Penn
UNIVERSITY of PENNSYLVANIA

Modern Benchmarking and the Search for Unusual Hospitals, Communities, and Cops

Greg Ridgeway

Department of Criminology

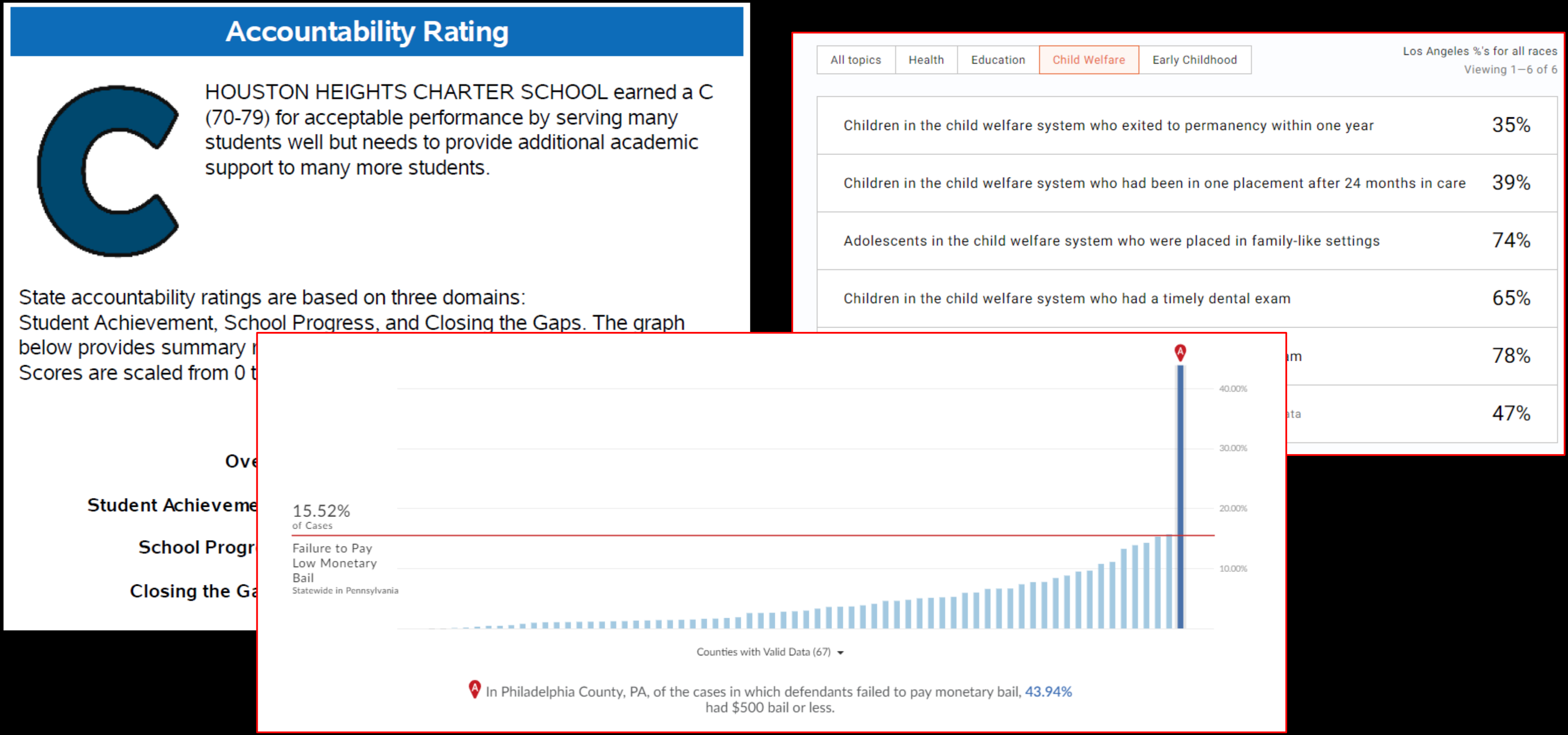
Department of Statistics

JSM Aug 2020

Scorecards are Popular

Texas Education Scorecard

California Children's Well-Being



Accountability Rating

HOUSTON HEIGHTS CHARTER SCHOOL earned a C (70-79) for acceptable performance by serving many students well but needs to provide additional academic support to many more students.

State accountability ratings are based on three domains: Student Achievement, School Progress, and Closing the Gaps. The graph below provides summary r

Scores are scaled from 0 t

Ove

Student Achieveme

School Progr

Closing the Ga

All topics | Health | Education | Child Welfare | Early Childhood

Los Angeles %'s for all races
Viewing 1–6 of 6

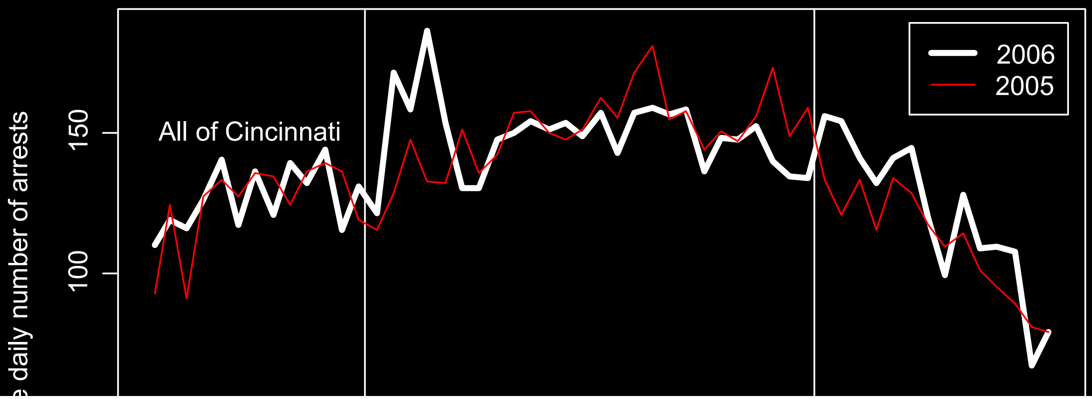
Children in the child welfare system who exited to permanency within one year	35%
Children in the child welfare system who had been in one placement after 24 months in care	39%
Adolescents in the child welfare system who were placed in family-like settings	74%
Children in the child welfare system who had a timely dental exam	65%
m	78%
ta	47%

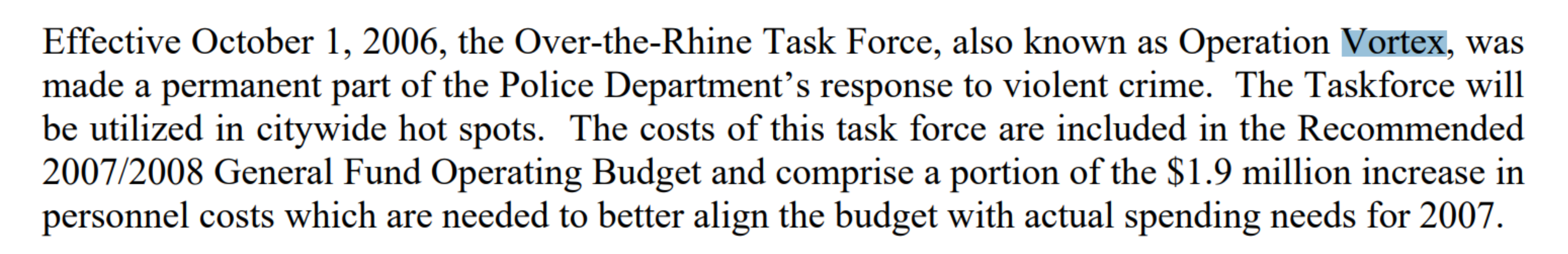
In Philadelphia County, PA, of the cases in which defendants failed to pay monetary bail, 43.94% had $500 bail or less.

Measures for Justice

Have these scorecards addressed fundamental differences?

Our Story Begins in 2006 with Operation Vortex




Effective October 1, 2006, the Over-the-Rhine Task Force, also known as Operation Vortex, was made a permanent part of the Police Department's response to violent crime. The Taskforce will be utilized in citywide hot spots. The costs of this task force are included in the Recommended 2007/2008 General Fund Operating Budget and comprise a portion of the $1.9 million increase in personnel costs which are needed to better align the budget with actual spending needs for 2007.

"highly visible proactive unit that has a zero-tolerance approach to street crimes, drug trafficking, and quality of life issues"

JSM Aug 2020

Does Operation Vortex Exacerbate Racial Disparities?

- Propensity score weight regular patrol stops to resemble stops involving Vortex officers
  - Time: hour, day of week, month of year
  - Place: block
  - Reason: moving violations, stolen auto, criminal suspect
- Compare Vortex and standard patrol stops on race of stopped drivers, searches, and hit rates

Operation Vortex Disproportionately Affects Black Drivers

- Propensity score weight regular patrol stops to resemble stops involving Vortex officers
  - Time: hour, day of week, month of year
  - Place: block
  - Reason: moving violations, stolen auto, criminal suspect
- Compare Vortex and standard patrol stops on race of stopped drivers, searches, and hit rates

Unit	% Black				
Vortex	71%				
Patrol	65%				

Black and white drivers equally likely to be searched

- Propensity score weight regular patrol stops to resemble stops involving Vortex officers
  - Time: hour, day of week, month of year
  - Place: block
  - Reason: moving violations, stolen auto, criminal suspect
- Compare Vortex and standard patrol stops on race of stopped drivers, searches, and hit rates

Unit	% Black	Search rate			
		Black	White		
Vortex	71%	22%	25%		
Patrol	65%	13%	14%		

Vortex less likely to recover contraband from searched black drivers

- Propensity score weight regular patrol stops to resemble stops involving Vortex officers
  - Time: hour, day of week, month of year
  - Place: block
  - Reason: moving violations, stolen auto, criminal suspect
- Compare Vortex and standard patrol stops on race of stopped drivers, searches, and hit rates

Unit	% Black	Search rate		Hit rate	
		Black	White	Black	White
Vortex	71%	22%	25%	23%	33%
Patrol	65%	13%	14%		

Vortex less likely to recover contraband from searched black drivers

- Propensity score weight regular patrol stops to resemble stops involving Vortex officers
  - Time: hour, day of week, month of year
  - Place: block
  - Reason: moving violations, stolen auto, criminal suspect
- Compare Vortex and standard patrol stops on race of stopped drivers, searches, and hit rates

Unit	% Black	Search rate		Hit rate	
		Black	White	Black	White
Vortex	71%	22%	25%	23%	33%
Patrol	65%	13%	14%		

Vortex has a racial disparity in hit rates not observed in standard patrol

- Propensity score weight regular patrol stops to resemble stops involving Vortex officers
  - Time: hour, day of week, month of year
  - Place: block
  - Reason: moving violations, stolen auto, criminal suspect
- Compare Vortex and standard patrol stops on race of stopped drivers, searches, and hit rates

Unit	% Black	Search rate		Hit rate	
		Black	White	Black	White
Vortex	71%	22%	25%	23%	33%
Patrol	65%	13%	14%	23%	23%

Three benchmarking applications

- Which officers stop black pedestrians at an unusual rate?
- Which communities are particularly dissatisfied with the police?
- Which hospitals have…
  - excessive opioid prescriptions?
  - unusually high mortality rates?

Three benchmarking applications

- Which officers stop black pedestrians at an unusual rate?
- Which communities are particularly dissatisfied with the police?
- Which hospitals have…
  - excessive opioid prescriptions?
  - unusually high mortality rates?

G. Ridgeway and J.M. MacDonald (2009). "Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops," *Journal of the American Statistical Association* 104:661–668

Is an Officer Who Stops 86% Black Pedestrians Unusual?

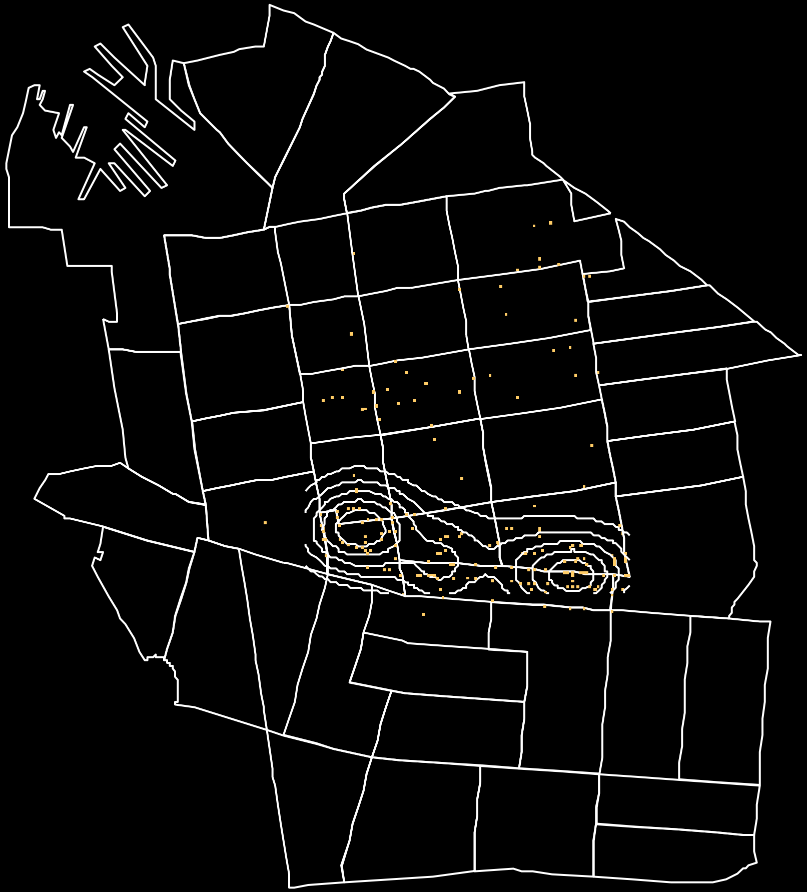
Stop Characteristic	Example Officer (%) n = 392	
% black pedestrians stopped	86%	

- Combine three statistical techniques to answer this question
  - Propensity score weighting
  - Doubly robust estimation
  - False discovery rate

We Know a Lot About the Environment of this Officer’s Stops

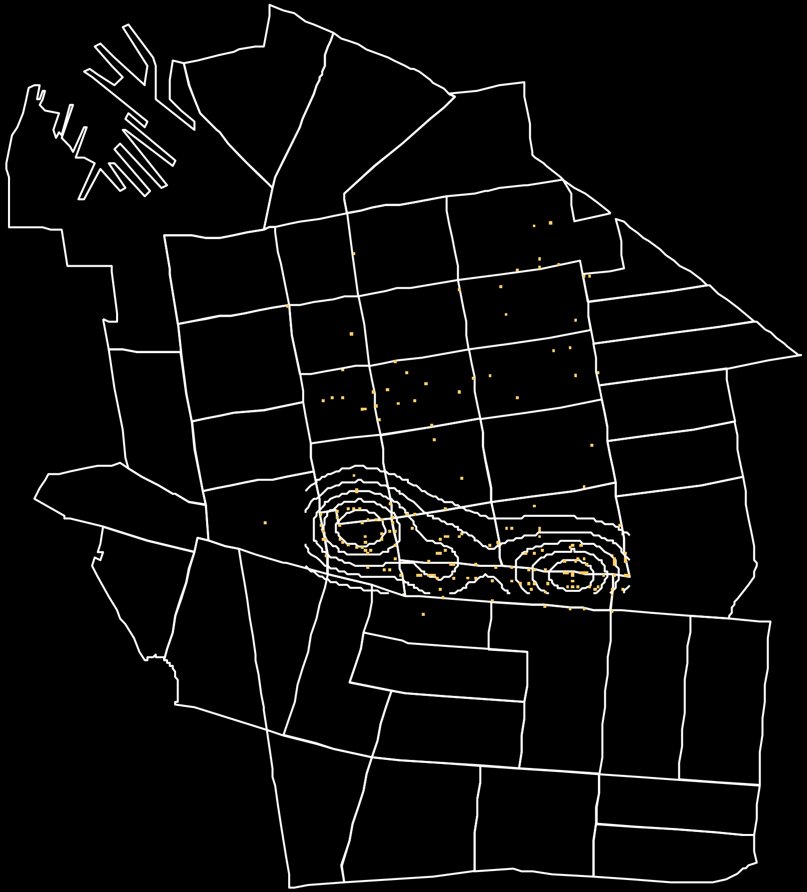
Stop Characteristic		Example Officer (%) n = 392	
% black pedestrians stopped		86%	
Month	January	3	
	February	4	
	March	8	
Day of the week	Monday	13	
	Tuesday	11	
	Wednesday	14	
Time of day	(4-6 p.m.]	9	
	(6-8 p.m.]	8	
	(8-10 p.m.]	23	
	(10 p.m. -12 a.m.]	17	
Patrol borough	Brooklyn North	100	
Precinct	B	98	
	C	1	
Outside		96	
In uniform	Yes	99	
Radio run	Yes	1	

We Also Know the Exact Location of This Officer's Stops



Example Officer

Idea: Reweight Stops Made By Other Officers to Resemble This Officer's Stops



Example Officer

- Align their distributions

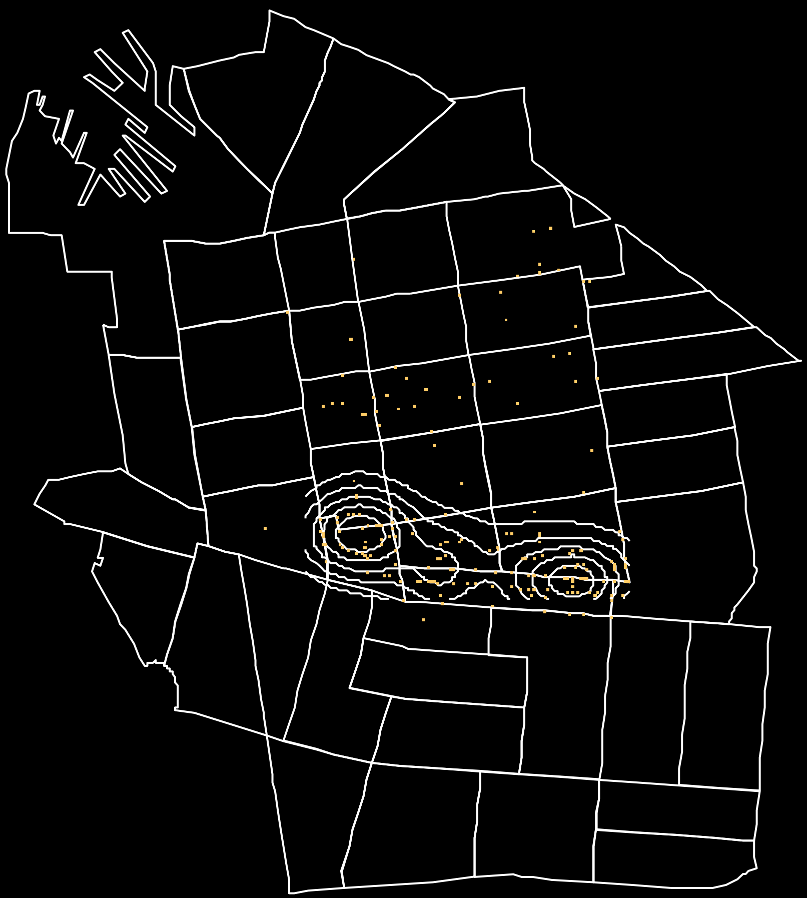
$$f(\mathbf{x}|\underbrace{t=1}_{\text{Example officer}}) = w(\mathbf{x})f(\mathbf{x}|\underbrace{t=0}_{\text{Other officers}})$$

- Solving for $w(\mathbf{x})$ yields the propensity score weight

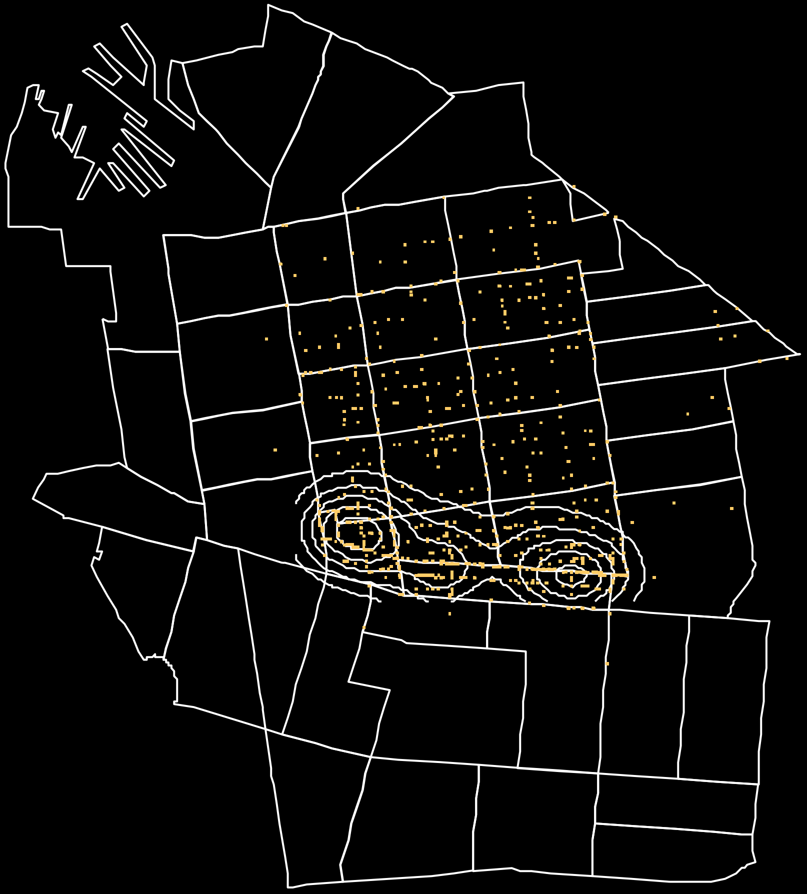
$$w(\mathbf{x}) \propto \frac{P(t=1|\mathbf{x})}{1-P(t=1|\mathbf{x})}$$

- Estimate $P(t=1|\mathbf{x})$ using boosted logistic regression as implemented in `gbm`

Reweighting Aligns the Distribution of Stop Locations



Example Officer



Matched Stops

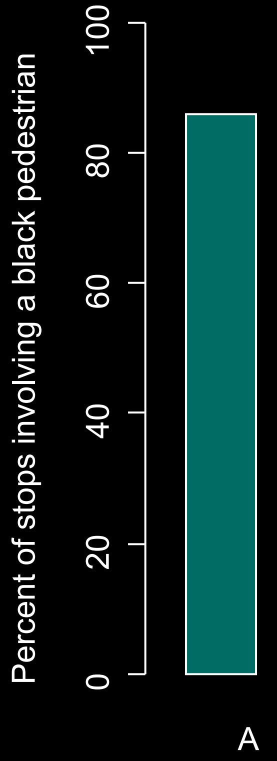
Reweighting Also Aligns the Distribution of All Other Stop Features

Stop Characteristic		Example Officer (%)	Internal Benchmark (%)
		n = 392	ESS = 3,676
% black pedestrians stopped		86%	
Month	January	3	3
	February	4	4
	March	8	9
Day of the week	Monday	13	13
	Tuesday	11	10
	Wednesday	14	15
Time of day	(4-6 p.m.]	9	10
	(6-8 p.m.]	8	8
	(8-10 p.m.]	23	23
	(10 p.m. -12 a.m.]	17	17
Patrol borough	Brooklyn North	100	100
Precinct	B	98	98
	C	1	1
Outside		96	94
In uniform	Yes	99	97
Radio run	Yes	1	3

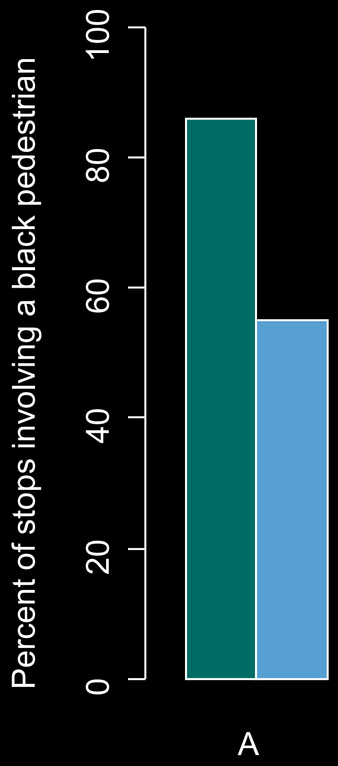
Colleagues at the Same Time, Place, and Context Stop 55% Black Pedestrians

Stop Characteristic		Example Officer (%)	Internal Benchmark (%)
		n = 392	ESS = 3,676
% black pedestrians stopped		86%	55%
Month	January	3	3
	February	4	4
	March	8	9
Day of the week	Monday	13	13
	Tuesday	11	10
	Wednesday	14	15
Time of day	(4-6 p.m.]	9	10
	(6-8 p.m.]	8	8
	(8-10 p.m.]	23	23
	(10 p.m. -12 a.m.]	17	17
Patrol borough	Brooklyn North	100	100
Precinct	B	98	98
	C	1	1
Outside		96	94
In uniform	Yes	99	97
Radio run	Yes	1	3

86% of the Officer's Stops Were Black…



…Compared with 55% for the Benchmark



- Doubly robust benchmark estimate obtainable from weighted logistic regression

$$\ell(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} w_i \left(y_i s(t_i, \mathbf{x}_i|\boldsymbol{\beta}) - \log\left(1 + e^{s(t_i,\mathbf{x}_i|\boldsymbol{\beta})}\right)\right)$$

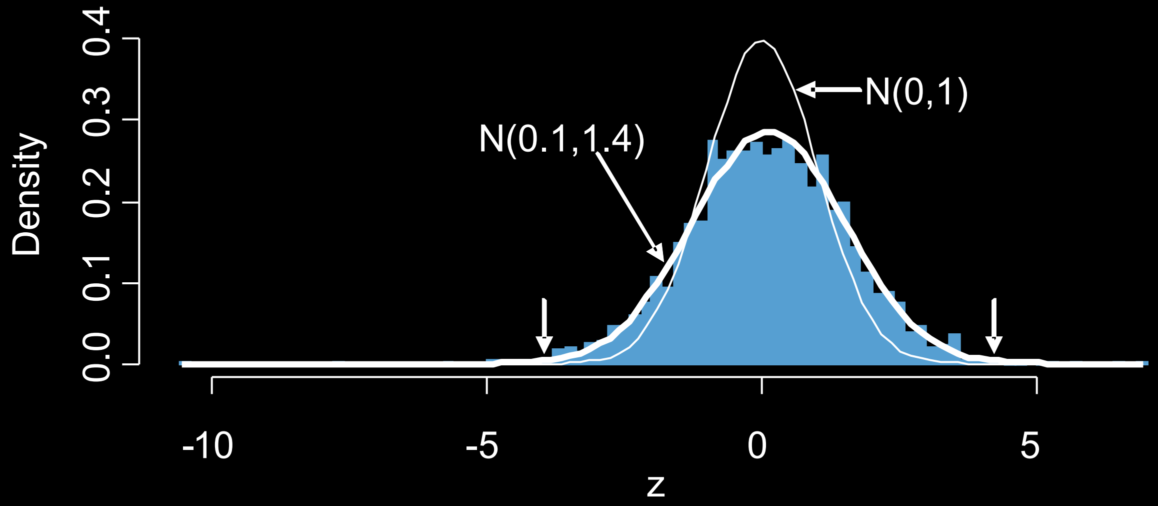
- Disparity computed as

$$\hat{\theta}_A^{DR} = \frac{\mathbf{1}}{\sum t_i} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_i \left(\frac{1}{1+\exp\left(-s(1,\mathbf{x}_i|\hat{\boldsymbol{\beta}})\right)} - \frac{1}{1+\exp\left(-s(0,\mathbf{x}_i|\hat{\boldsymbol{\beta}})\right)}\right)$$

Predicted probability stopped pedestrian is black for example officer

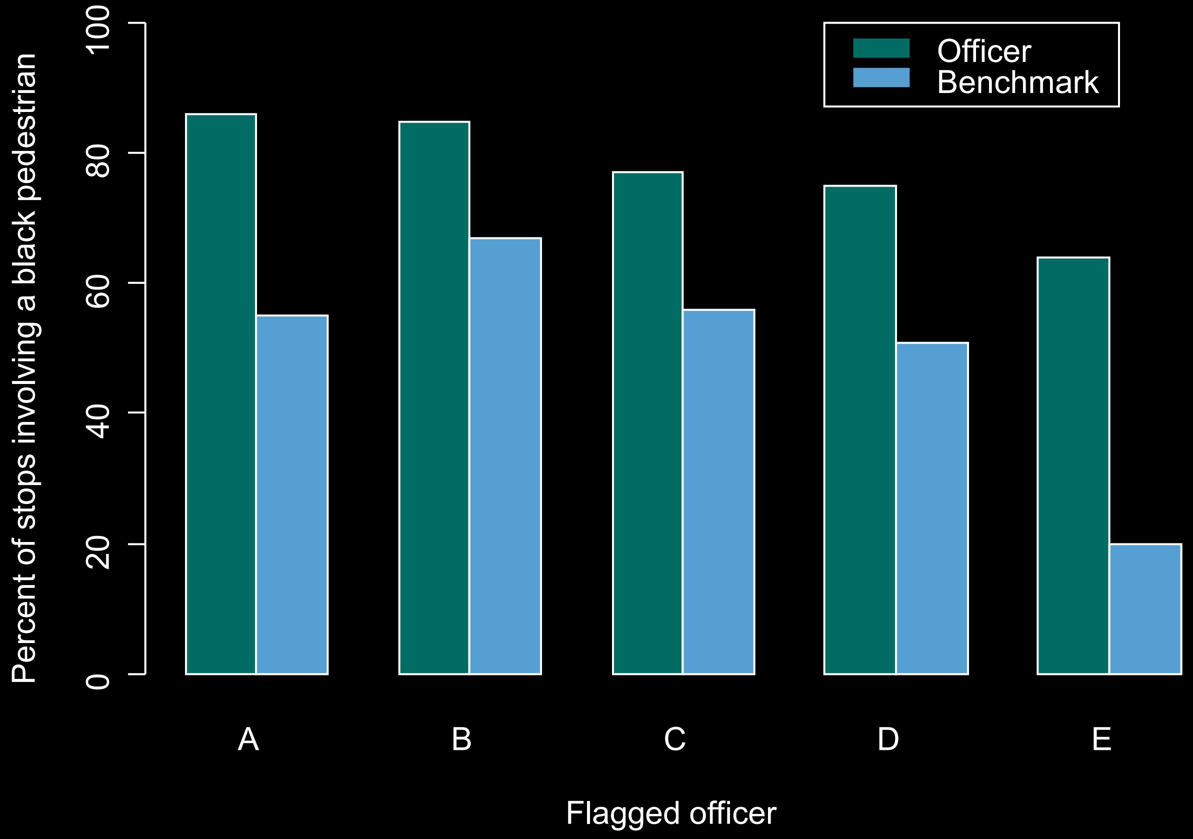
Predicted probability stopped pedestrian is black for other officers

Repeat for Nearly 3,000 NYPD Officers Actively Involved in Stops



- $$P(\text{problem}|z) = 1 - \frac{f(z|\text{no problem})f(\text{no problem})}{f(z)}$$
$$\geq 1 - \frac{f_0(z)}{f(z)}$$
- Right tail consists of 5 officers with "problem officer" probabilities in excess of 50%
- Standard cutoff of z > 2.0 flags 242 officers, 90% of which have fdr estimated to be greater than 0.999

Analysis in NYPD Flagged Five Officers



Benchmarking

1. Apply propensity score weights so benchmark is based on activities in similar context
2. Compute z-statistic from a propensity score weighted regression
3. Repeat for all units, customizing benchmark for each
4. Compute false discovery rate based on empirical distribution of z-statistics

Three benchmarking applications

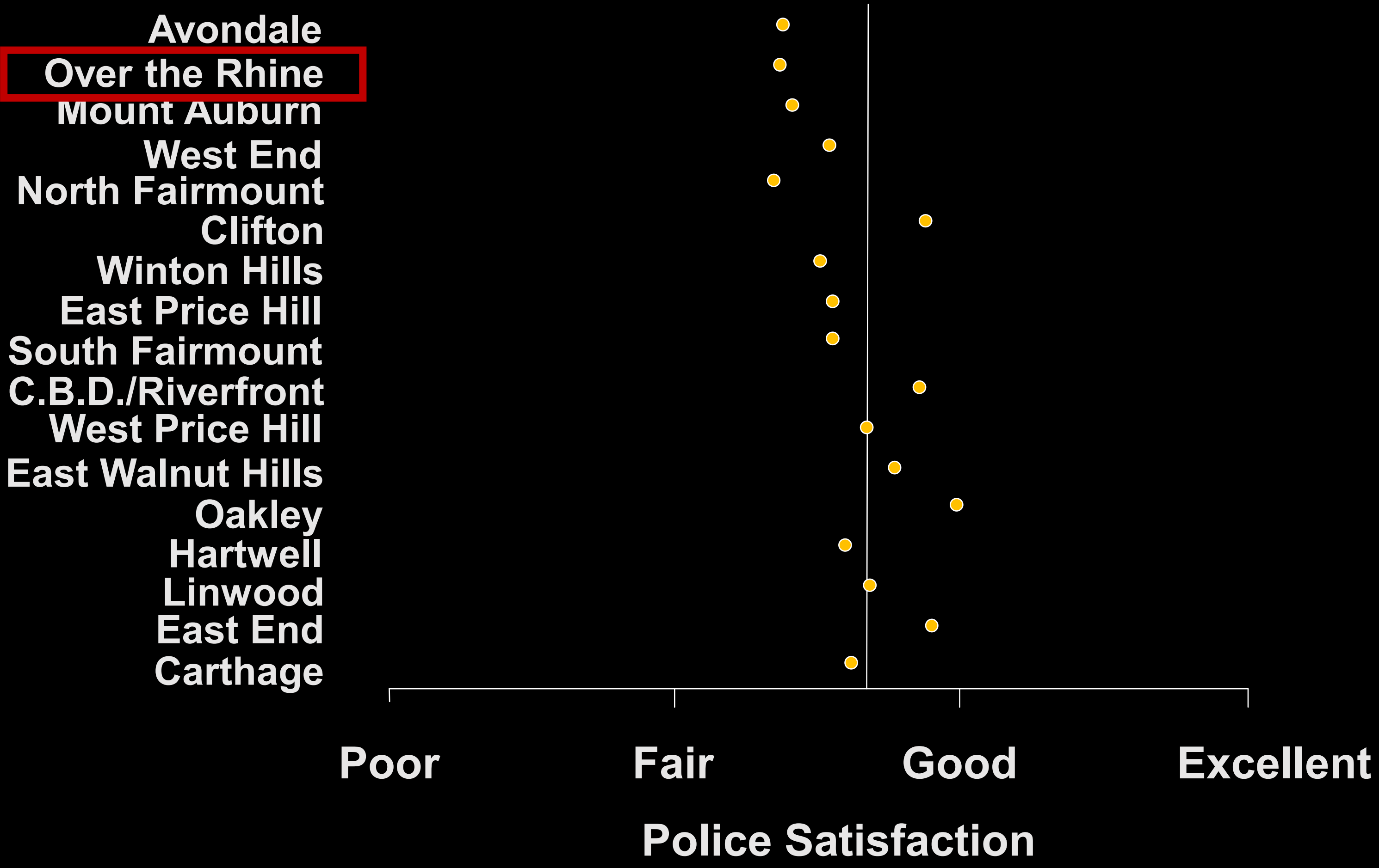
- Which officers stop black pedestrians at an unusual rate?
- Which communities are particularly dissatisfied with the police?
- Which hospitals have…
    - excessive opioid prescriptions?
    - unusually high mortality rates?

G. Ridgeway and J.M. MacDonald (2014). “A Method for Internal Benchmarking of Criminal Justice System Performance,” *Crime & Delinquency* 60(1):145-162

In Which Neighborhoods Are Police Underperforming?

- Cincinnati Police Department sponsored a citywide survey of citizens
  - Citizen satisfaction with the police
  - Perceptions of racially discriminatory police practices
  - Whether residents felt that they had personally experienced racial profiling
- 6,000 residents in Cincinnati selected via random-digit dialing and list-assisted sampling methods
  - Stratified to cover 45 defined Cincinnati neighborhoods
  - Respondents were 18 years or older

Few Neighborhoods Differ from Benchmarks



JSM Aug 2020

Respondents Differ on Key Features Associated with Police Satisfaction

Respondent characteristics	Respondents from Over-the-Rhine (N=146)	Respondents from other neighborhoods (N=5,671)	
Less than HS	21	10	

Respondents Differ on Key Features Associated with Police Satisfaction

Respondent characteristics	Respondents from Over-the-Rhine (N=146)	Respondents from other neighborhoods (N=5,671)
Less than HS	21	10
College degree+	23	33
Black	66	42
White	30	53
$20,000 or less	47	25
$100,000 or more	6	11
Employed (%)	60	58
Married (%)	15	38
Male (%)	43	36
Age 22-29	16	8
Age 65+	13	25
Homeowner (%)	20	60
Children at home (%)	40	31

Constructed Benchmark Matches Neighborhoods on These Features

Respondent characteristics	Respondents from Over-the-Rhine (N=146)	Respondents from other neighborhoods (N=5,671)	Weighted respondents from other neighborhoods (N=422)
Less than HS	21	10	**21**
College degree+	23	33	**22**
Black	66	42	**65**
White	30	53	**32**
$20,000 or less	47	25	**45**
$100,000 or more	6	11	**5**
Employed (%)	60	58	**58**
Married (%)	15	38	**16**
Male (%)	43	36	**42**
Age 22-29	16	8	**17**
Age 65+	13	25	**13**
Homeowner (%)	20	60	**21**
Children at home (%)	40	31	**38**

Constructed Benchmark Matches Neighborhoods on These Features

Respondent characteristics	Respondents from Over-the-Rhine (N=146)	Respondents from other neighborhoods (N=5,671)	Weighted respondents from other neighborhoods (N=422)
Less than HS	21	10	**21**
College degree+	23	33	**22**
Black	66	42	**65**
White	30	53	**32**
$20,000 or less	47	25	**45**
$100,000 or more	6	11	**5**
Employed (%)	60	58	**58**
Married (%)	15	38	**16**
Male (%)	43	36	**42**
Age 22-29	16	8	**17**
Age 65+	13	25	**13**
Homeowner (%)	20	60	**21**
Children at home (%)	40	31	**38**

Police Satisfaction in Over-the-Rhine Is Close to Its Benchmark

Respondent characteristics	Respondents from Over-the-Rhine (N=146)	
Satisfaction with the Police	2.37	

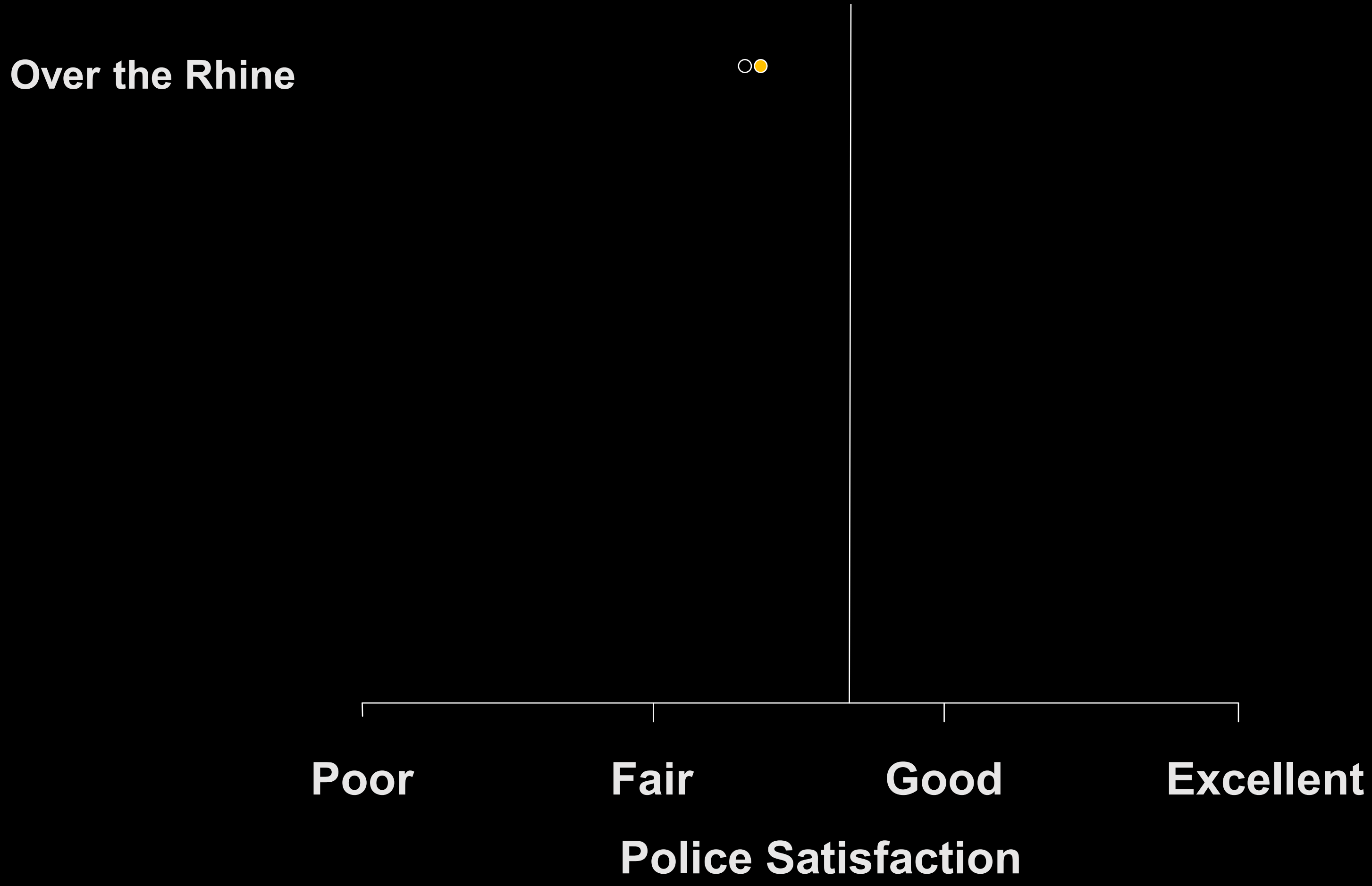
Police Satisfaction in Over-the-Rhine Is Close to Its Benchmark

Respondent characteristics	Respondents from Over-the-Rhine (N=146)	Respondents from other neighborhoods (N=422)
Satisfaction with the Police	2.37	2.31

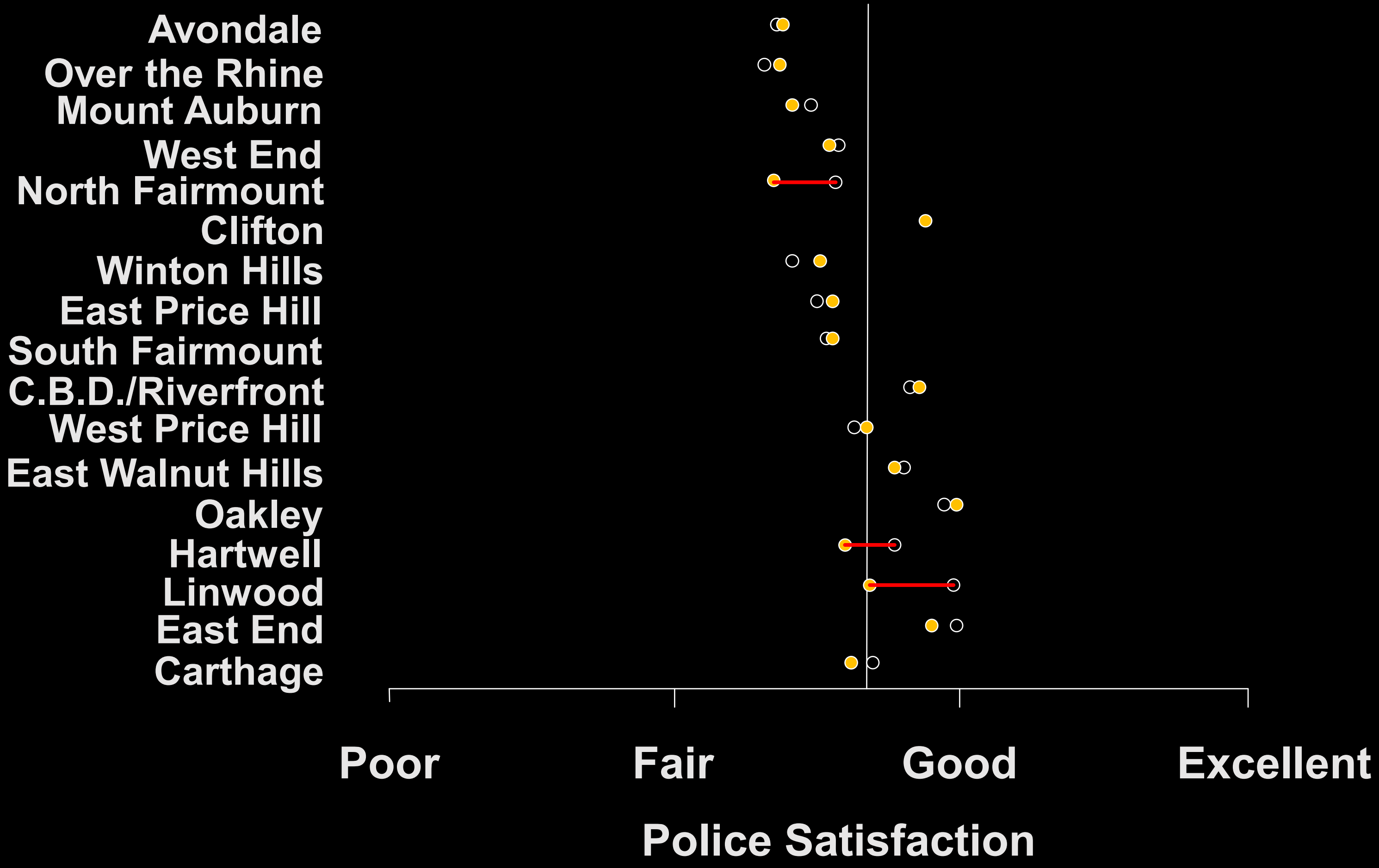
Police Satisfaction in Over-the-Rhine Is Close to Its Benchmark

Respondent characteristics	Respondents from Over-the-Rhine (N=146)	Respondents from other neighborhoods (N=422)
Satisfaction with the Police	2.37	2.31
Perception of Racial Profiling	2.59	2.65
Personal Racial Profiling Experience	32%	30%

Police Satisfaction in Over the Rhine is Close to Expectation



Few Neighborhoods Differ from Benchmarks



JSM Aug 2020

Three benchmarking applications

- Which officers stop black pedestrians at an unusual rate?
- Which communities are particularly dissatisfied with the police?
- Which hospitals have…
  - unusually high mortality and readmission rates?
  - excessive opioid prescriptions?

G. Ridgeway, M. Nørgaard, T.B. Rasmussen, W.D. Finkle, L. Pedersen, H.E. Bøtker, and H.T. Sørensen (2019). “Benchmarking Danish Hospitals on Mortality and Readmission Rates After Cardiovascular Admission,” *Clinical Epidemiology* 11:67-80

Compare Performance of 26 Danish Hospitals

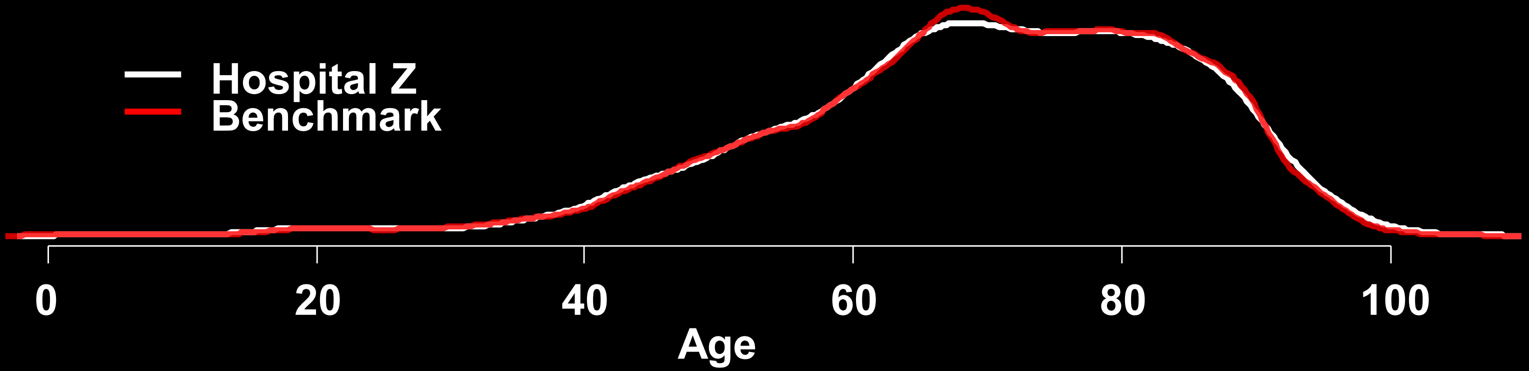
- Data from all Danish hospitals
    - 331,513 patients
    - Danish National Patient Registry and the Danish National Health Service Prescription Database
    - discharged with a primary cardiovascular diagnosis
    - from one of 26 Danish hospitals during 2011-2015
- Main outcome measures
    - 30-day post-admission mortality rates
    - 30-day post-discharge readmission rates
- Patient features
    - age, sex
    - primary discharge diagnosis
    - diagnosis history
    - medications
    - previous cardiac procedures
    - comorbidities

Benchmark Patients at Other Hospitals Resemble Hospital Z's Patients

	Hospital Z	Benchmark patients	All other hospitals
Age, average	69.9	69.9	**68.6**
Male, %	55.7	55.2	**57.4**

Distributions Match, Not Only Means

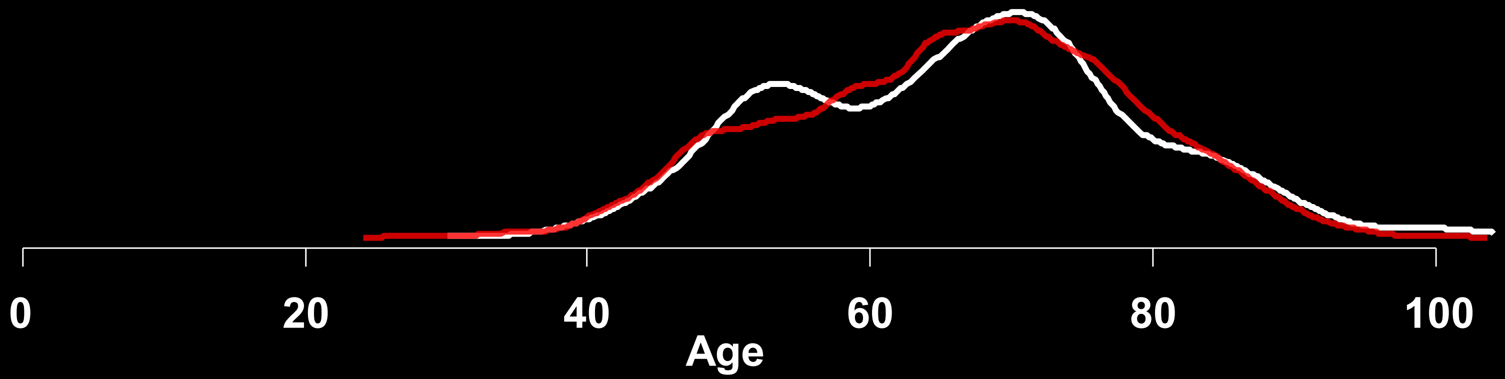
	Hospital Z	Benchmark patients	All other hospitals
Age, average	69.9	69.9	**68.6**
Male, %	55.7	55.2	**57.4**



And Multivariate Marginals Match

	Hospital Z	Benchmark patients	All other hospitals
Age, average	69.9	69.9	**68.6**
Male, %	55.7	55.2	**57.4**

I20.0 (Unstable angina) patients on statins



Patients Match on 105 Discharge Diagnoses

	Hospital Z	Benchmark patients	All other hospitals
Myocardial infarction (any)	8.8	8.9	**10.5**

Patients Match on 105 Discharge Diagnoses

	Hospital Z	Benchmark patients	All other hospitals
Myocardial infarction (any)	8.8	8.9	**10.5**
STEMI	0.5	0.5	**3.1**
Unstable angina	4.2	4.2	**2.4**
Stable coronary artery disease	15.7	15.7	**11.4**
Arterial hypertension	8.2	8.2	**5.4**
Atrial fibrillation or flutter	27.7	27.9	**23.8**
Ischemic stroke	4.7	4.7	**11.4**
…			

Patients Match on 5-year Cardiovascular Diagnosis History

	Hospital Z	Benchmark patients	All other hospitals
Myocardial infarction	9.3	9.1	**9.4**
Heart Failure	11.2	11.6	**16.0**
Arterial hypertension	27.8	28.3	**33.0**
Valvular heart disease	5.0	5.2	**8.1**
Stroke (any)	7.0	7.1	**8.3**
…			

Patients Match on Current Cardiovascular Medication

	Hospital Z	Benchmark patients	All other hospitals
Current use of prescribed cardiovascular medications			
Betablockers	44.3	44.1	**40.9**
Diuretics	44.2	43.6	**37.5**

Patients Match on Procedures

	Hospital Z	Benchmark patients	All other hospitals
Current use of prescribed cardiovascular medications			
Betablockers	44.3	44.1	**40.9**
Diuretics	44.2	43.6	**37.5**
Previous cardiac procedures			
Implantable cardiac defibrillator	1.4	1.4	**2.0**
Aortic valve surgery	1.1	1.0	**1.6**

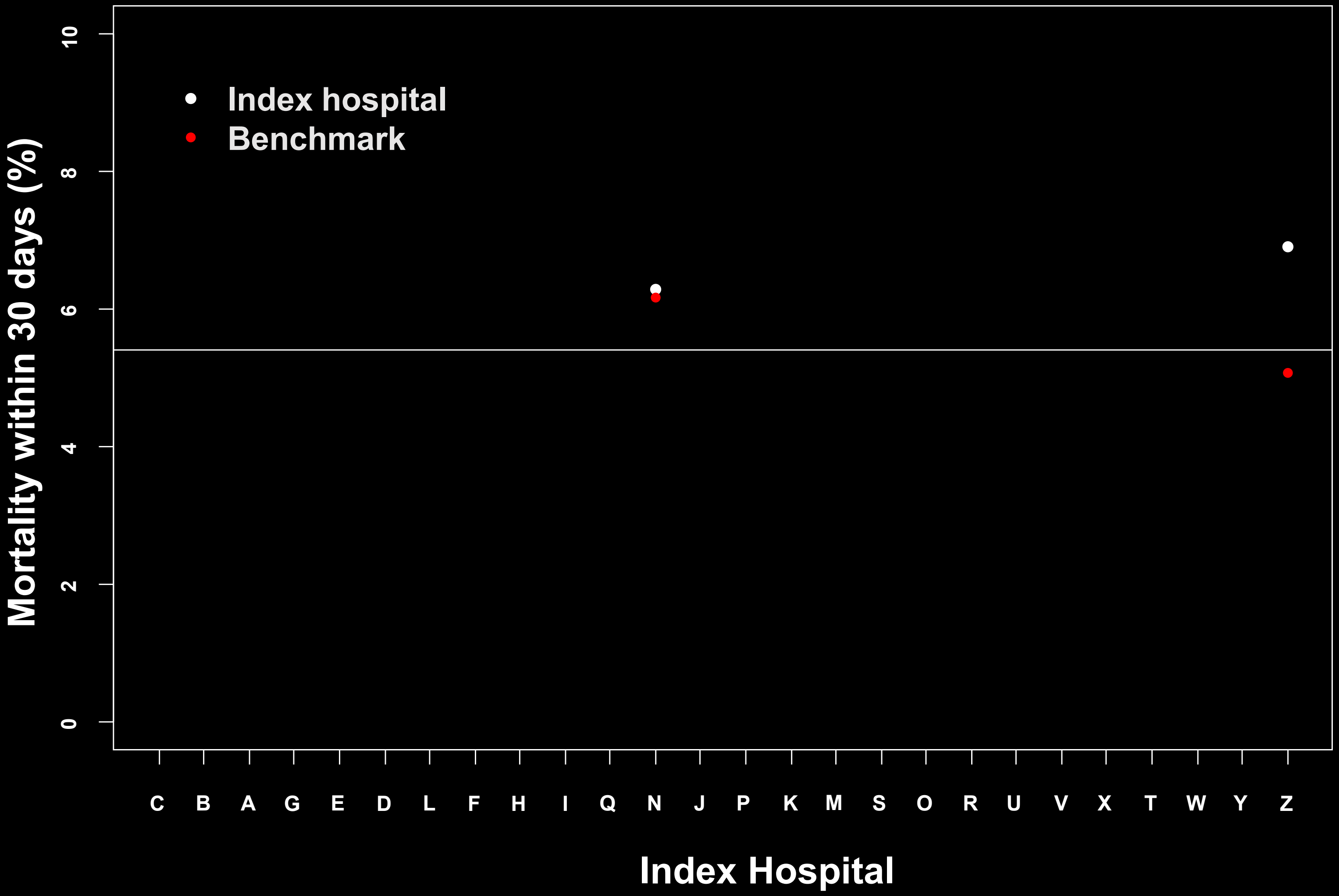
Patients Match on Comorbidities

	Hospital Z	Benchmark patients	All other hospitals
Current use of prescribed cardiovascular medications			
Betablockers	44.3	44.1	**40.9**
Diuretics	44.2	43.6	**37.5**
Previous cardiac procedures			
Implantable cardiac defibrillator	1.4	1.4	**2.0**
Aortic valve surgery	1.1	1.0	**1.6**
Selected comorbidity diagnosis history			
Diabetes	11.1	11.4	**12.9**
Liver disease	0.8	0.8	**1.5**

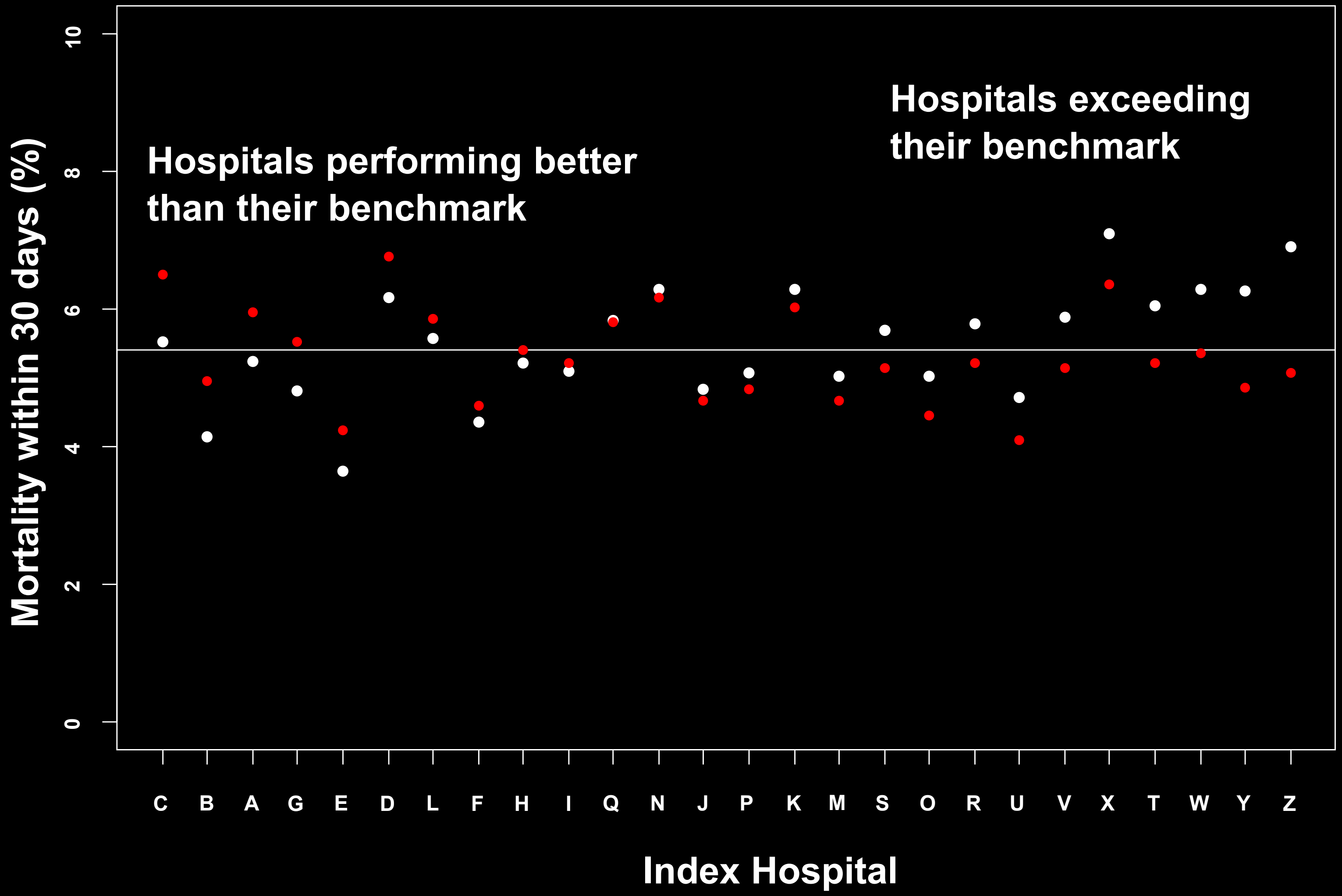
Patients Match on Other Prescribed Medication

	Hospital Z	Benchmark patients	All other hospitals
Current use of prescribed cardiovascular medications			
Betablockers	44.3	44.1	**40.9**
Diuretics	44.2	43.6	**37.5**
Previous cardiac procedures			
Implantable cardiac defibrillator	1.4	1.4	**2.0**
Aortic valve surgery	1.1	1.0	**1.6**
Selected comorbidity diagnosis history			
Diabetes	11.1	11.4	**12.9**
Liver disease	0.8	0.8	**1.5**
Current use of selected prescribed other medications			
Antidepressants	9.8	9.4	**7.9**
Antidiabetics	13.5	13.8	**14.1**

Compare Every Hospital to Its Customized Benchmark

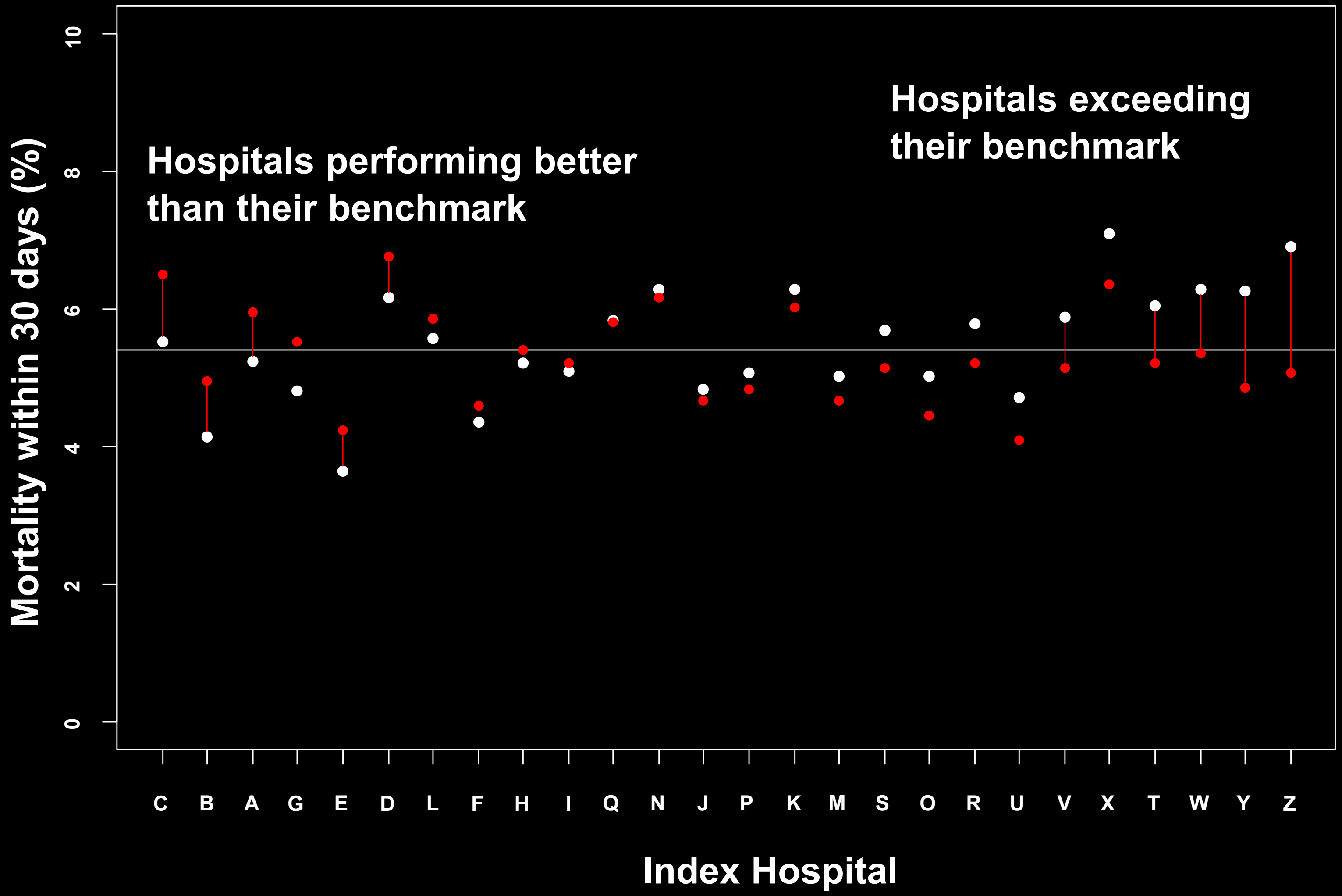


Compare Every Hospital to Its Customized Benchmark



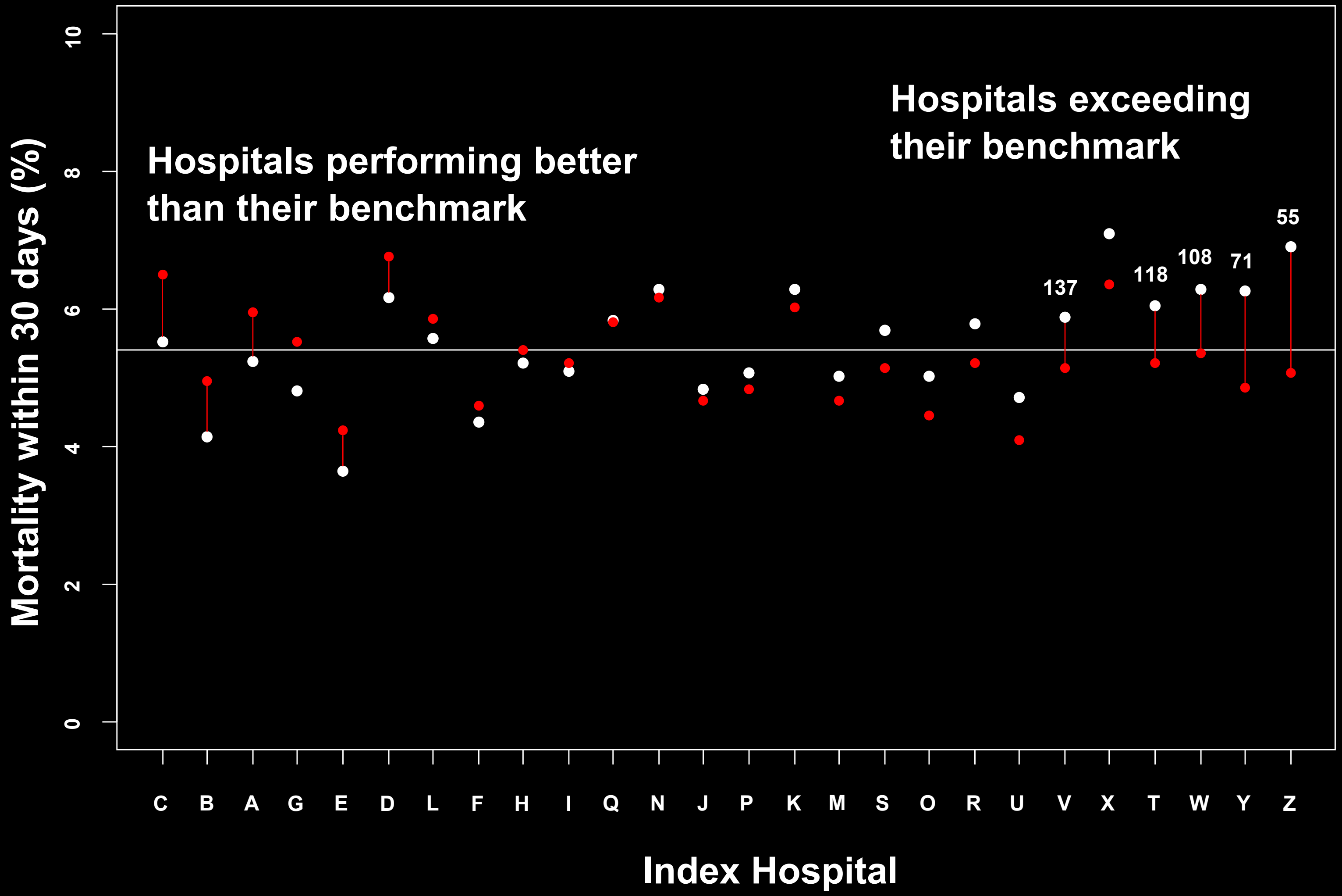
JSM Aug 2020

False Discovery Rate Below 5% for Five Hospitals Exceeding Benchmark

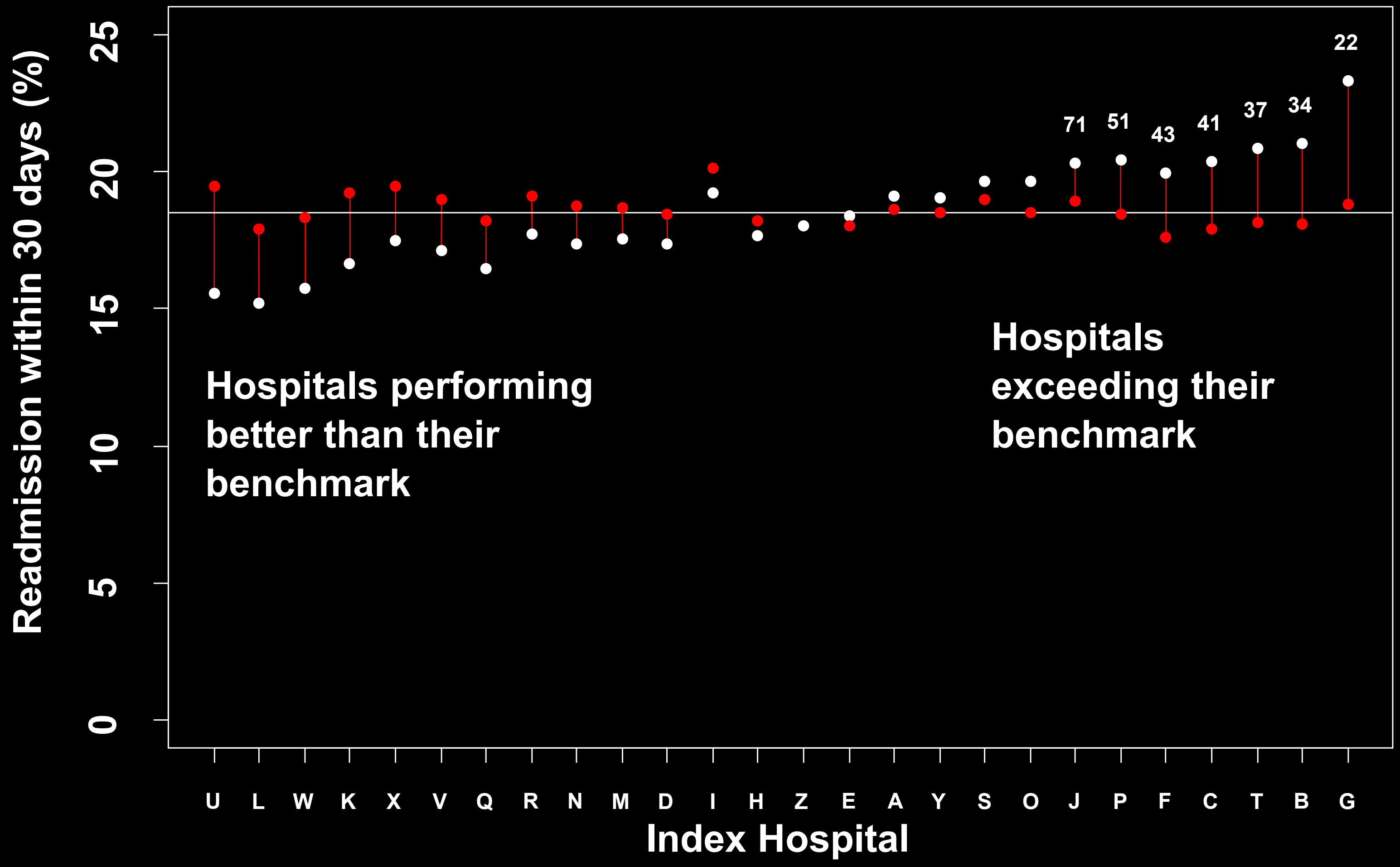


JSM Aug 2020

Number Needed to Harm is Low at Hospital Z



Hospital T Also Has High 30-day Readmission Rates



JSM Aug 2020

Broad Applicability in Creating Hospital Scorecards

	Hospital X	Benchmark	All Patients
30-day readmission	15.8%	11.7%	7.1%

Consider 287 hospitals

- MarketScan Medicaid Multi-State Database
- Admissions between January 2012-September 2014

Broad Applicability in Creating Hospital Scorecards

	Hospital X	Benchmark	All Patients
30-day readmission	15.8%	11.7%	7.1%
Oxygen expense (90-day)	$12.63	$5.30	$2.97
Oxygen prescribed (per 100)	9.7	9.7	6.2

Broad Applicability in Creating Hospital Scorecards

	Hospital X	Benchmark	All Patients
30-day readmission	15.8%	11.7%	7.1%
Oxygen expense (90-day)	$12.63	$5.30	$2.97
Oxygen prescribed (per 100)	9.7	9.7	6.2
Oxycodone supply (30-day)	5.7	5.0	2.5
Oxycodone supply (90-day)	12.3	11.4	5.2
Opiate supply (30-day)	10.1	12.1	6.5
Opiate supply (90-day)	23.5	29.0	14.1
Any opiate prescribed	49.2%	57.0%	42.2%

Traditional Regression Approach Flags Hospital X on Several Outcomes

	Hospital X	Benchmark	All Patients
30-day readmission	15.8%	11.7%	7.1%
Oxygen expense (90-day)	$12.63	$5.30	$2.97
Oxygen prescribed (per 100)	9.7	9.7	6.2
Oxycodone supply (30-day)	5.7	5.0	2.5
Oxycodone supply (90-day)	12.3	11.4	5.2
Opiate supply (30-day)	10.1	12.1	6.5
Opiate supply (90-day)	23.5	29.0	14.1
Any opiate prescribed	49.2%	57.0%	42.2%

But the False Discovery Rate is Low Only for Oxygen Expense

	Hospital X	Benchmark	FDR
30-day readmission	15.8%	11.7%	1.00
Oxygen expense (90-day)	$12.63	$5.30	0.06
Oxygen prescribed (per 100)	9.7	9.7	1.00
Oxycodone supply (30-day)	5.7	5.0	1.00
Oxycodone supply (90-day)	12.3	11.4	1.00
Opiate supply (30-day)	10.1	12.1	0.39
Opiate supply (90-day)	23.5	29.0	0.49
Any opiate prescribed	49.2%	57.0%	0.27

Identify Hospitals with Unusual Opioid Prescription Patterns

ID	Hospital	Benchmark	Hospital # Patients	Benchmark # Patients	False Discovery Rate
	Rate of prescription per 100 discharges				
XP	62.1	51.8	642	28,104	0.01
XD	36.6	31.8	4,270	28,744	0.01
XH	63.6	36.1	228	3,827	0.01
ZA	61.4	46.7	526	5,366	0.01
	Days supply 30 days post-discharge				
XD	3.1	2.5	4,270	28,744	0.08
XP	13.4	9.4	642	28,104	0.14
XH	8.5	4.6	228	3,827	0.14

Broad Applicability of General Approach

- Justice
  - Racial profiling
  - Police performance
  - Sentencing disparities
  - Judicial decision making
- Healthcare
  - Mortality
  - Expenses
  - Prescription practice
- Education?
- Transportation?

Modern Benchmarking and the Search for Unusual Hospitals, Communities, and Cops

Greg Ridgeway

Department of Criminology

Department of Statistics

JSM Aug 2020