Case study

William Lowe

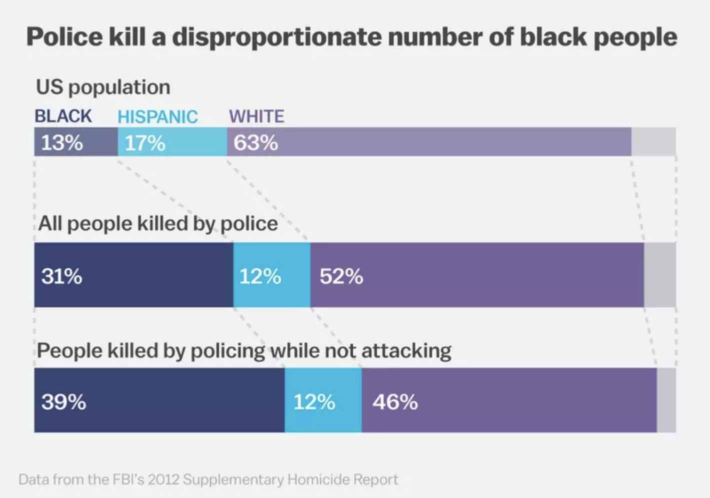
Hertie School

22nd November 2020

Plan

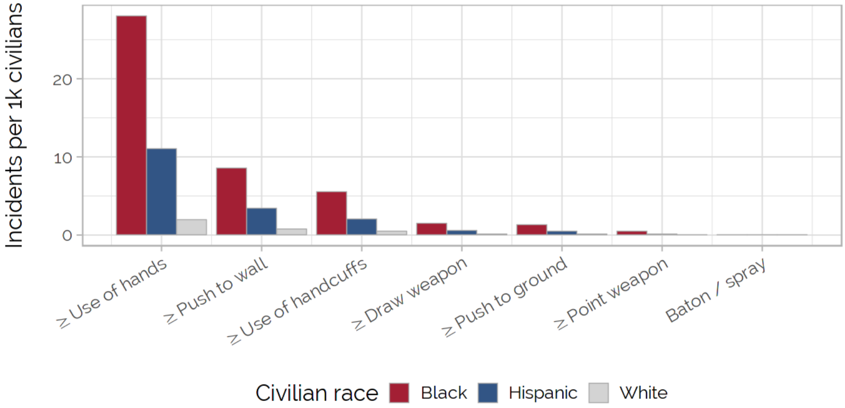
- → A policy problem in the US
- → The causal inference of the policy problem
- → What is the effect of race?
- → What can be estimated?
- → What should be estimated?
- → A minimal approach (Knox et al., 2020a)
- → Principal stratification, assumptions, results

Police shootings in the US



Alvin Chang/Vox

NYPD STOP AND FRISK, 2003-2013



Evidence-based policy

Evidence from causal inference is necessary for

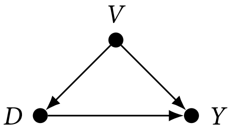
- realizing there is a problem at all (looking at you Wall Street Journal editorial page)
- Doing something effective about it (implicit-bias training, scenario training, etc.)

Aside: America is an interesting place to study this

- Policing is organized very locally, so lots of variation in techniques, training, etc.
- Lots of variation in outcomes
- Much interest in data collection (body and surveillance cameras, Stingrays, Shotspotter) and data-oriented solutions ('predictive policing')
- Sometimes very revealing data, e.g. vehicle stop transcripts

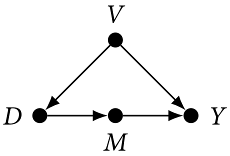
Graphs

Too simply



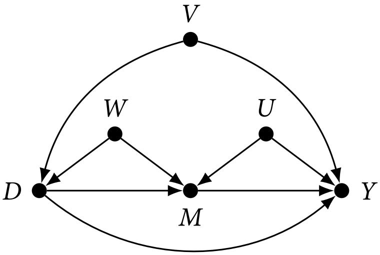
- D Race
- Y Force
- V Confounders

Slightly more realistically



- M Police stop

(Nearly) all the things that could go wrong



- U, W Confounders
- Direct effect $D \longrightarrow Y$

Wait what?

How is there an arrow going into race (D)?

- → Our unit of analysis is the *encounter* or *sighting* involving a person of some race
- → Not the person
- → Scenario: Officer sees person, then decides whether to stop them

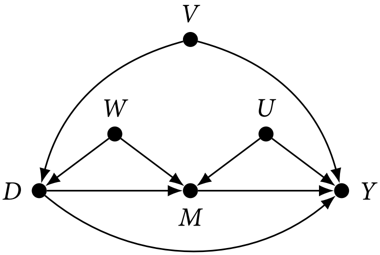
This is reflected in police data which is organised by stop, not suspect

- → Sidestep issues about the *manipulability* of race
- → We can manipulate race by (experimentally, even) switching in a similarly situated person of a different race into the encounter

Arrow into D means: factors that change the balance of race across encounters, e.g. neighbourhood indicators

Nearly all the things

(Nearly) all the things that could go wrong



What is the causal effect of race?

It's often argued that

→ race (gender etc.) are fundamentally *not manipulable* (Kohler-Hausmann, 2019)
→ non-manipulable variables cannot be causes, because they have no well-defined counterfactuals (Holland, 2003)

What is the causal effect of race?

It's often argued that

→ race (gender etc.) are fundamentally *not manipulable* (Kohler-Hausmann, 2019)
→ non-manipulable variables cannot be causes, because they have no well-defined counterfactuals (Holland, 2003)

And argued back either that

→ That doesn't matter because the *correlates* of race, e.g. name, dress, accent, etc. are manipulable (Bertrand & Mullainathan, 2004; Greiner & Rubin, 2010)
→ there may not be much more to race than this anyway (Sen & Wasow, 2016)
→ it wouldn't be a problem if there were (Pearl, 2018)
→ the problem is misidentified as ontological (VanderWeele & Hernán, 2012)

(gender provides a natural comparison case for these responses)

A bundle of sticks

Operationalization of race	"Immutable characteristics"	"Bundle of sticks"
Underlying theory	Essentialist	Constructivist
Race manipulable?	No, race is an immutable characteristic	Yes, race contains mutable and manipulable elements
Always post-treatment bias?	Yes, race is assigned at conception	No, some constitutive elements of race are assigned after conception
Race unstable?	No, race is homogenous and measurable	Yes, race demands disaggregation
Measurement?	Race is typically coded as a binary or categorical variable	Race is a composite variable in which an element of race is the key variable and determines coding

From (Sen & Wasow, 2016)

What is the causal effect of race?

The opposite view is potentially difficult:

- → Strong essentialism (‘essence’ vs. ‘accident’) last popular in the medieval period
- → Switching race would be (effectively) a ‘transformative treatment’ (Paul & Healy, 2018)

(psychotherapy or alcoholism treatment provide natural comparison cases for Paul and Healy’s response)

This apparently abstruse theoretical questions matters for policy

What is the causal effect of race?

The opposite view is potentially difficult:

- → Strong essentialism ('essence' vs. 'accident') last popular in the medieval period
- → Switching race would be (effectively) a 'transformative treatment' (Paul & Healy, 2018)

(psychotherapy or alcoholism treatment provide natural comparison cases for Paul and Healy's response)

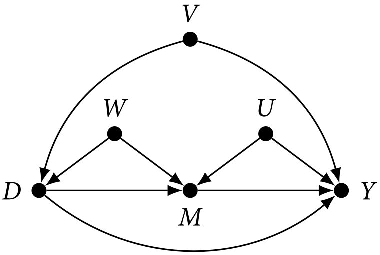
This apparently abstruse theoretical questions matters for policy

Aside:

- → If race, gender, etc. are protected characteristics, then so much for counterfactual theories of fairness as a way of dealing with them

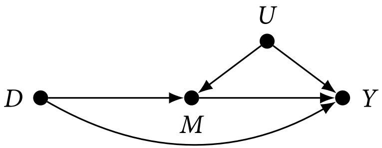
Nearly all the things

(Nearly) all the things that could go wrong

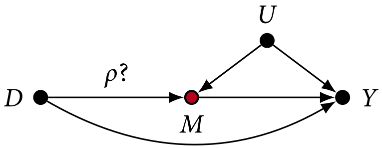


Even in the best case scenario

Assume we can *measure and control for* all these confounders



However, our data conditions on M



This is a mediation problem:

- → Direct effect: Conditional on being stopped (M=1), race (D) affects use of force (Y)
- → Indirect effect: Force is only applied in stops (M=1)
- → Interaction: M=1 implies Y=0 (but not vice versa)

This is a missing data problem:

- → M is a missing data indicator. If $M = 1$ we get to see the case, otherwise not

Lots of collider bias potential…

Estimands

It's natural to ask

→ What is the causal effect of race on use of force

Turns out there are a lot of ways to answer this

We'll need to figure out

→ What are the possible answers
→ Which of them can be estimated from data
→ What kind of data we would need to estimate them

Even before we start to ask what to do about the answer

Estimands

Consider a simple scenario with 10 people (courtesy Macartan Humphreys [link])

→ Assumption: Race (D) is unrelated to suspicious behaviour (U)

Population

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

$M = 1$ if $3D + U \geq 4$

	$D=0$	1
$U=1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force, if M=1

$Y = 1$ if $D + U \geq 3$

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

Estimands: ATE

Population

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

	$D=0$	1
$U=1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force if m=1

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

The causal effect of D on Y

→ Proportion of people would (not) have had force applied if they had been the other race

→ 2/5

Estimands: naive estimate

Population

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

	$D=0$	1
$U=1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force if m=1

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

→ Proportion of stopped D=1 that experience force: 3/4

→ Proportion of stopped D=0 that experience force: 1

Apparent effect: −1/4

Estimands: ATE (stopping)

Population

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

	$D=0$	1
$U=1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force if m=1

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

→ 3/5

→ Proportion of the population for whom D affects stopping

Estimands: CDE

Population

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

	$D=0$	1
$U=1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force if m=1

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

The effect of D on Y if *everyone* were stopped ($M = 1$)

- → The proportion of people would (not) have had force applied if they had been the other race
- → 1/5

Estimands: ATE(M=1)

Population

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

	$D=0$	1
$U=1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force if M=1

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

- 2/5
- The effect of D on those who actually were stopped $M=1$
- e B, C, D, and E are stopped
- C and D have $Y=1$

Estimands: ATE(M=1)

Population

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

	$D=0$	1
$U=1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force if M=1

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

Note on ATE(M=1)

- → C would not have had $Y=1$ *even if* $D=0$ (from Table 3)
- → D would *would* have had $Y=1$ even if $D=0$, but then she wouldn't have been stopped at all

Estimands: CDE(M=1)

Population

	$D = 0$	1
$U = 1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

	$D = 0$	1
$U = 1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force if M=1

	$D = 0$	1
$U = 1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D = 0$	1
$U = 1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

- 1/5
- Note the same as ATE(M=1)!
- Imagine changing D but with M *fixed* to its observed value
- D is now counted in the stopped crowd, regardless that they would not have been had they been the other race

Estimands: attribution

Population

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	C
4	d	D
5	e	E

Stopping

	$D=0$	1
$U=1$	a	A
2	b	**B**
3	c	**C**
4	d	**D**
5	**e**	**E**

Force if m=1

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	**d**	**D**
5	**e**	**E**

Observed force

	$D=0$	1
$U=1$	a	A
2	b	B
3	c	**C**
4	d	**D**
5	**e**	**E**

→ Of all uses of force on minorities, how many were due to being a minority?

→ Subpopulation C, D, E

→ Only E's race was irrelevant to the use of force

→ 2/3

What should be estimated?

Which (if any) of these quantities is relevant

→ for public policy
→ for studying race
→ for studying bias
→ for causal inference

What can be estimated?

…without knowing $D \longrightarrow M$

- → Everything? (Fryer, 2018)
- → Everything sometimes? (Gaebler et al., n.d.)
- → Very little without extra assumptions (Knox et al., 2020a, 2020b)

KLM 2020 in a nutshell

Principal stratification

Divide units into principal strata

- → Would never have been stopped regardless of race
- → Would be stopped if D=1 but not if D=0 (anti-minority 'racial stops')
- → Would be stopped if D=0 but not if D=1 (anti-white 'racial stops')
- → Would be stopped regardless of race

If we knew these we could condition on them as pre-treatment covariates (Rubin, 2006)

All causal effects are weighted averages of them

Solution: principal stratification

→ If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$		
$M_i(1) = 0$		

Types of police-civilian encounters

→ If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	
$M_i(1) = 0$		

Types of police-civilian encounters

→ If $D \to M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	
$M_i(1) = 0$		"never stop" (inconspicuous)

Types of police-civilian encounters

→ If $D \to M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$		"never stop" (inconspicuous)

Types of police-civilian encounters

→ If $D \to M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

Types of police-civilian encounters

→ If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

What do we get to see in police data?

Types of police-civilian encounters

→ If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

For black civilians …

Types of police-civilian encounters

→ If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

For white civilians …

Assumptions

- Mandatory reporting
- Mediator monotonicity: No anti-white 'racial stops'
- Relative non-severity of racial stops
- Treatment ignorability

Unsurprisingly we can't get the ATE

Naive estimator is biased for ATE(M=1)

- even without unobserved U in the way
- bias is always non-positive

Bounding

→ Bias can be re-written in terms of all things that can be directly estimated from data except two:

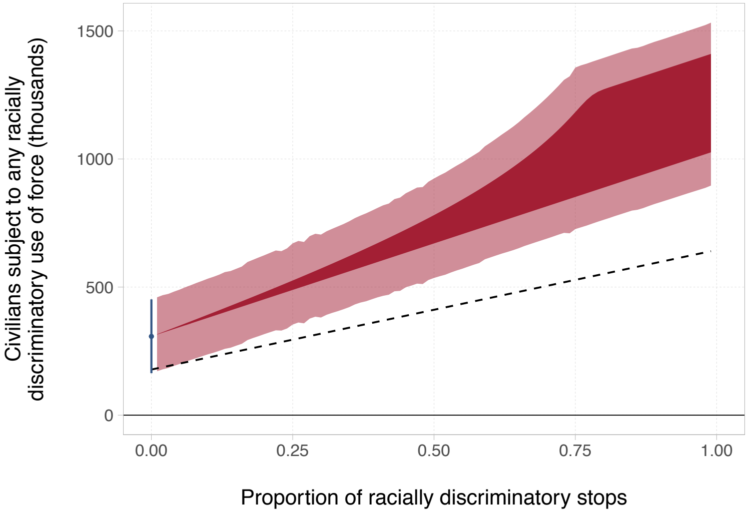
1. $\rho = \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)$: share of minority stops due to race (unknown)
2. $\theta = \mathbb{E}[Y(1,1) | D_i = 1, M_i(1) = 1, M_i(0) = 0]$ violence rate among racially stopped minorities

→ If we knew the joint distribution $\Pr(Y(1,1), M_i(0) = 0 | D_i = 1, M_i(1) = 1) = \Pr(A, B)$, we could then back out θ

→ $\theta = P(A|B) = \frac{\Pr(A,B)}{\Pr(B)} = \frac{\Pr(A,B)}{\rho}$

→ We don't, but we can place Fréchet bounds on $\Pr(A, B)$

Bounds on race effects, black vs. white



What is ρ?

What is the share of minority stops that would not have happened if civilians had been white?

- → Can be anywhere in $[0, 1)$. If $\rho = 0$, bias disappears.
- → Two prior studies estimate this using data on “Stop, Question and Frisk” in
- → Gelman, Fagan & Kiss (2007) and Goel, Rao and Schroff (2016)
- → Studies take totally different approaches
- → Results imply ρ is at least .32 or .34, respective
- → We use ρ =.32 to be conservative

Bounds for force thresholds, black vs. white

	TE_S for encounters with black civilians (vs. white)			
	No covariates		Full specification	
Minimum force	bounds	naïve	bounds	naïve
Use of hands	**(112.66, 124.59)**	**61.69**	**(86.99, 96.74)**	**23.53**
	(84.6, 151.84)	*(32.89, 90.63)*	*(81.7, 102.15)*	*(16.41, 30.61)*
Push to wall	**(24.15, 27.75)**	**4.2**	**(26.48, 30.21)**	**6.67**
	(15.5, 37.35)	*(-5.29, 14.02)*	*(24.29, 32.38)*	*(3.73, 9.52)*
Use of handcuffs	**(14.6, 16.92)**	**1.32**	**(16.56, 19.02)**	**3.9**
	(9.45, 22.61)	*(-4.83, 7.53)*	*(15.05, 20.55)*	*(1.87, 5.88)*
Draw weapon	**(4.52, 5.14)**	**1.26**	**(4.71, 5.35)**	**1.46**
	(3.13, 6.67)	*(-0.33, 2.83)*	*(4.22, 5.86)*	*(0.79, 2.13)*
Push to ground	**(4.04, 4.58)**	**1.22**	**(4.11, 4.66)**	**1.26**
	(2.79, 5.97)	*(-0.21, 2.66)*	*(3.68, 5.09)*	*(0.68, 1.82)*
Point weapon	**(1.49, 1.7)**	**0.36**	**(1.64, 1.86)**	**0.55**
	(0.96, 2.29)	*(-0.29, 1)*	*(1.37, 2.13)*	*(0.18, 0.91)*
Baton or pepper spray	**(0.17, 0.19)**	**0.08**	**(0.17, 0.19)**	**0.07**
	(0.1, 0.26)	*(-0.01, 0.15)*	*(0.12, 0.24)*	*(-0.01, 0.14)*

Bounds for force thresholds, black vs. white

	TE$_S$ for encounters with black civilians (vs. white)			
	No covariates		Full specification	
Minimum force	bounds	naïve	bounds	naïve
Use of hands	**(112.66, 124.59)**	**61.69**	**(86.99, 96.74)**	**23.53**
	(84.6, 151.84)	*(32.89, 90.63)*	*(81.7, 102.15)*	*(16.41, 30.61)*
Push to wall	**(24.15, 27.75)**	**4.2**	**(26.48, 30.21)**	**6.67**
	(15.5, 37.35)	*(-5.29, 14.02)*	*(24.29, 32.38)*	*(3.73, 9.52)*
Use of handcuffs	**(14.6, 16.92)**	**1.32**	**(16.56, 19.02)**	**3.9**
	(9.45, 22.61)	*(-4.83, 7.53)*	*(15.05, 20.55)*	*(1.87, 5.88)*
Draw weapon	**(4.52, 5.14)**	**1.26**	**(4.71, 5.35)**	**1.46**
	(3.13, 6.67)	*(-0.33, 2.83)*	*(4.22, 5.86)*	*(0.79, 2.13)*
Push to ground	**(4.04, 4.58)**	**1.22**	**(4.11, 4.66)**	**1.26**
	(2.79, 5.97)	*(-0.21, 2.66)*	*(3.68, 5.09)*	*(0.68, 1.82)*
Point weapon	**(1.49, 1.7)**	**0.36**	**(1.64, 1.86)**	**0.55**
	(0.96, 2.29)	*(-0.29, 1)*	*(1.37, 2.13)*	*(0.18, 0.91)*
Baton or pepper spray	**(0.17, 0.19)**	**0.08**	**(0.17, 0.19)**	**0.07**
	(0.1, 0.26)	*(-0.01, 0.15)*	*(0.12, 0.24)*	*(-0.01, 0.14)*

Bounds for force thresholds, black vs. white

	TE$_S$ for encounters with black civilians (vs. white)			
	No covariates		Full specification	
Minimum force	bounds	naïve	bounds	naïve
Use of hands	**(112.66, 124.59)**	**61.69**	**(86.99, 96.74)**	**23.53**
	(84.6, 151.84)	*(32.89, 90.63)*	*(81.7, 102.15)*	*(16.41, 30.61)*
Push to wall	**(24.15, 27.75)**	**4.2**	**(26.48, 30.21)**	**6.67**
	(15.5, 37.35)	*(-5.29, 14.02)*	*(24.29, 32.38)*	*(3.73, 9.52)*
Use of handcuffs	**(14.6, 16.92)**	**1.32**	**(16.56, 19.02)**	**3.9**
	(9.45, 22.61)	*(-4.83, 7.53)*	*(15.05, 20.55)*	*(1.87, 5.88)*
Draw weapon	**(4.52, 5.14)**	**1.26**	**(4.71, 5.35)**	**1.46**
	(3.13, 6.67)	*(-0.33, 2.83)*	*(4.22, 5.86)*	*(0.79, 2.13)*
Push to ground	**(4.04, 4.58)**	**1.22**	**(4.11, 4.66)**	**1.26**
	(2.79, 5.97)	*(-0.21, 2.66)*	*(3.68, 5.09)*	*(0.68, 1.82)*
Point weapon	**(1.49, 1.7)**	**0.36**	**(1.64, 1.86)**	**0.55**
	(0.96, 2.29)	*(-0.29, 1)*	*(1.37, 2.13)*	*(0.18, 0.91)*
Baton or pepper spray	**(0.17, 0.19)**	**0.08**	**(0.17, 0.19)**	**0.07**
	(0.1, 0.26)	*(-0.01, 0.15)*	*(0.12, 0.24)*	*(-0.01, 0.14)*

Summing up

This particular bit of applied causal inference opened up a lot of conceptually difficult and socially contentious issues:

- → How to think about race
- → How to think about fairness
- → How to think about effective use of force
- → The limits of inference from data
- → What data should be collected

When you get yelled at after being written up in 538, you're either doing something very wrong…

Or very right…

References

Bertrand, M. & Mullainathan, S. (2004). ‘Are {e}mily and {g}reg more employable than {l}akisha and {j}amal? {a} field experiment on labor market discrimination’. *The American Economic Review*, *94*(4), 991–1013.

Fryer, R. G. (2018). ‘An empirical analysis of racial differences in police use of force’. *Journal of Political Economy*, *127*(3), 1210–1261.

Gaebler, J., Cai, W. & Basse, G. (n.d.). *Deconstructing claims of post-treatment bias in observational studies of discrimination*.

Greiner, D. J. & Rubin, D. B. (2010). ‘Causal effects of perceived immutable characteristics’. *The Review of Economics and Statistics*, *93*(3), 775–785.

Holland, P. (2003). ‘Causation and race’. *ETS Research Report Series*, *2003*.

Knox, D., Lowe, W. & Mummolo, J. (2020a). ‘Administrative records mask racially biased policing’. *American Political Science Review*, *114*(3), 619–637.

References

Knox, D., Lowe, W. & Mummolo, J. (2020b, August 24). *Can racial bias in policing be credibly estimated using data contaminated by post-treatment selection?*

Kohler-Hausmann, I. (2019, January 1). *Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination* (SSRN Scholarly Paper No. ID 3050650). Social Science Research Network. Rochester, NY.

Paul, L. A. & Healy, K. (2018). 'Transformative treatments'. *Noûs*, *52*(2), 320–335.

Pearl, J. (2018). 'Does obesity shorten life? or is it the soda? on non-manipulable causes'. *Journal of Causal Inference*, *6*(2).

Rubin, D. B. (2006). 'Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death'. *Statistical Science*, *21*(3), 299–309.

Sen, M. & Wasow, O. (2016). 'Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics'. *Annual Review of Political Science*, *19*(1), 499–522.

References

VanderWeele, T. J. & Hernán, M. A. (2012, June 25). Causal effects and natural laws: Towards a conceptualization of causal counterfactuals for nonmanipulable exposures, with application to the effects of race and sex. In C. Berzuini, P. Dawid & L. Bernardinelli (Eds.), *Wiley series in probability and statistics* (pp. 101–113). John Wiley & Sons, Ltd.