Overcooked models

Mixing prediction, explanation, confounders, and mediators

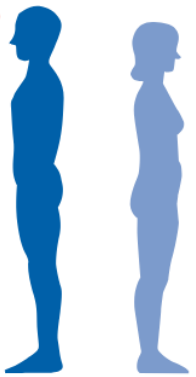
Travis Gerke, ScD
@travisgerke

MOFFITT CANCER CENTER

Motivating example: prostate cancer

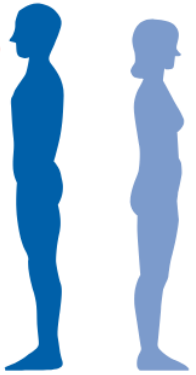
Estimated New Cases

Male			Female		
Prostate	174,650	20%	Breast	268,600	30%
Lung & bronchus	116,440	13%	Lung & bronchus	111,710	13%
Colon & rectum	78,500	9%	Colon & rectum	67,100	7%
Urinary bladder	61,700	7%	Uterine corpus	61,880	7%
Melanoma of the skin	57,220	7%	Melanoma of the skin	39,260	5%
Kidney & renal pelvis	44,120	5%	Thyroid	37,810	4%
Non-Hodgkin lymphoma	41,090	5%	Non-Hodgkin lymphoma	33,110	4%
Oral cavity & pharynx	38,140	4%	Kidney & renal pelvis	29,700	3%
Leukemia	35,920	4%	Pancreas	26,830	3%
Pancreas	29,940	3%	Leukemia	25,860	3%
All sites	**870,970**		**All sites**	**891,480**	



Estimated Deaths

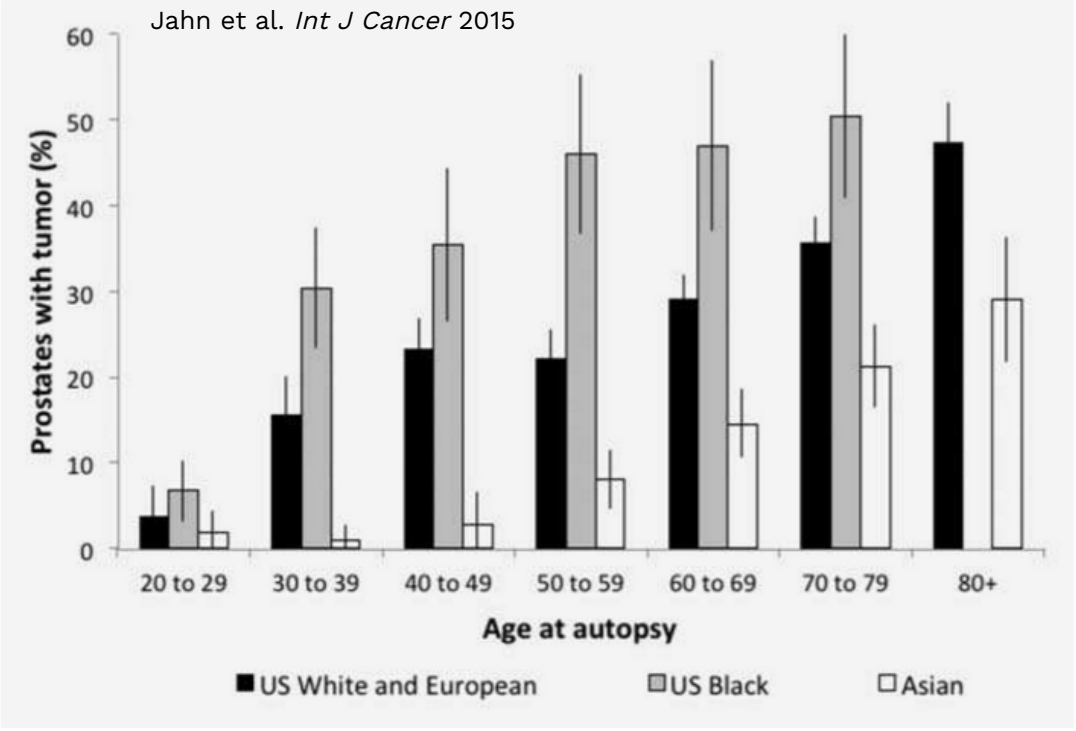
Male			Female		
Lung & bronchus	76,650	24%	Lung & bronchus	66,020	23%
Prostate	31,620	10%	Breast	41,760	15%
Colon & rectum	27,640	9%	Colon & rectum	23,380	8%
Pancreas	23,800	7%	Pancreas	21,950	8%
Liver & intrahepatic bile duct	21,600	7%	Ovary	13,980	5%
Leukemia	13,150	4%	Uterine corpus	12,160	4%
Esophagus	13,020	4%	Liver & intrahepatic bile duct	10,180	4%
Urinary bladder	12,870	4%	Leukemia	9,690	3%
Non-Hodgkin lymphoma	11,510	4%	Non-Hodgkin lymphoma	8,460	3%
Brain & other nervous system	9,910	3%	Brain & other nervous system	7,850	3%
All sites	**321,670**		**All sites**	**285,210**	



Estimates are rounded to the nearest 10, and cases exclude basal cell and squamous cell skin cancers and in situ carcinoma except urinary bladder. Estimates do not include Puerto Rico or other US territories. Ranking is based on modeled projections and may differ from the most recent observed data.

More men die with prostate cancer than from it

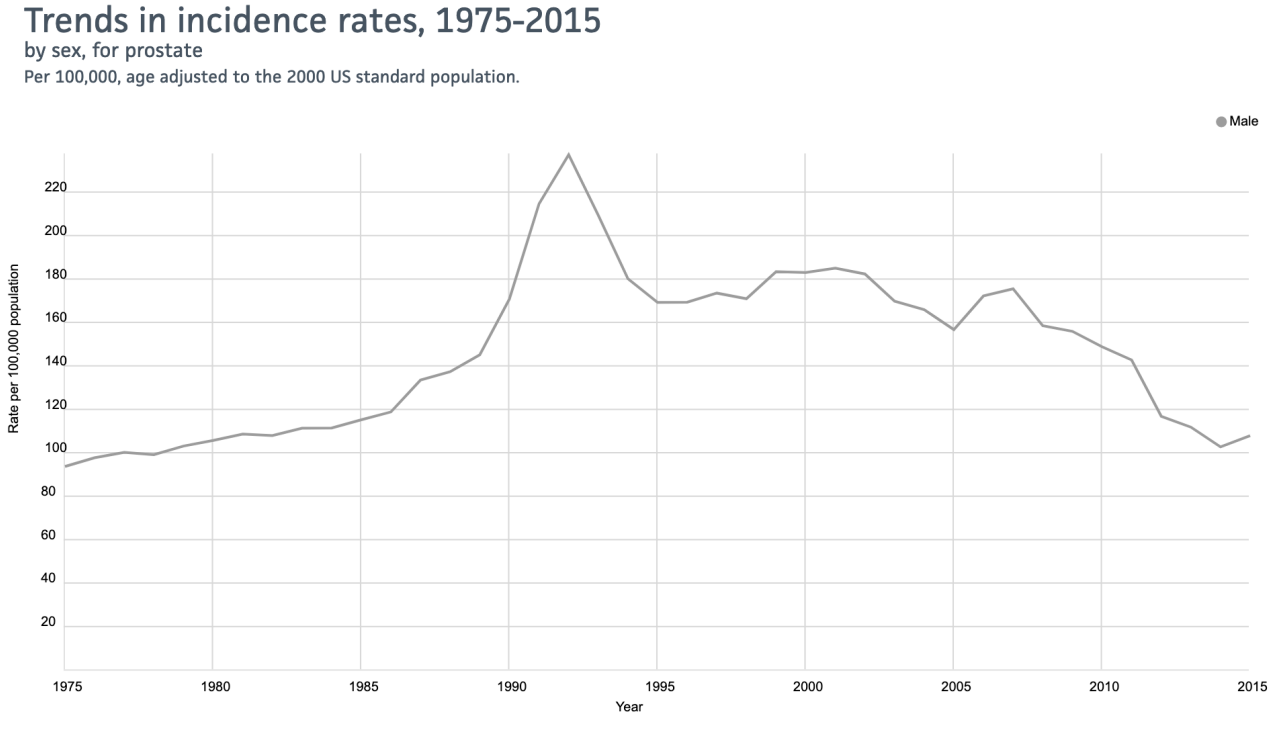
- 5-year survival ≈ 98% and <10% of prostate cancer patients have fatal disease
- 2.9 million men living with a diagnosis in US
  - 42 million latent (undiagnosed) cases!



PSA screening trends determine cancer incidence

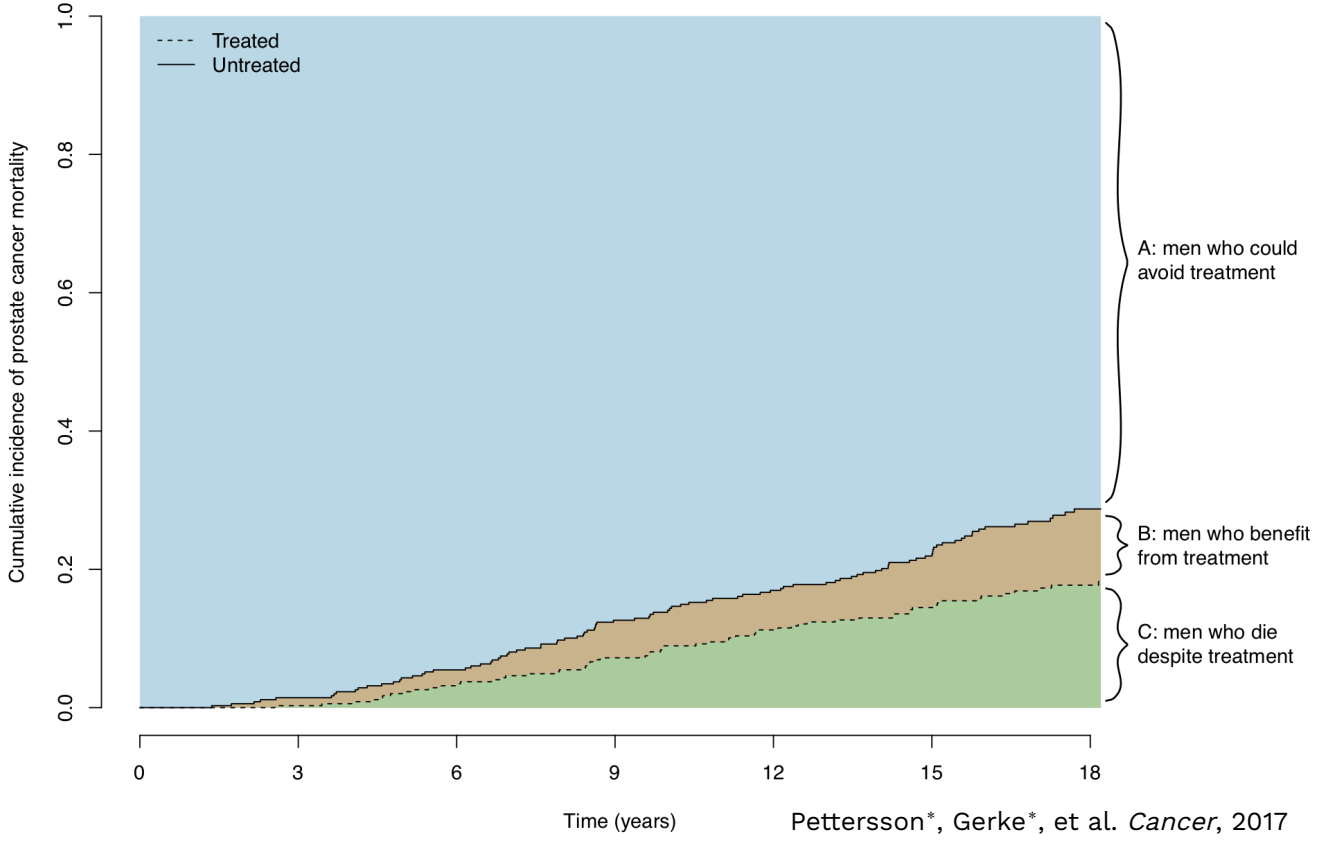
- PSA screening ⟶ high rates of overdiagnosis and overtreatment
  - Up to 2/3 of prostate cancers are overdiagnosed, most are treated[1]

1. Loeb et al. *Eur Urol* 2014



An urgent clinical challenge

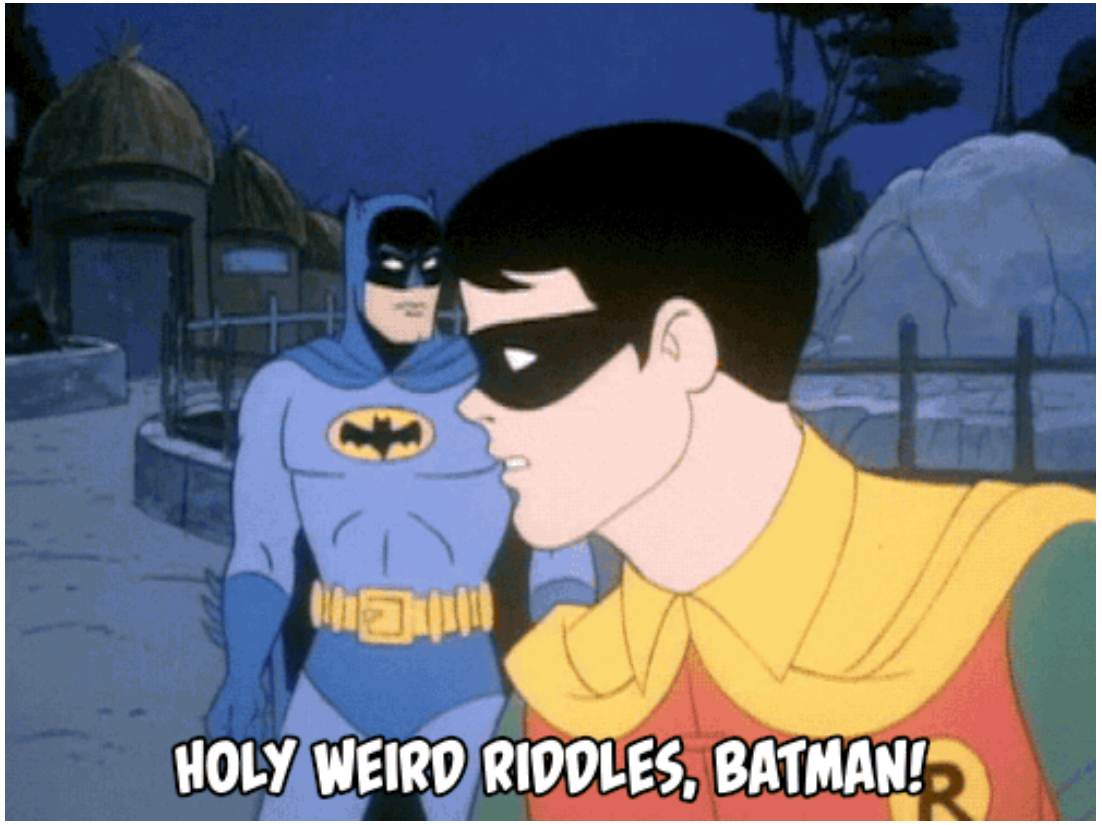
- How can we distinguish indolent from lethal disease?
  - Patients with indolent tumors could avoid overtreatment
  - Patients with potentially lethal tumors could receive timely treatment



Pettersson*, Gerke*, et al. *Cancer*, 2017

A different(???) clinical challenge

- What are the causes of lethal disease?
  - If we know these, won't we also know the answer to the previous question, "How can we distinguish indolent from lethal disease?"



It's complicated.

Good prediction models $\neq$ good causal models

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the

Explanatory modeling = causal inference

- Test causal hypotheses for mechanistic understanding
  - Randomized experiments/trials are a gold standard
  - Increasingly, causal inference is conducted by evaluating association patterns within observational data according to specific rules
  - Success: an understandable statistical model (e.g. regression) that fits data well according to an expert-guided mechanistic theory

Predictive modeling = predicting future events

- Models that use input values to accurately predict future outputs
  - Study design includes training and validation data sets
  - Success: a model built in the training data which need not be easily interpretable (e.g. neural net) works well in the validation data

Proof that good prediction $\neq$ good explanation

- TL;DR version: prediction error is a tradeoff between bias and variance. You can use a biased model (in a causal sense) that has low variance to reduce error

APPENDIX: IS THE "TRUE" MODEL THE BEST PREDICTIVE MODEL? A LINEAR REGRESSION EXAMPLE

Consider $\mathcal{F}$ to be the true function relating constructs $\mathcal{X}$ and $\mathcal{Y}$ and let us assume that f is a valid operationalization of $\mathcal{F}$. Choosing an intentionally biased function f^* in place of f is clearly undesirable from a theoretical–explanatory point of view. However, we will show that f^* can be preferable to f from a predictive standpoint.

To illustrate this, consider the statistical model $f(x) = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ which is assumed to be correctly specified with respect to $\mathcal{F}$. Using data, we obtain the estimated model $\hat{f}$, which has the properties

$$\text{Bias} = 0, \tag{2}$$

$$\begin{aligned} \text{Var}(\hat{f}(x)) &= \text{Var}(x_1\hat{\beta}_1 + x_2\hat{\beta}_2) \\ &= \sigma^2 x'(X'X)^{-1}x, \end{aligned} \tag{3}$$

where x is the vector $x = [x_1, x_2]'$, and X is the de-

fied model that leaves out q predictors has a lower EPE when the following inequality holds:

$$q\sigma^2 > \beta_2' X_2'(I - H_1)X_2\beta_2. \tag{6}$$

This means that the underspecified model produces more accurate predictions, in terms of lower EPE, in the following situations:

- when the data are very noisy (large σ);
- when the true absolute values of the left-out parameters (in our example β_2) are small;
- when the predictors are highly correlated; and
- when the sample size is small or the range of left-out variables is small.

The bottom line is nicely summarized by Hagerty and Srinivasan (1991): "We note that the practice in applied research of concluding that a model with a higher predictive validity is "truer," is not a valid inference. This paper shows that a parsimonious but less true model can have a higher predictive validity than a truer but less parsimonious model."

Simpler for good prediction $\neq$ good explanation

- You can build good prediction models with variables that have nothing to do with mechanism

CRIME | A BLOG ABOUT MURDER, THEFT, AND OTHER WICKEDNESS. | JULY 9 2013 2:59 PM

When Ice Cream Sales Rise, So Do Homicides. Coincidence, or Will Your Next Cone Murder You?

By Justin Peters

614 49 0

*Crime is **Slate**'s crime blog. Like us on **Facebook**, and follow us on Twitter **@slatecrime**.*

The New Orleans *Times-Picayune* **ran a piece last Friday** attempting to answer a question the entire world has been asking: Should ice cream be blamed for murders? "The

Slate

The Erupting Earth, the Wheeling Heavens

Slate Money on Amazon's "Local Register" and Listener Questions

SPONSORED CONTENT

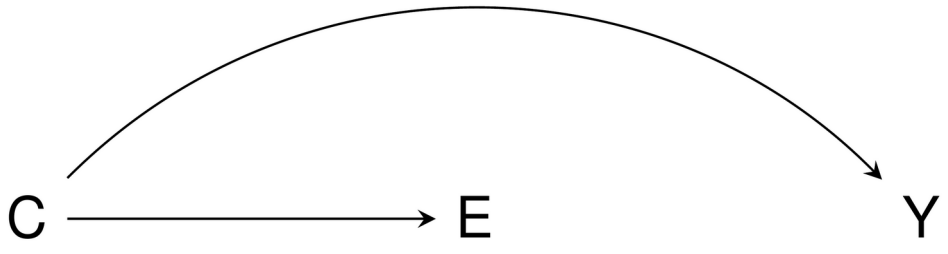
When Patagonia Says 'Don't Buy This Jacket,' Here's Why Everyone Buys That Jacket

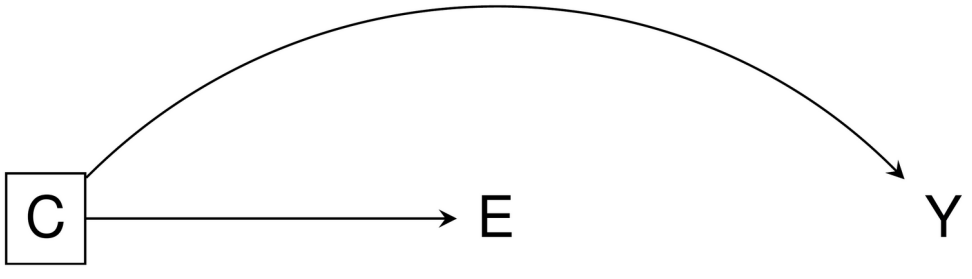
How Obama Can Leave a Meaningful Foreign Policy Legacy With the Time He Has Left

Correlation is not causation, except when it is

- The previous slide gave a good example of non-causative correlation
  - But I also just said we can use associational patterns to infer causation
  - So, which is it?

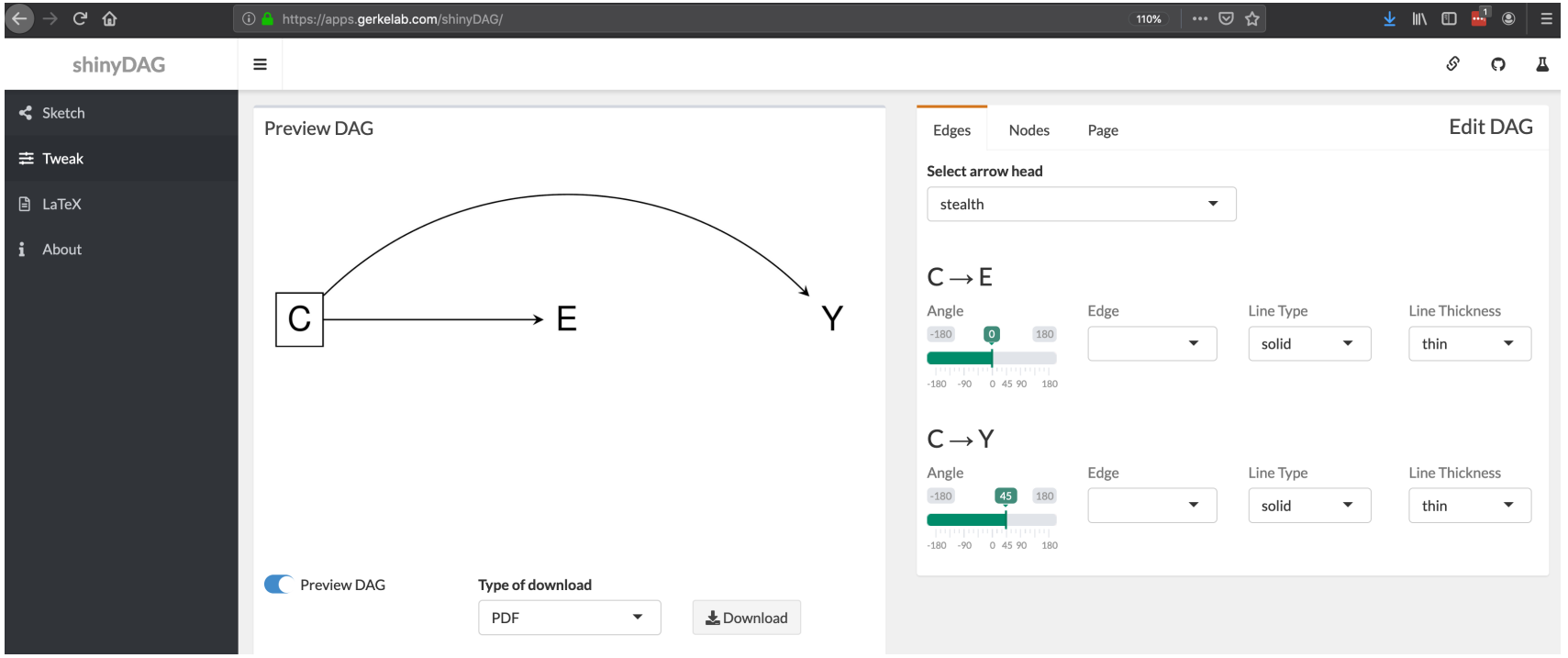
- Directed acyclic graphs (DAGs) help understand when correlation == causation
  - The most basic rule is that association flows through edges
  - When two nodes are connected, we observe a statistical association
  - When association persists when all spurious edges between two variables are blocked, the observed association is causal





DAGs: there's an app for that

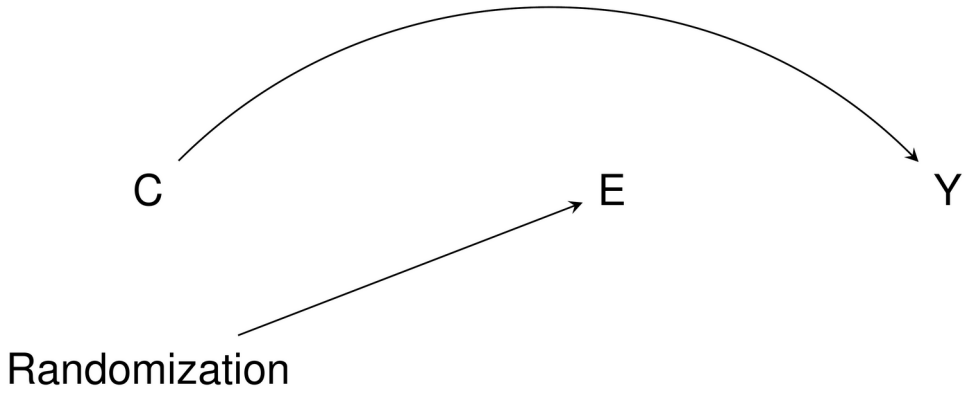
- There are rules governing how association flows beyond the scope of this talk
  - https://apps.gerkelab.com/shinyDAG/



Example: Randomized trials as a gold stardard

- Randomized experiments provide causal effect estimates
  - Here's a DAG for an RCT under the null
  - Boilerplate explanation: "Because randomization adjusts for all confounders"

- *Very important definition:* A confounder is a common cause of exposure and outcome



Confounding

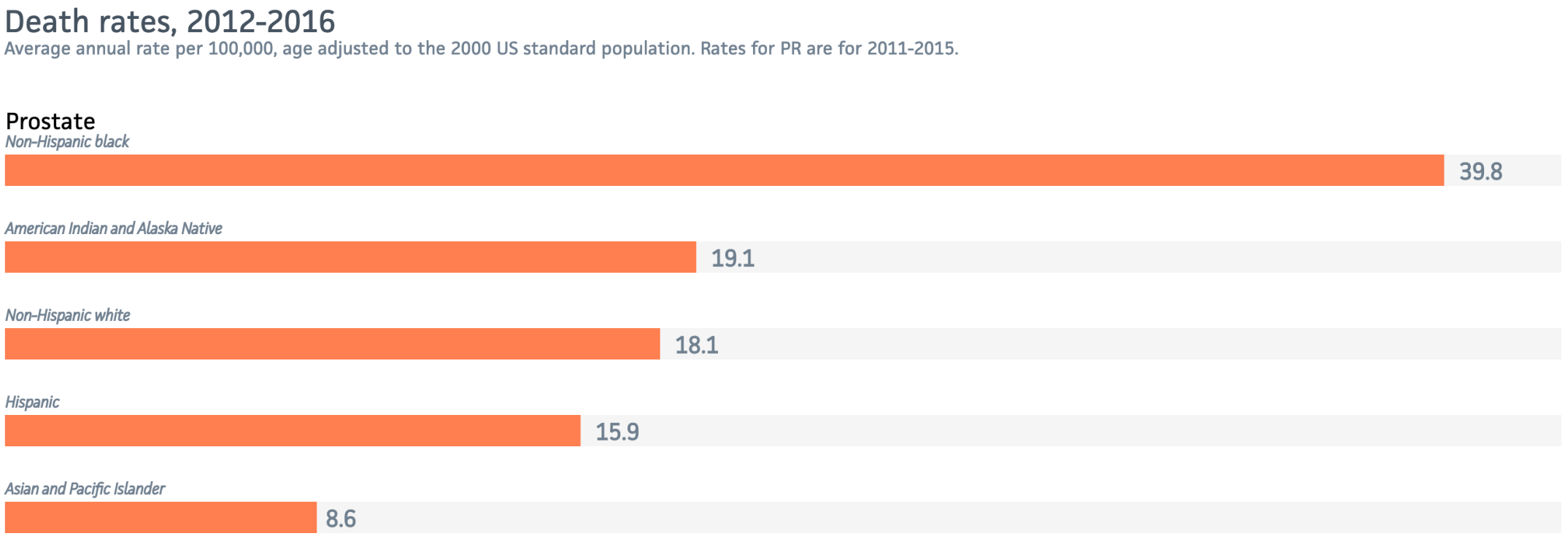
From Wikipedia, the free encyclopedia

"Confounding factor" redirects here. For the defunct British video games company, see Confounding Factor (games company).

In statistics, a **confounding variable** (also **confounding factor**, a **confound**, a **lurking variable** or a **confounder**) is a variable in a statistical model that correlates (directly or inversely) with both the dependent variable and an independent variable, in a way that "explains away" some or all of the correlation between these two variables.

NO!

Back to the motivating example: PCa disparities



Data Source: National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention, 2018

CancerStatisticsCenter.cancer.org

An artifact? "Confounding"?

https://www.phillyvoice.com/black-men-not-more-likely-die-prostate-cancer/

VOICE

There's actually *not* a racial disparity when it comes to prostate cancer deaths, study finds

The reason seems to be tied to access, not genetics, as researchers debunk medical myth

BY **BAILEY KING**
PhillyVoice Staff

MEN'S HEALTH Prostate Cancer

JAMA Oncology | Search All | Enter Search Term

Original Investigation

May 23, 2019

Association of Black Race With Prostate Cancer-Specific and Other-Cause Mortality

Robert T. Dess, MD[1]; Holly E. Hartman, MS[2]; Brandon A. Mahal, MD[3]; et al

Author Affiliations | Article Information

JAMA Oncol. 2019;5(7):975-983. doi:10.1001/jamaoncol.2019.0826

NEWS

Black men not at higher risk of dying from prostate cancer, study finds

SHARE THIS –

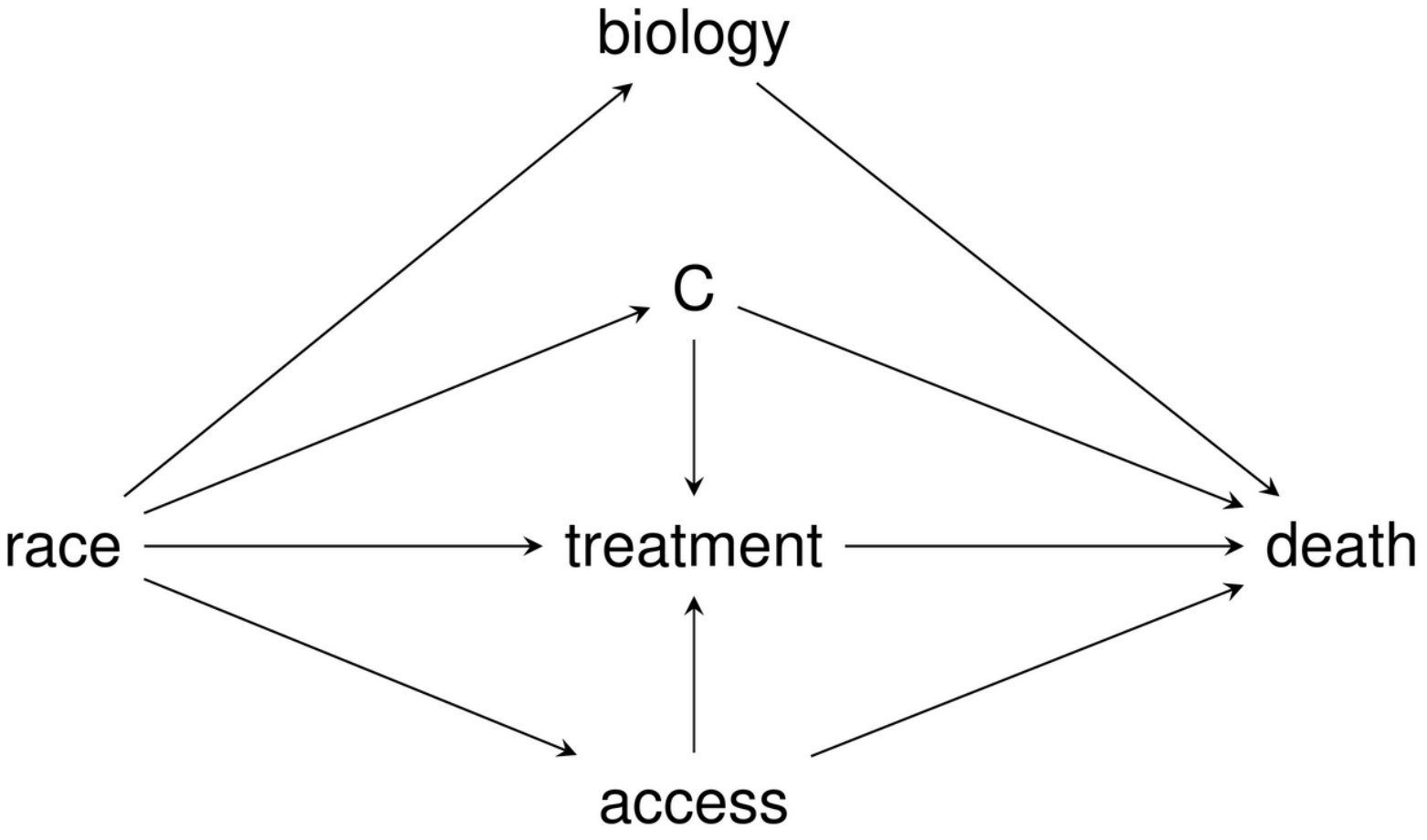
MEN'S HEALTH

Black men not at higher risk of dying from prostate cancer, study finds

The new study debunks the myth that genetics plays a larger role than health disparities in risk of death from prostate cancer for black men.

Let the DAG help us know what to do!

- Are there confounders of race?



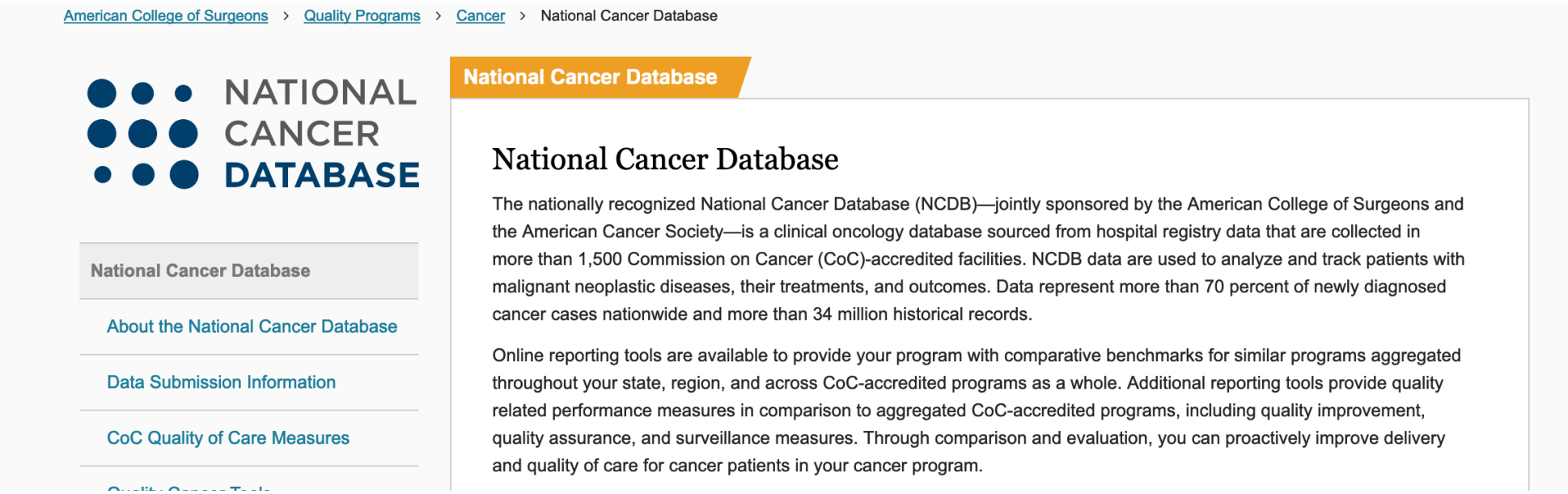
Oops.

JAMA Oncology | Search All | Enter Search Term

Second, our approach highlights the challenges of interpreting population-based data.[24] We adjusted for age, insurance, and a newly released validated socioeconomic status variable. Moreover, we adjusted for cancer- and treatment-related confounders, including the newly released quality-assured PSA values, which were a significant limitation in prior SEER analyses.[25] Inclusion of these crucial prognostic factors substantially decreased the estimated age-only PCSM hazard for black men, but we still came to a slightly different conclusion with respect to the unexplained significant association of black race when using SEER population data compared with the VA and RCT cohorts. Residual group imbalances,[26] unmeasured confounders,[5] cause of death attribution bias,[27] and issues with coding treatment-related variables[28] may all contribute to the difficulty in interpreting outcomes from population registries.[6] We included several cohorts for comparison with explicit stepwise adjustment, and the overall findings provide evidence against an increased biological risk

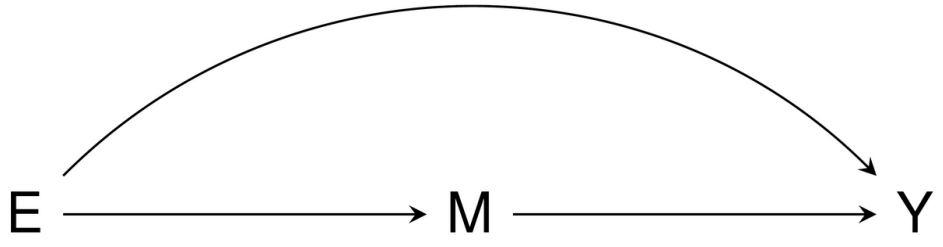
Let's try again

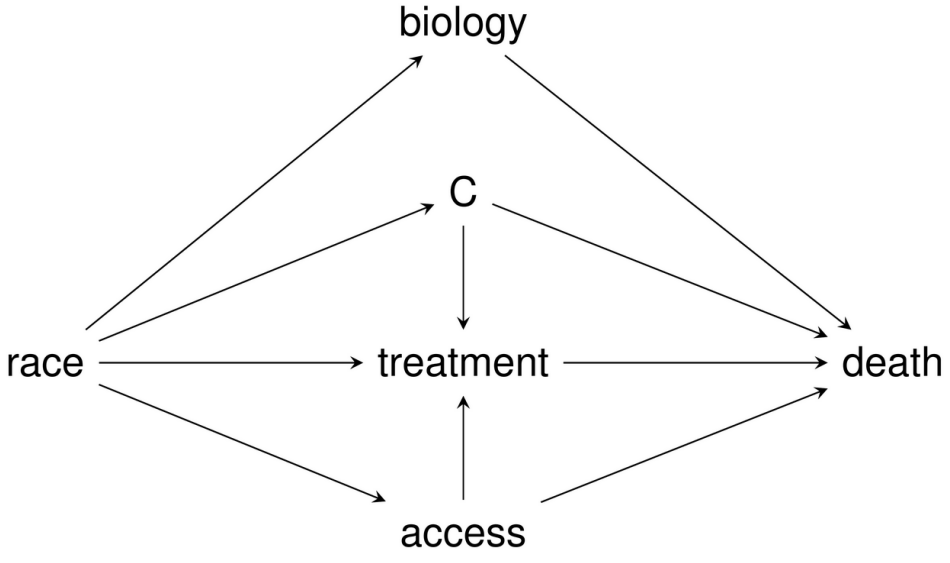
- 1,380,357 prostate cancer patients in NCDB
  - After subsetting to those with complete follow-up, created a density-matched 1:1 case-control sample of 12,256 patients
  - Crude analysis by race: HR for death of 1.30 (95% CI: 1.18-1.44) comparing black to white patients



How to properly think about the role of access

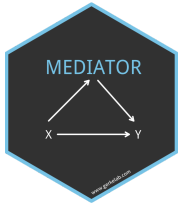
- Another *very important definition*: A mediator is a variable caused by exposure, which in turn, causes the outcome
  - Effects can then be separated into *direct* and *indirect* components
  - Methods for mediation analyses are a distinct branch of statistics
  - Recommended resource: VanderWeele 2015 *Explanation in Causal Inference*
  - For implementation in R, https://github.com/GerkeLab/mediator





Effect of race is minimally mediated by access

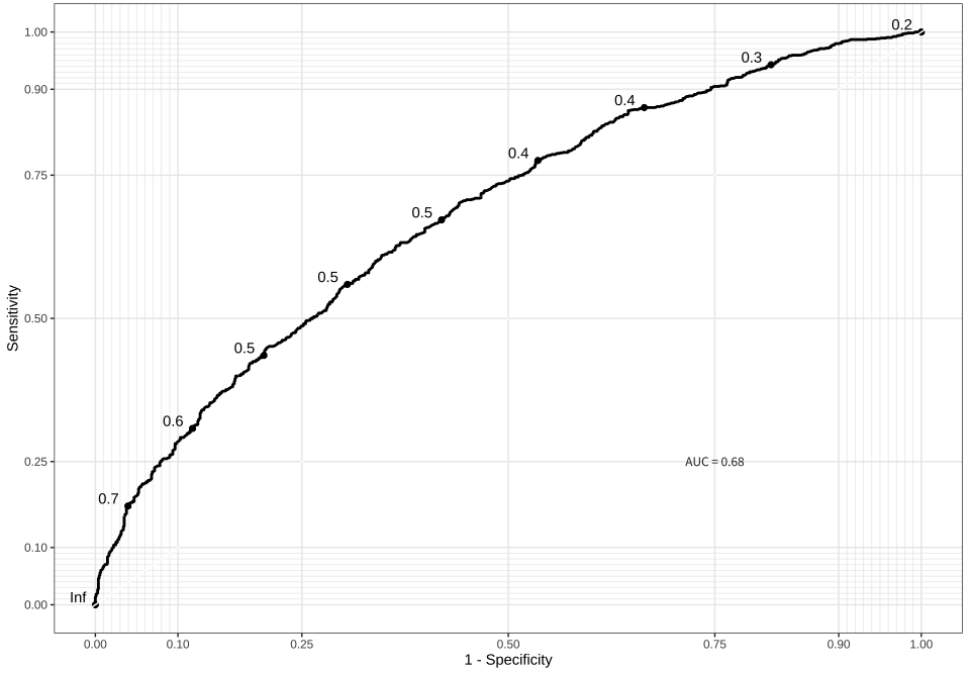
- One way to measure this is through insurance status
  - The proportion of race effect mediated through insurance status $\approx 1\%$
  - We needed to adjust for many mediator-outcome confounders
  - [Hand-wave over other nuances 👋]
  - [Can discuss these at conclusion]



term	estimate	std.error	statistic	p.value	conf.low	conf.high
Race == “Black”TRUE	1.4768801	0.0571501	6.8229446	0.0000000	1.3205646	1.6522114
AgeCat(65,75]	1.7866940	0.0447925	12.9567952	0.0000000	1.6366709	1.9508316
AgeCat[34,55]	0.6888132	0.0570961	-6.5290844	0.0000000	0.6156888	0.7701460
AgeCat>75	3.2186762	0.1483401	7.8803391	0.0000000	2.4199028	4.3323936
Stage2	1.0501864	0.0948629	0.5161939	0.6057190	0.8717953	1.2646338
Stage2A	1.0080892	0.0550367	0.1463864	0.8836164	0.9049648	1.1228805
Stage2B	1.1705871	0.0669021	2.3542662	0.0185593	1.0266501	1.3345302
Stage3	1.2994194	0.1352738	1.9362036	0.0528428	0.9981476	1.6969694
Stage4	3.8001621	0.3032038	4.4031237	0.0000107	2.1619211	7.1581361
Stage99	1.3476367	0.0956530	3.1191111	0.0018140	1.1174639	1.6260412
Gleason2	1.2021596	0.0535434	3.4386990	0.0005845	1.0824086	1.3352124
Gleason3	1.3872915	0.0672614	4.8668840	0.0000011	1.2160003	1.5828958
Gleason4	1.5814921	0.0760083	6.0305127	0.0000000	1.3628966	1.8360285
Gleason5	3.1912406	0.0910467	12.7452147	0.0000000	2.6729352	3.8197414
PSA(6,10]	0.8167643	0.0677657	-2.9868319	0.0028188	0.7150437	0.9326330
PSA[0,6]	0.6670257	0.0632058	-6.4064832	0.0000000	0.5891813	0.7548597
PSA>20	0.8580002	0.0922104	-1.6608860	0.0967363	0.7161881	1.0280983
TTT	0.9996285	0.0003610	-1.0293232	0.3033278	0.9989128	1.0003291
FacilityCommunity Cancer Program	1.5131036	0.0960230	4.3131638	0.0000161	1.2542179	1.8277359
FacilityComprehensive Community Cancer Program	1.2007048	0.0436496	4.1903905	0.0000278	1.1022706	1.3079789
FacilityIntegrated Network Cancer Program	1.1371362	0.0671733	1.9131569	0.0557280	0.9968091	1.2971361
InsuranceNot Insured	1.5692756	0.1505921	2.9922834	0.0027690	1.1695347	2.1121887
SurgeryOther/Unk	1.0759997	0.6024995	0.1215772	0.9032339	0.3425714	3.7883261
SurgeryProstatectomy	0.3938498	0.3444875	-2.7048463	0.0068336	0.1907853	0.7454435
RadiationRadiaiton	1.3407091	0.0800677	3.6618858	0.0002504	1.1465815	1.5694806
RadiationUnknown	1.1321455	0.2391102	0.5190684	0.6037130	0.7093763	1.8170487

The published effort is still useful

- Turns out, it's a decent predictive model!
  - The below is AUC on a 20% hold-out validation data set from a simple logistic regression with all included factors
  - And we be even better if we used more modern machine learning



The real heroes

Jordan Creed

Jordan.Creed@moffitt.org

jhcreed

Garrick Aden-Buie

Garrick.Aden-Buie@moffitt.org

gadenbuie

Further details

- https://www.gerkelab.com/
- https://github.com/gerkelab