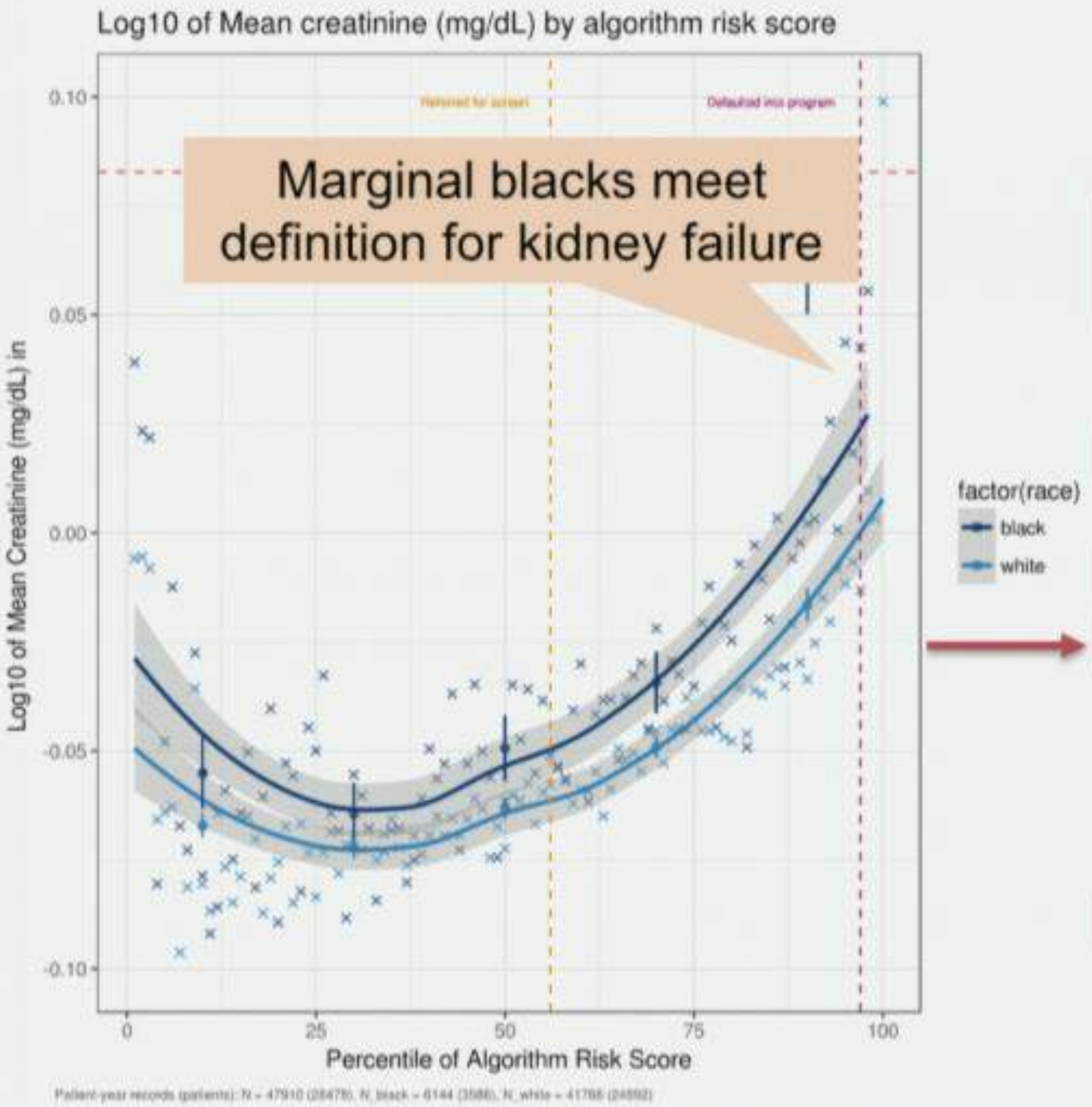


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

Black vs white, same risk score: Worse kidney function

Kidney function (creatinine) vs risk score

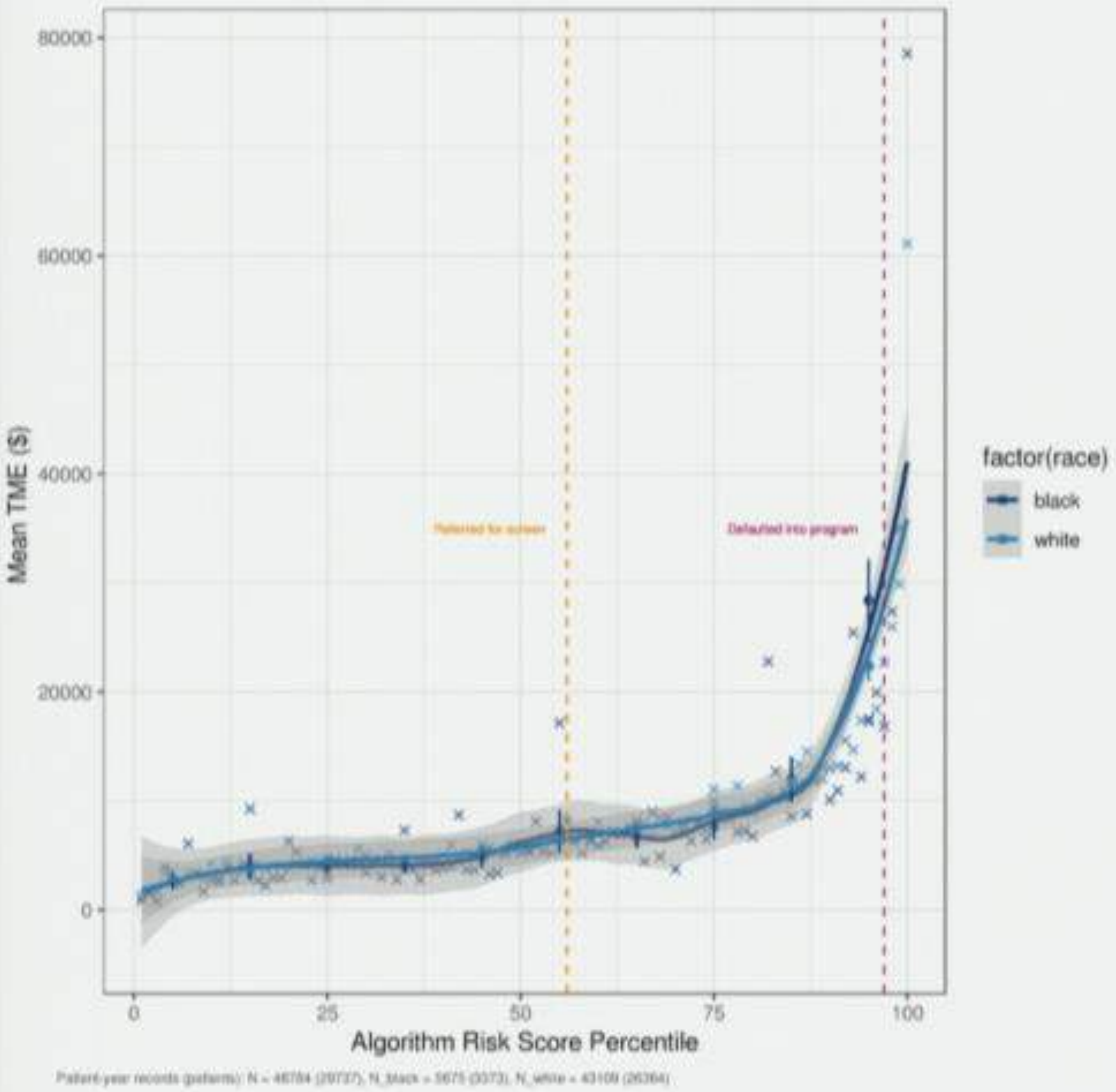


Chronic medical conditions

	All	White	Black
n (patient-years)	100,001	88,072	11,929
n (patients)	49,618	43,539	6,079
Prevalence of chronic illnesses			
Total number of illnesses	1.3	1.2	1.9
Hypertension	0.30	0.29	0.44
Diabetes, uncomplicated	0.14	0.08	0.22
Arrythmia	0.09	0.09	0.08
Hypothyroid	0.08	0.09	0.05
Obesity	0.08	0.07	0.18
Pulmonary disease	0.08	0.07	0.11
Cancer	0.07	0.07	0.06
Depression	0.06	0.06	0.07
Anemia	0.06	0.05	0.10
Arthritis	0.04	0.04	0.04
Renal failure	0.04	0.03	0.07
Electrolyte disorder	0.03	0.03	0.05
Heart failure	0.03	0.03	0.05
Psychosis	0.03	0.03	0.05
Valvular disease	0.03	0.03	0.02
Stroke	0.02	0.02	0.03
Peripheral vascular disease	0.02	0.02	0.02
Complicated diabetes	0.02	0.01	0.07
Heart attack	0.01	0.01	0.02
Liver disease	0.01	0.01	0.02

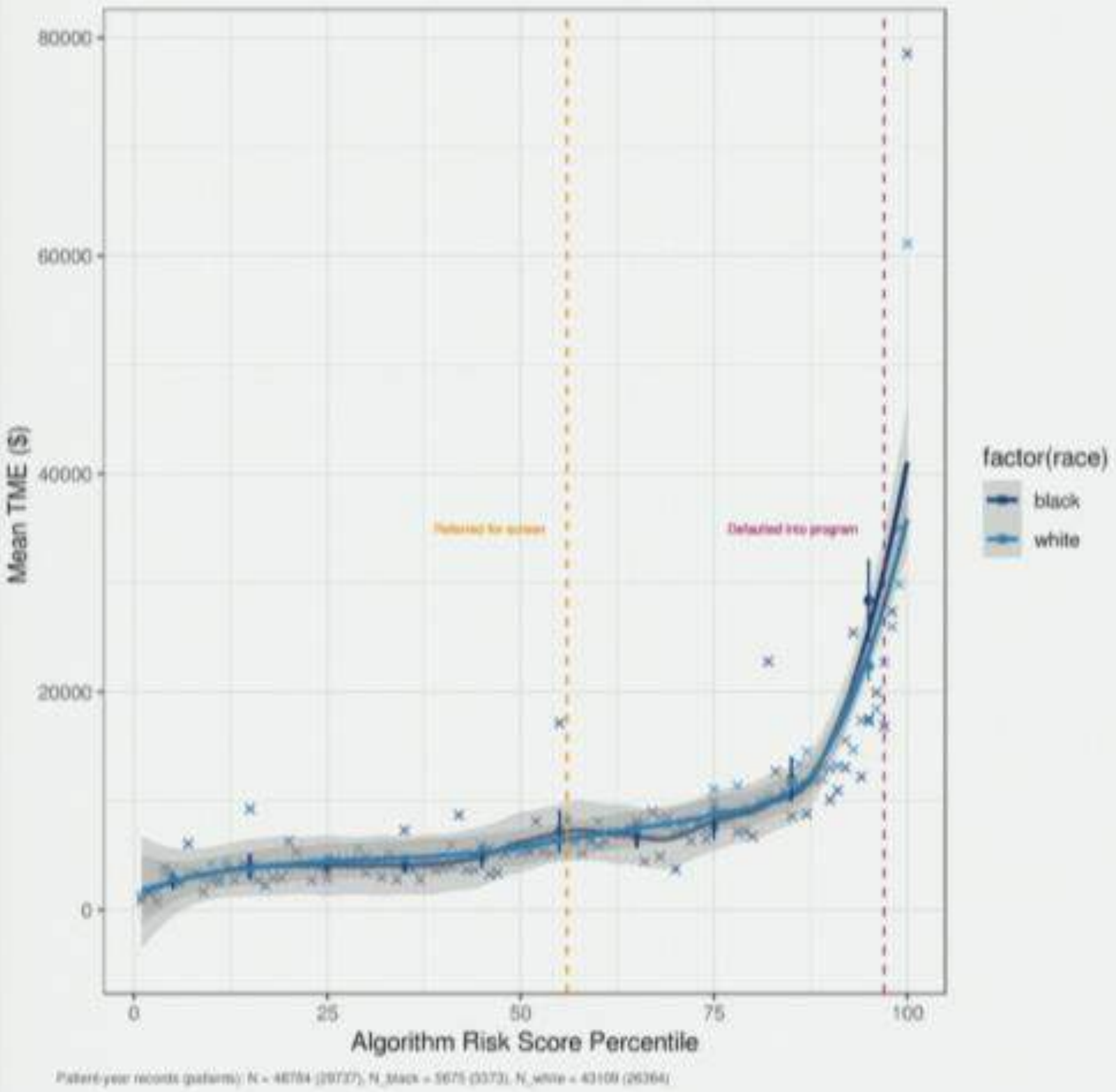
Where is the algorithm going wrong?

A clue is where it's going right:
*unbiased **cost** prediction*

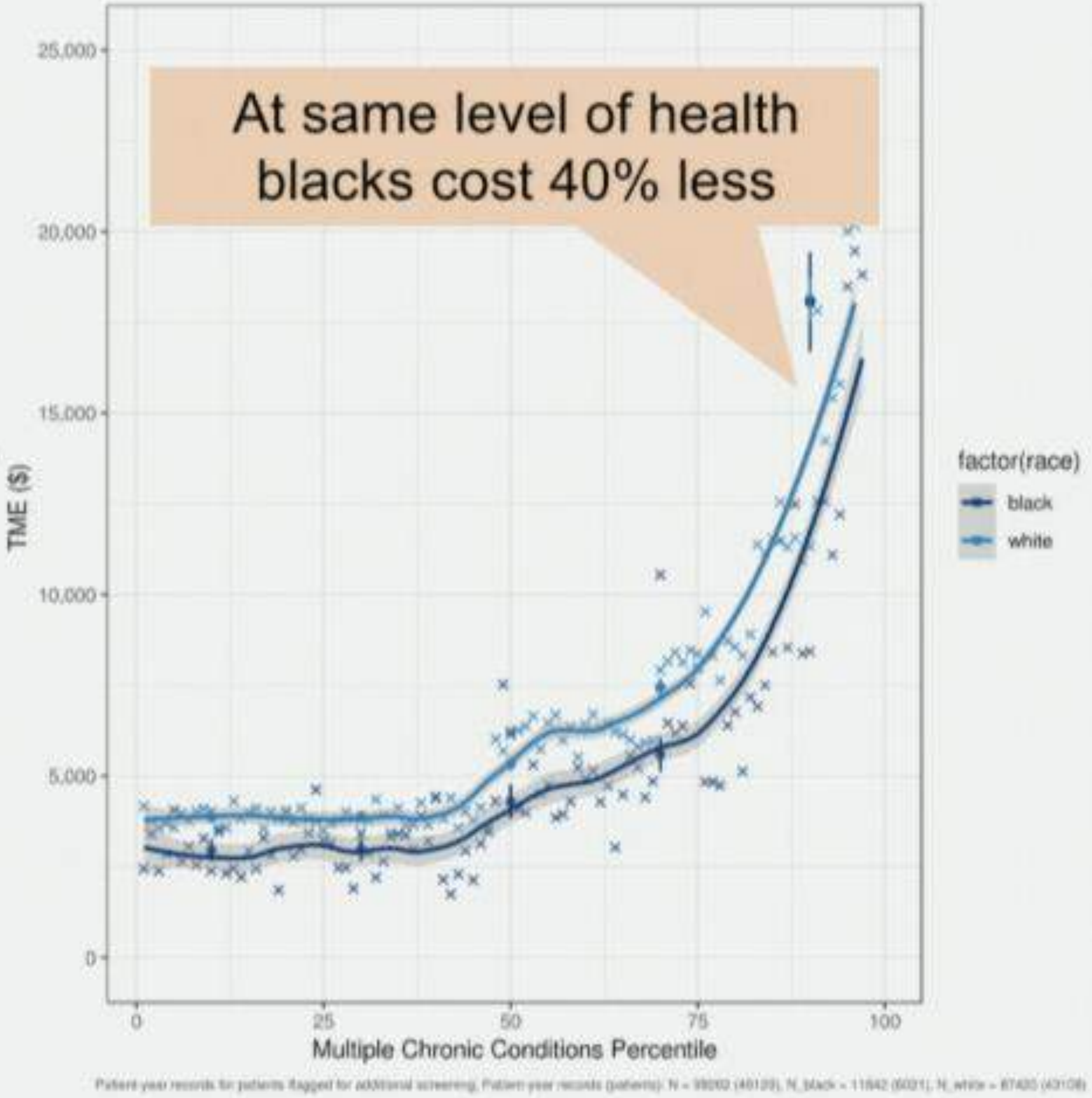


Where is the algorithm going wrong?

*A clue is where it's going right: unbiased **cost** prediction*



*Race differences in cost–health relationship: biased **health** prediction*



Composition of costs also very different

Category	$[B–W]	
Overall cost	*-1808*	***
IP Surgical	-640	***
OP Specialists	-423	***
OP Surgery	-291	***
IP Medical	-286	***
Other	-208	***
Pharmacy	-193	***
Physical therapy	-72	***
Radiology	-47	***
Home health	-30	
OP Primary care	-27	
Laboratory	-20	*
Dialysis	109	***
Emergency	111	***

- Not just less cost
  - Different kinds

Composition of costs also very different

Category	$[B–W]	
Overall cost	*-1808*	***
IP Surgical	-640	***
OP Specialists	-423	***
OP Surgery	-291	***
IP Medical	-286	***
Other	-208	***
Pharmacy	-193	***
Physical therapy	-72	***
Radiology	-47	***
Home health	-30	
OP Primary care	-27	
Laboratory	-20	*
Dialysis	109	***
Emergency	111	***

- Not just less cost
  - Different kinds
- Blacks have
  - Less aggressive care--inpatient, surgical
  - Fewer medications

Composition of costs also very different

Category	$[B–W]	
Overall cost	*-1808*	***
IP Surgical	-640	***
OP Specialists	-423	***
OP Surgery	-291	***
IP Medical	-286	***
Other	-208	***
Pharmacy	-193	***
Physical therapy	-72	***
Radiology	-47	***
Home health	-30	
OP Primary care	-27	
Laboratory	-20	*
Dialysis	109	***
Emergency	111	***

- Not just less cost
  - Different kinds
- Blacks have
  - Less aggressive care--inpatient, surgical
  - Fewer medications
- More emergency visits and dialysis

Possible channels for this effect

Possible channels for this effect

Socioeconomic

- Even with the same insurance:

Possible channels for this effect

Socioeconomic

- Even with the same insurance:

- Financial barriers

Possible channels for this effect

Socioeconomic

- Even with the same insurance:

- Financial barriers

- Logistical barriers
  - Inflexible work schedules
  - Transportation
  - …

Possible channels for this effect

Socioeconomic

- Even with the same insurance:
- Financial barriers
- Logistical barriers
  - Inflexible work schedules
  - Transportation
  - …

Direct effect of race

Alsan et al. (2018), Alsan and Wannamaker (2016):

Possible channels for this effect

Socioeconomic

- Even with the same insurance:
- Financial barriers
- Logistical barriers
  - Inflexible work schedules
  - Transportation
  - …

Direct effect of race

- RCT: assignment to black vs non-black PCP
  - 18% ↑ uptake of preventive care
  - …50-70% ↑ when needles involved

Alsan et al. (2018), Alsan and Wannamaker (2016):

Possible channels for this effect

Socioeconomic

- Even with the same insurance:
- Financial barriers
- Logistical barriers
  - Inflexible work schedules
  - Transportation
  - …

Direct effect of race

- RCT: assignment to black vs non-black PCP
  - 18% ↑ uptake of preventive care
  - …50-70% ↑ when needles involved
- 1SD distance from Tuskeegee post 1972
  - 20-30% ↓ medical care

Alsan et al. (2018), Alsan and Wannamaker (2016):

Dissecting the problem

- Summary so far: Predicting <u>cost</u> accurately means
  - Under-predicting <u>health</u> risk for black patients
  - Healthier whites prioritized before sicker blacks

Dissecting the problem

- Summary so far: Predicting <u>cost</u> accurately means
  - Under-predicting <u>health</u> risk for black patients
  - Healthier whites prioritized before sicker blacks
- Proximal cause
  - The choice of cost as the outcome to predict

Dissecting the problem

- Summary so far: Predicting <u>cost</u> accurately means
  - Under-predicting <u>health</u> risk for black patients
  - Healthier whites prioritized before sicker blacks
- Proximal cause
  - The choice of cost as the outcome to predict
- Deeper cause
  - Why was cost chosen?

Dissecting the problem

- Story 1: Follow the money

Dissecting the problem

- **Story 1: Follow the money**
  - Society cares about health H

Dissecting the problem

- **Story 1: Follow the money**
  - Society cares about health H
  - Hospitals, insurers care about dollars C

Dissecting the problem

- **Story 1: Follow the money**
  - Society cares about health H
  - Hospitals, insurers care about dollars C
  - Many health care problems arise from $(C - H)$
    - Choice of cost is privately optimal
    - But socially suboptimal
  - Another instance of profit distorting health care

A thought experiment

A thought experiment

- A self-interested algorithm developer could just be making a trade-off
  - Willing to tolerate racial bias
  - To get better cost prediction, cost savings

A thought experiment

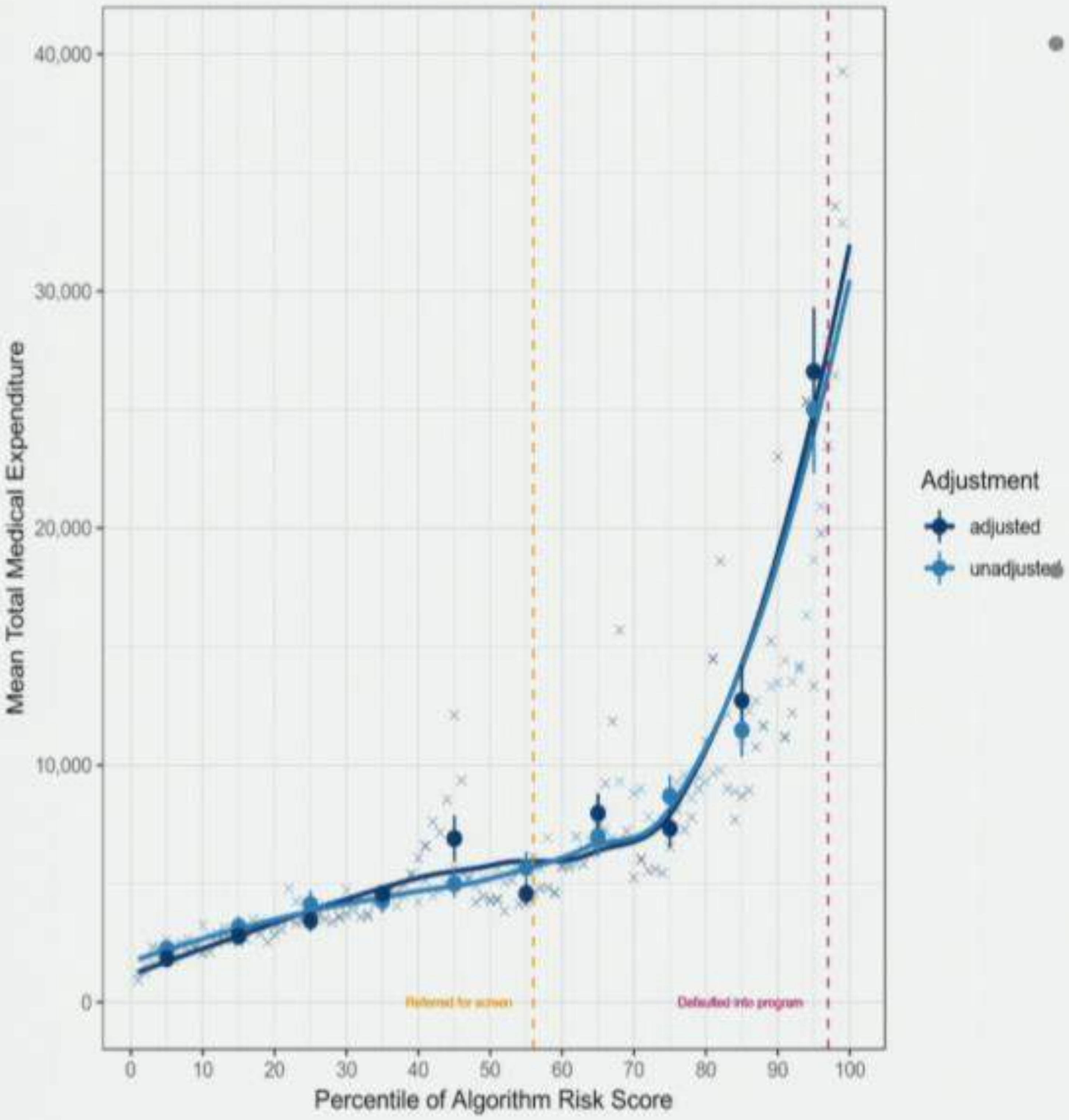
- A self-interested algorithm developer could just be making a trade-off
  - Willing to tolerate racial bias
  - To get better cost prediction, cost savings
- But what if we could construct an algorithm that
  - Achieves these goals just as well
  - Is less biased

Performance of a modified algorithm

Performance of a modified algorithm

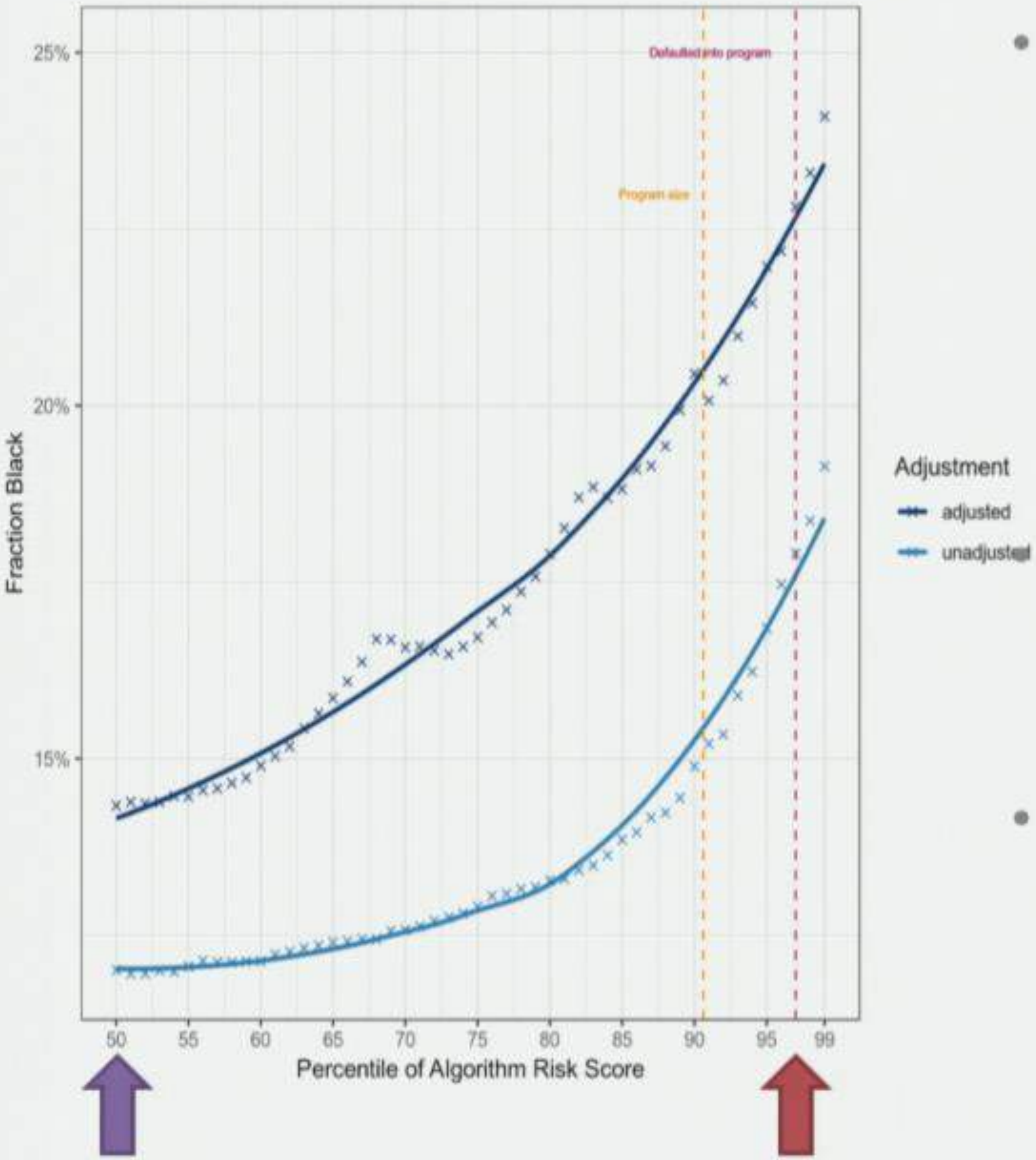
- Procedure
  - Rank by health
  - Then rank by cost
  - Results from holdout

Performance of a modified algorithm



- Procedure
  - Rank by health
  - Then rank by cost
  - Results from holdout
- Still predicts cost
  - R^2: 0.088 vs 0.077

Performance of a modified algorithm



- Procedure
  - Rank by health
  - Then rank by cost
  - Results from holdout
- Still predicts cost
  - R^2: 0.088 vs 0.077
- **Far less biased**
  - At **≥50th** percentile
  - At **97th** percentile: **61.3% ↑ blacks**

What could be going on here?

- Changing the objective function to incorporate health
  - Reductions in bias
  - Without hurting cost prediction

What could be going on here?

- Changing the objective function to incorporate health
  - Reductions in bias
  - Without hurting cost prediction
- Why didn't the algorithm developer do this?

What could be going on here?

- Changing the objective function to incorporate health
  - Reductions in bias
  - Without hurting cost prediction
- Why didn't the algorithm developer do this?
- Why do online algorithms predict ad clicks?

What could be going on here?

What could be going on here?

- We've been thinking of one tradeoff
  - Prediction accuracy vs. racial bias

What could be going on here?

- We've been thinking of one tradeoff
  - Prediction accuracy vs. racial bias
- There is another tradeoff
  - Prediction accuracy vs. cost of accuracy

What could be going on here?

- We've been thinking of one tradeoff
  - Prediction accuracy vs. racial bias
- There is another tradeoff
  - Prediction accuracy vs. cost of accuracy
- Is cost in $t+1$ the most accurate prediction target?
  - Capitalized cost over likely coverage period?
  - Avoidable cost savings from health problems that could be averted by the program?
  - …

Dissecting the problem

- Story 1: Follow the money

Dissecting the problem

- Story 1: Follow the money
- Story 2: Learning

Dissecting the problem

- Story 1: Follow the money
- **Story 2: Learning**
  - Costs C are a convenient proxy for an underlying complex health target H^*
    - This is logical: H^* in general $\rightarrow C$
    - Far easier to measure C than H^*

Dissecting the problem

- Story 1: Follow the money
- **Story 2: Learning**
  – Costs C are a convenient proxy for an underlying complex health target H^*
    - This is logical: H^* in general $\rightarrow C$
    - Far easier to measure C than H^*
  – Cost was chosen because of
    1. Convenience
    2. Not appreciating consequences in terms of bias

This is a generic problem in health and social policy

Passi and Barocas 2019

This is a generic problem in health and social policy

- Health algorithms: complex data x in, label y out
  - X-ray → radiologist's interpretation
  - Medical history → costs

Passi and Barocas 2019

This is a generic problem in health and social policy

- Health algorithms: complex data x in, label y out
  - X-ray $\rightarrow$ radiologist's interpretation
  - Medical history $\rightarrow$ costs
- But 'problem formulation' choices mean that y is often a convenient proxy

Passi and Barocas 2019

This is a generic problem in health and social policy

- Health algorithms: complex data x in, label y out
  - X-ray → radiologist's interpretation
  - Medical history → costs

- But 'problem formulation' choices mean that y is often a convenient proxy

- Can mean automation of errors, biases, inequalities
  - Physicians have incentives and biases
  - Structural inequalities affect diagnosis and utilization

Passi and Barocas 2019

Mis-measurement framework for prediction problems

Mullainathan and Obermeyer, *AER P&P* 2017

Mis-measurement framework for prediction problems

- Target variable: $Y = Y^* + \Delta$

Mullainathan and Obermeyer, *AER P&P* 2017

Mis-measurement framework for prediction problems

- Target variable: $Y = Y^* + \Delta$
- We only observe Y, so we predict

$$\hat{y} \approx E[Y|x] = E[Y^*|x] + E[\Delta|x]$$

Mullainathan and Obermeyer, *AER P&P* 2017

Mis-measurement framework for prediction problems

- Target variable: $Y = Y^* + \Delta$
- We only observe Y, so we predict

$$\hat{y} \approx E[Y|x] = E[Y^*|x] + E[\Delta|x]$$

- Two scenarios

Mullainathan and Obermeyer, *AER P&P* 2017

Mis-measurement framework for prediction problems

- Target variable: $Y = Y^* + \Delta$
- We only observe Y, so we predict

$$\hat{y} \approx E[Y|x] = E[Y^*|x] + E[\Delta|x]$$

- Two scenarios

1) Δ is white noise

Mullainathan and Obermeyer, *AER P&P* 2017

Mis-measurement framework for prediction problems

- Target variable: $Y = Y^* + \Delta$
- We only observe Y, so we predict

$$\hat{y} \approx E[Y|x] = E[Y^*|x] + E[\Delta|x]$$

- Two scenarios

1) Δ is white noise

$$\hat{y} \approx \underbrace{\underbrace{E[Y^*|x]}_{\text{Signal}} + \underbrace{\cancel{E[\Delta|x]}^{\,0}}_{\text{Noise}}}_{\text{Reduce Error}}$$

Mis-measurement framework for prediction problems

- Target variable: $Y = Y^* + \Delta$
- We only observe Y, so we predict

$$\hat{y} \approx E[Y|x] = E[Y^*|x] + E[\Delta|x]$$

- Two scenarios

1) Δ is white noise

$$\hat{y} \approx \underbrace{\underbrace{E[Y^*|x]}_{\text{Signal}} + \underbrace{\cancelto{0}{E[\Delta|x]}}_{\text{Noise}}}_{\text{Reduce Error}}$$

2) Δ is not white noise

Mullainathan and Obermeyer, *AER P&P* 2017

Mis-measurement framework for prediction problems

- Target variable: $Y = Y^* + \Delta$
- We only observe Y, so we predict

$$\hat{y} \approx E[Y|x] = E[Y^*|x] + E[\Delta|x]$$

- Two scenarios

1) Δ is white noise

$$\hat{y} \approx \underbrace{\underbrace{E[Y^*|x]}_{\text{Signal}} + \underbrace{\cancel{E[\Delta|x]}}_{\text{Noise}}}_{\text{Reduce Error}} \to 0$$

2) Δ is not white noise

$$\hat{y} \approx \underbrace{\underbrace{E[Y^*|x]}_{\text{Signal}} + \underbrace{\overbrace{E[\Delta|x]}^{\neq 0}}_{\text{Automate Error}}}_{\text{Ambiguous}}$$

Mullainathan and Obermeyer, *AER P&P* 2017

Mis-measurement framework for prediction problems

- Target variable: $Y = Y^* + \Delta$
- We only observe Y, so we predict

$$\hat{y} \approx E[Y|x] = E[Y^*|x] + E[\Delta|x]$$

- Two scenarios

1) Δ is white noise

$$\hat{y} \approx \underbrace{\underbrace{E[Y^*|x]}_{\text{Signal}} + \underbrace{\cancel{E[\Delta|x]}^{\,0}}_{\text{Noise}}}_{\text{Reduce Error}}$$

2) Δ is not white noise

$$\hat{y} \approx \underbrace{\underbrace{E[Y^*|x]}_{\text{Signal}} + \underbrace{\overbrace{E[\Delta|x]}^{\neq 0}}_{\text{Automate Error}}}_{\text{Ambiguous}}$$

- Predictions depends on balance of $E[Y^*|x]$ and $E[\Delta|x]$

Mullainathan and Obermeyer, *AER P&P* 2017

Why we might worry about this in general

Mullainathan and Obermeyer, *AER P&P* 2017

Why we might worry about this in general

- We often might suspect $Cov[\Delta,x] > Cov[y^*,x]$

Mullainathan and Obermeyer, *AER P&P* 2017

Why we might worry about this in general

- We often might suspect $Cov[\Delta, x] > Cov[y^*, x]$
- In general, y^* is complex
  - Complex biological phenomena
  - Overall health, health needs
  - …

Mullainathan and Obermeyer, *AER P&P* 2017

Why we might worry about this in general

- We often might suspect $Cov[\Delta,x] > Cov[y^*,x]$
- In general, y^* is complex
  - Complex biological phenomena
  - Overall health, health needs
  - …
- Δ can be very simple
  - Biases, errors
  - Often correlated with age, insurance, race, …

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #1: Predicting better may be worse

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #1: Predicting better may be worse

- Evaluating quality of predictions

$$\begin{aligned} E[L(\hat{y}, y)] &\approx E[(\hat{y} - y)^2] \\ &= E[(\hat{y^*} - y^*)^2] + E[(\hat{\Delta} - \Delta)^2] + \\ &\qquad Cov(\hat{y^*} - y^*, \hat{\Delta} - \Delta) \end{aligned}$$

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #1: Predicting better may be worse

- Evaluating quality of predictions

$$
\begin{aligned}
E[L(\hat{y}, y)] &\approx E[(\hat{y} - y)^2] \\
&= E[(\hat{y^*} - y^*)^2] + E[(\hat{\Delta} - \Delta)^2] + \\
&\qquad Cov(\hat{y^*} - y^*, \hat{\Delta} - \Delta)
\end{aligned}
$$

- Imagine algorithms A and B, with $\boldsymbol{L}(\hat{y}^A, y) < \boldsymbol{L}(\hat{y}^B, y)$
  - Is $\hat{y}^A$ predicting better because:

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #1: Predicting better may be worse

- Evaluating quality of predictions

$$E[L(\hat{y},y)] \approx E[(\hat{y}-y)^2]$$
$$= E[(\hat{y^*}-y^*)^2] + E[(\hat{\Delta}-\Delta)^2] +$$
$$Cov(\hat{y^*}-y^*, \hat{\Delta}-\Delta)$$

- Imagine algorithms A and B, with $\boldsymbol{L}(\hat{y}^A,y) < \boldsymbol{L}(\hat{y}^B,y)$
  - Is $\hat{y}^A$ predicting better because:

$$\underbrace{E[(\hat{y^*}-y^*)^2]}_{\uparrow \text{ Signal}}$$

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #1: Predicting better may be worse

- Evaluating quality of predictions

$$
\begin{aligned}
E[L(\hat{y}, y)] &\approx E[(\hat{y} - y)^2] \\
&= E[(\hat{y^*} - y^*)^2] + E[(\hat{\Delta} - \Delta)^2] + \\
&\qquad Cov(\hat{y^*} - y^*, \hat{\Delta} - \Delta)
\end{aligned}
$$

- Imagine algorithms A and B, with $\boldsymbol{L}(\hat{y}^A, y) < \boldsymbol{L}(\hat{y}^B, y)$
  - Is $\hat{y}^A$ predicting better because:

$$
\underbrace{E[(\hat{y^*} - y^*)^2]}_{\uparrow \text{ Signal}} \quad \text{or} \quad \underbrace{E[(\hat{\Delta} - \Delta)^2]}_{\uparrow \text{ Error}}
$$

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #1: Predicting better may be worse

- Evaluating quality of predictions

$$E[L(\hat{y}, y)] \approx E[(\hat{y} - y)^2]$$
$$= E[(\hat{y^*} - y^*)^2] + E[(\hat{\Delta} - \Delta)^2] +$$
$$Cov(\hat{y^*} - y^*, \hat{\Delta} - \Delta)$$

- Imagine algorithms A and B, with $\boldsymbol{L}(\hat{y}^A, y) < \boldsymbol{L}(\hat{y}^B, y)$
  - Is $\hat{y}^A$ predicting better because:

$$\underbrace{E[(\hat{y^*} - y^*)^2]}_{\uparrow \text{Signal}} \quad \text{or} \quad \underbrace{E[(\hat{\Delta} - \Delta)^2]}_{\uparrow \text{Error}}$$

- We may actually prefer 'bad' predictions!
  - Predictors that fit to errors could look good
  - Predictors that get Δ out of $\hat{y}$ could look bad

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #2: This reinforces inequalities

$$\hat{y} \approx \qquad E[Y^*|x] + \underbrace{E[\Delta|x]}_{\neq 0}$$

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #2: This reinforces inequalities

$$\hat{y} \approx \quad E[Y^*|x] + \underbrace{E[\Delta|x]}_{\neq 0}$$

- In ‘prediction policy problems’ we often wish to allocate resources according to $\hat{y}$
  - Diagnostic effort
  - Policies to help high-need patients
  - …

Mullainathan and Obermeyer, *AER P&P* 2017

General insight #2: This reinforces inequalities

$$\hat{y} \approx \qquad E[Y^*|x] + \underbrace{E[\Delta|x]}_{\neq 0}$$

- In ‘prediction policy problems’ we often wish to allocate resources according to $\hat{y}$
  - Diagnostic effort
  - Policies to help high-need patients
  - …
- **This leads us to allocate more to those who have more to begin with**

Mullainathan and Obermeyer, *AER P&P* 2017

Epilogue

Epilogue

- Why not contact the algorithm manufacturer?

Epilogue

- Why not contact the algorithm manufacturer?
- Was super responsive

Epilogue

- Why not contact the algorithm manufacturer?
- Was super responsive
- Replication of results in data on 3.7 million nationwide

Epilogue

- Why not contact the algorithm manufacturer?
- Was super responsive
- Replication of results in data on 3.7 million nationwide
- Wanted to work with us to
  - Understand the problem
  - Try to fix it

Epilogue

Epilogue

- In collaboration with manufacturer, we are developing methods to adjust predictions
  - Predict basket of outcomes H rather than C

Epilogue

- In collaboration with manufacturer, we are developing methods to adjust predictions
  - Predict basket of outcomes H rather than C
- Early results: substantial reductions in bias
  - Original algorithm: 48,772 excess chronic conditions in blacks vs. white, conditional on risk score

Epilogue

- In collaboration with manufacturer, we are developing methods to adjust predictions
  - Predict basket of outcomes H rather than C
- Early results: substantial reductions in bias
  - Original algorithm: 48,772 excess chronic conditions in blacks vs. white, conditional on risk score
  - Adjusted algorithm: 7758 excess chronic conditions – **84% reduction in bias**

Epilogue

- In collaboration with manufacturer, we are developing methods to adjust predictions
  - Predict basket of outcomes H rather than C
- Early results: substantial reductions in bias
  - Original algorithm: 48,772 excess chronic conditions in blacks vs. white, conditional on risk score
  - Adjusted algorithm: 7758 excess chronic conditions – **84% reduction in bias**
- Recognition that C is a proxy for H means we can work to find better proxies (i.e., learning view)

Notice a few things about this modified algorithm

Notice a few things about this modified algorithm

- The training data are the same
  - Did not need to ‘enrich’ with more non-white populations

Notice a few things about this modified algorithm

- The training data are the same
  - Did not need to ‘enrich’ with more non-white populations
- The predictors are the same
  - Did not need to remove race (directly or indirectly)
  - In fact, we would rather have more info on race
- **Only the prediction target (label) changed**

Summary

- Racial bias in a live, scaled commercial algorithm
  - Same risk score: blacks less healthy
  - Less access to extra help (by 2x)
  - Root cause: cost prediction vs. health prediction

Summary

- Racial bias in a live, scaled commercial algorithm
  - Same risk score: blacks less healthy
  - Less access to extra help (by 2x)
  - Root cause: cost prediction vs. health prediction
- Original algorithm reinforced structural inequalities
  - Patients who already use care get more care

Summary

- Racial bias in a live, scaled commercial algorithm
  - Same risk score: blacks less healthy
  - Less access to extra help (by 2x)
  - Root cause: cost prediction vs. health prediction
- Original algorithm reinforced structural inequalities
  - Patients who already use care get more care
- Modified algorithm does not solve these problems
  - But it is built on understanding them
  - And works to reverse them