Empirical Approaches to Physician Decision Making in Economics

Janet Currie

CHW
CENTER FOR HEALTH AND WELLBEING
PRINCETON UNIVERSITY

To Err is Human… Much Evidence of Poor Medical Decision Making

- "*To Err is Human*," (National Institute of Medicine, 2000): ~98,000 U.S. hospital patients die from preventable medical errors annually.
- The Harvard Medical Practice Study (NEJM 1991): 1% of hospital admissions involve an adverse event due to an error; but this only includes mistakes in which an adverse outcome occurred shortly after the error.
- McGlynn et al. (NEJM, 2003): U.S. adults receive only 55% of recommended care. Fisher (NEJM, 2003): 30% of U.S. care may be unnecessary. Many tests and procedures are overused.
- Unnecessary and harmful drug prescriptions, e.g. opioids and antibiotics.

More Broadly, Many Examples of Similar Patients Receiving Different Care

- Why does this happen?
  - Patient demand/unobserved patient differences
  - Fear of lawsuits (defensive medicine)
  - Profit motive (more broadly, incentives)
  - Peer effects
  - Discrimination/communications/homophily
  - Training
  - Cognitive limitations/skills

- What can be done about it?
  - Guidelines and algorithms

Patient Demand: Patients Differ in Terms of Unobserved (to Researcher) Health or Tastes

- Finkelstein, Gentzkow, and Williams (2014) use a large sample of elderly patients who move. Conclude that 40-50% of the variation is patient specific, but this mostly reflects differences in health rather than differences in tastes.
- Cutler et al. (2019) use vignettes and conclude that much variation in medical care is driven by differences in physician beliefs, not differences in patients.

Defensive Medicine: Do Doctors Do Too Much to Protect Against Being Sued for Doing Too Little?

- Baicker and Chandra (2006) find no evidence that treatment use responds to malpractice liability at the state level, except for some screening procedures.
- Currie and MacLeod (2008) find that limiting malpractice liability through tort reform *increases* unnecessary C-sections.
- The marginal women who receive C-sections are actually the low risk. If the marginal C-section does more harm than good, then limiting liability *increases* procedure use.

Profit motive and other incentives

- Price matters. E.g. Many studies have shown that doctors are more likely to do C-sections when the gap in fees between C-section and normal deliveries is larger (e.g. Gruber, Kim, and Mayzlin, JHE 1999, see also Clemens and Gottlieb, 2014).
- Chan (2018) – patients arriving at the ER at the end of a shift have shorter stays and get more tests and treatments. Also are more likely to be admitted to hospital and have higher costs.

Peers and Area-level Specialization

- Preceding explanations can't explain why doctors behave differently in similar situations.
- Chandra and Staiger (2007) develop a model in which physicians learn from colleagues.
- Some areas specialize in high intensity and some in low intensity treatments.
- High intensity areas are better at the high intensity procedure and vice versa (practice makes perfect).
- *Implies more uniformity within areas than across areas, as well as convergence of doctors to a regional practice style over time.*

Evidence re: Spillovers from Peers is Mixed

- Epstein and Nicholson (2009), Dranove, Ramanarayanan and Sfekas (2011) examine spillovers in C-section rates. Find no convergence in practice styles among physicians within a hospital.
- Chan (2016) finds that the practice style of attending physicians has little impact on the the practice style of physicians junior to them in the same hospital.

Evidence re: Spillovers is Mixed

- Ahomaki et al. (2020) Effect of a Precautionary letter to all physicians in Finland who had prescribed >tablets of paracetamol-codeine to a new patient? 12.8% reduction in treated group. But is this a peer effect or a threat effect?
- Molitor (2018) studies cardiologists who move and finds rapid convergence to practice style of destination within one year.

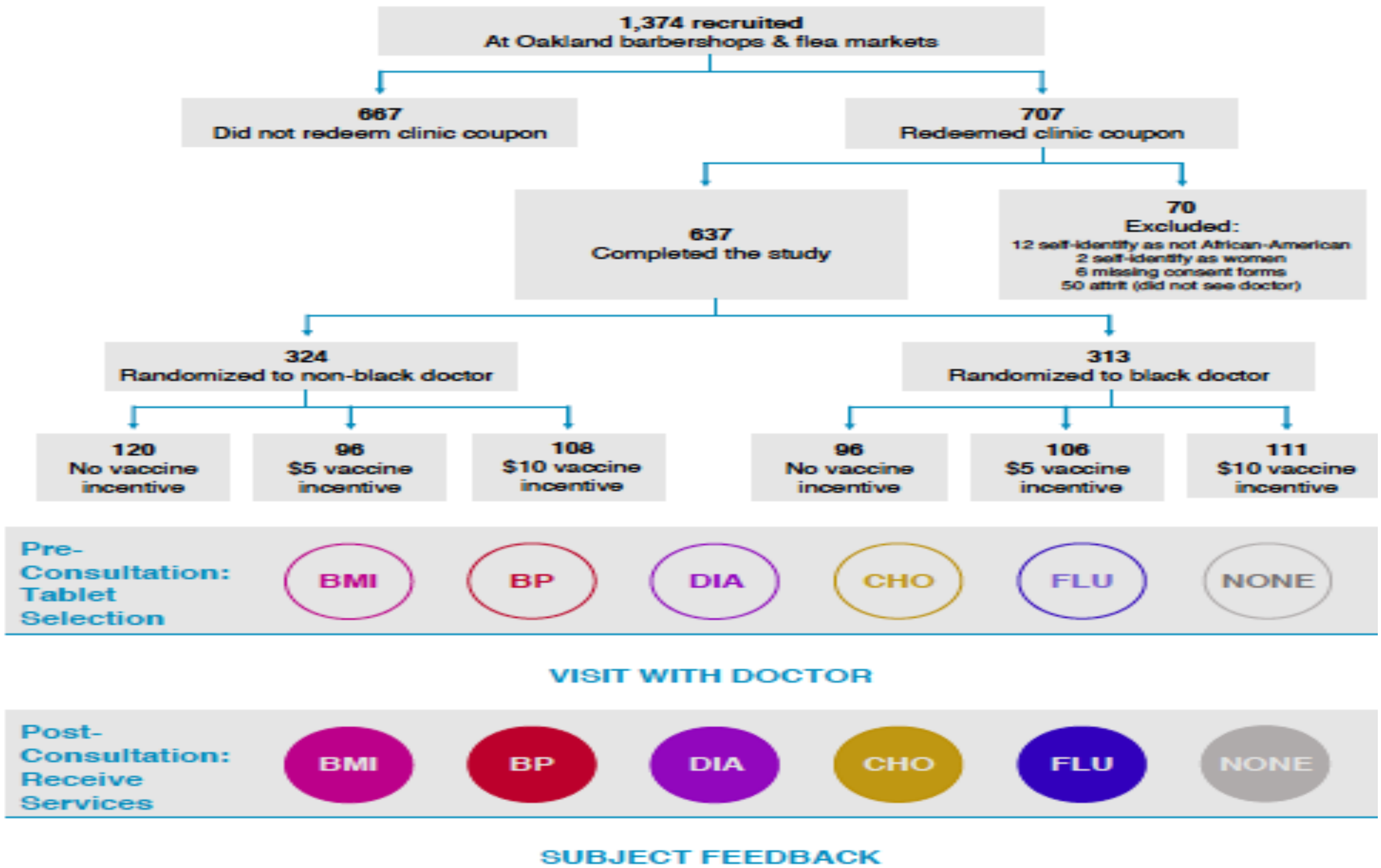
→ Spillovers may be more important in some contexts than others.

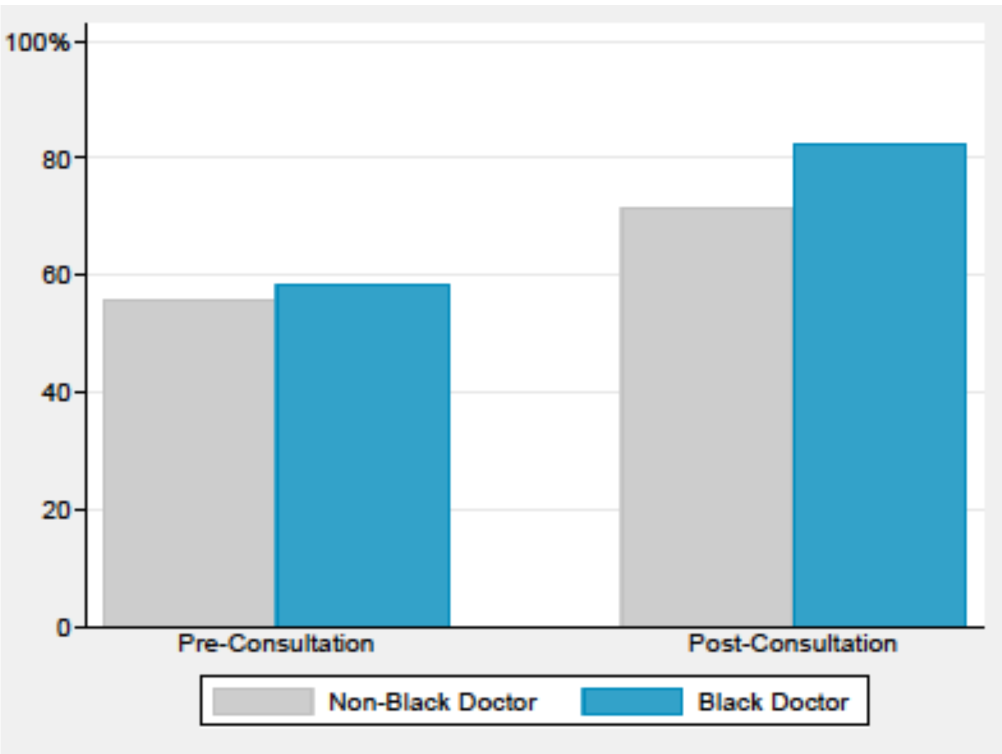
Discrimination/Communication/Affinity

- The same doctor may treat patients differently depending on factors like gender, race, or education.
- Alsan et al. (2019) created clinics in Oakland CA staffed by white and Black doctors and recruited low income Black patients.
- Doctors were to provide preventive care services.
- Importantly, patients were given information about the doctor including a photograph before seeing them and were also asked about their preferences about preventive care services.
- So experiment measures doctors success in persuading patients to receive recommended services.

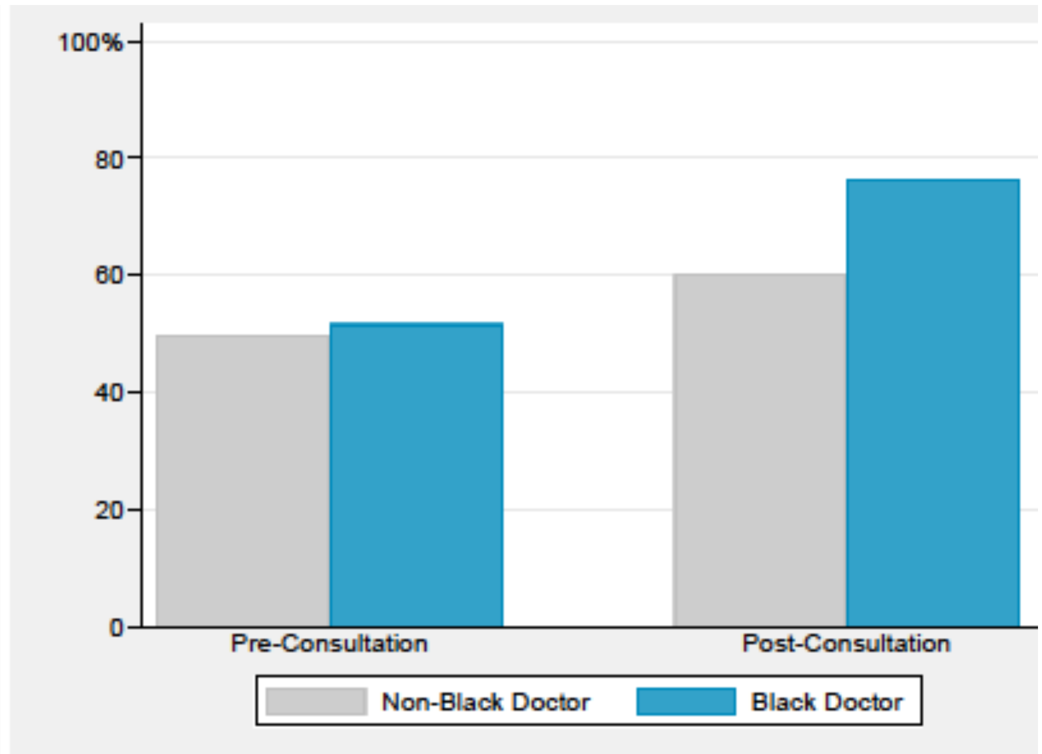
Alsan et al.

Figure 1: Study Design and Flow

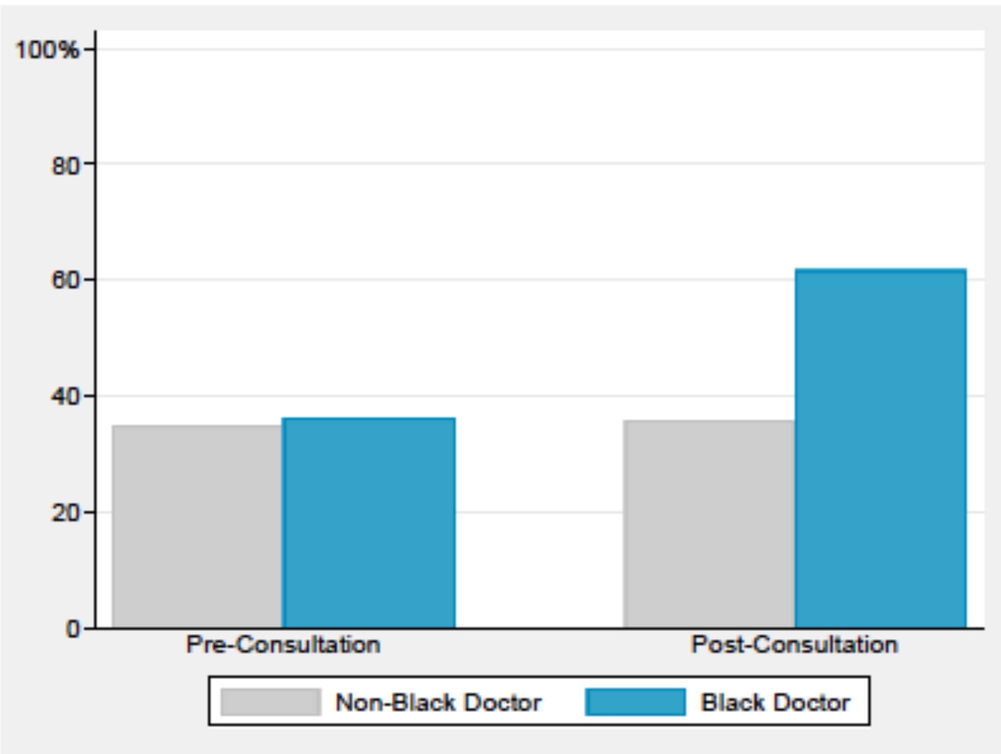




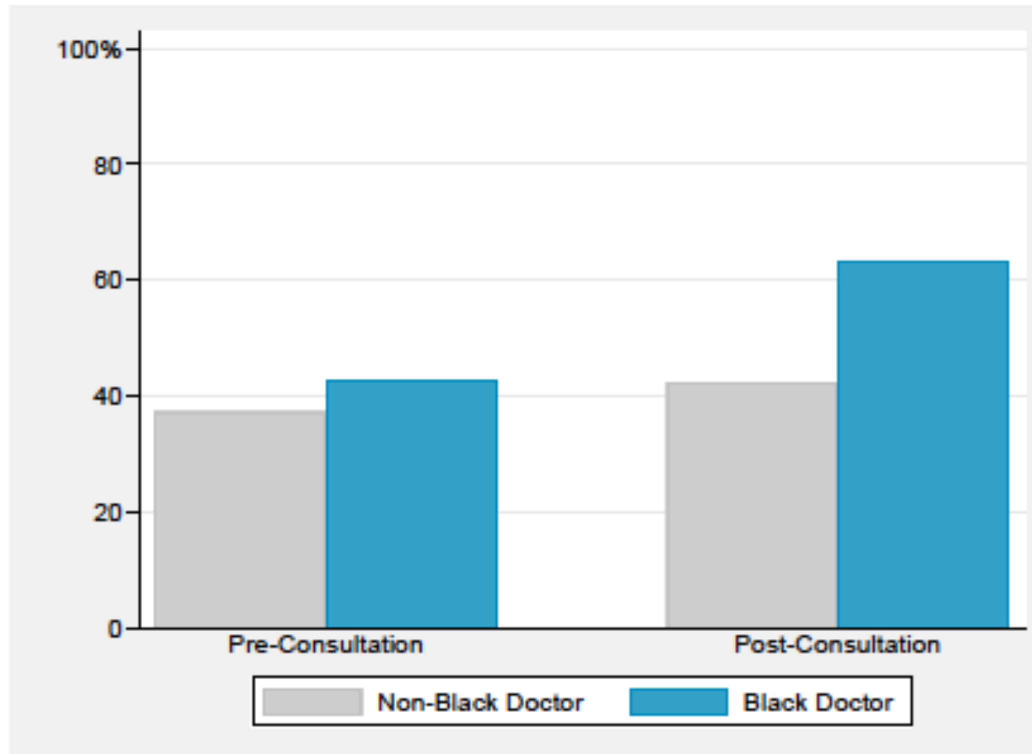
(a) Blood Pressure



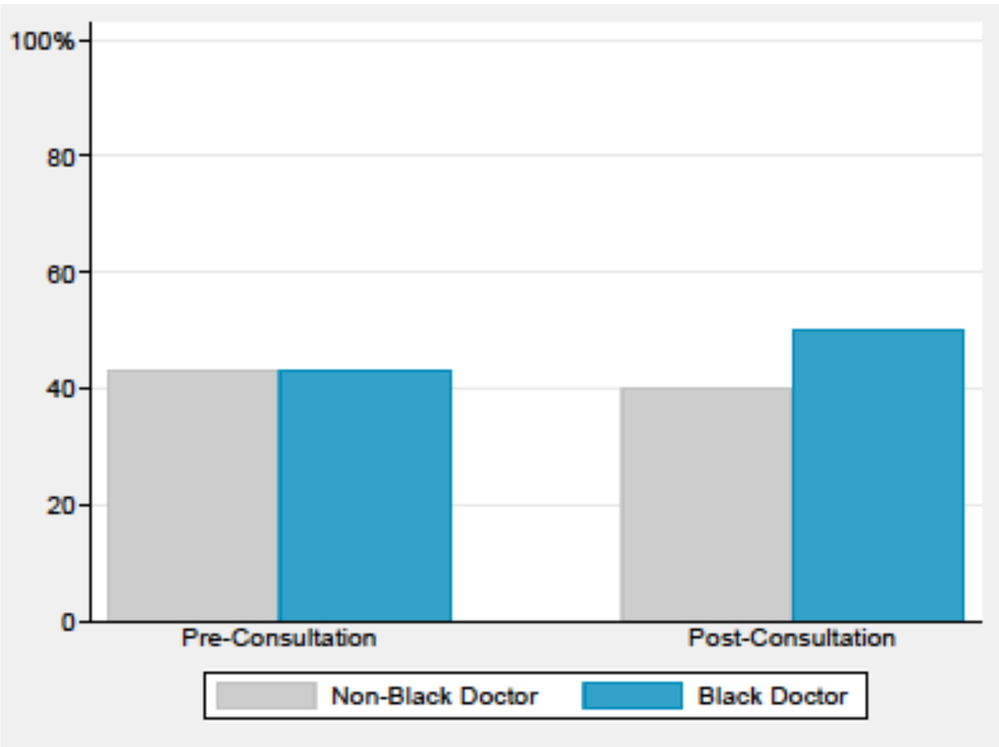
(b) BMI



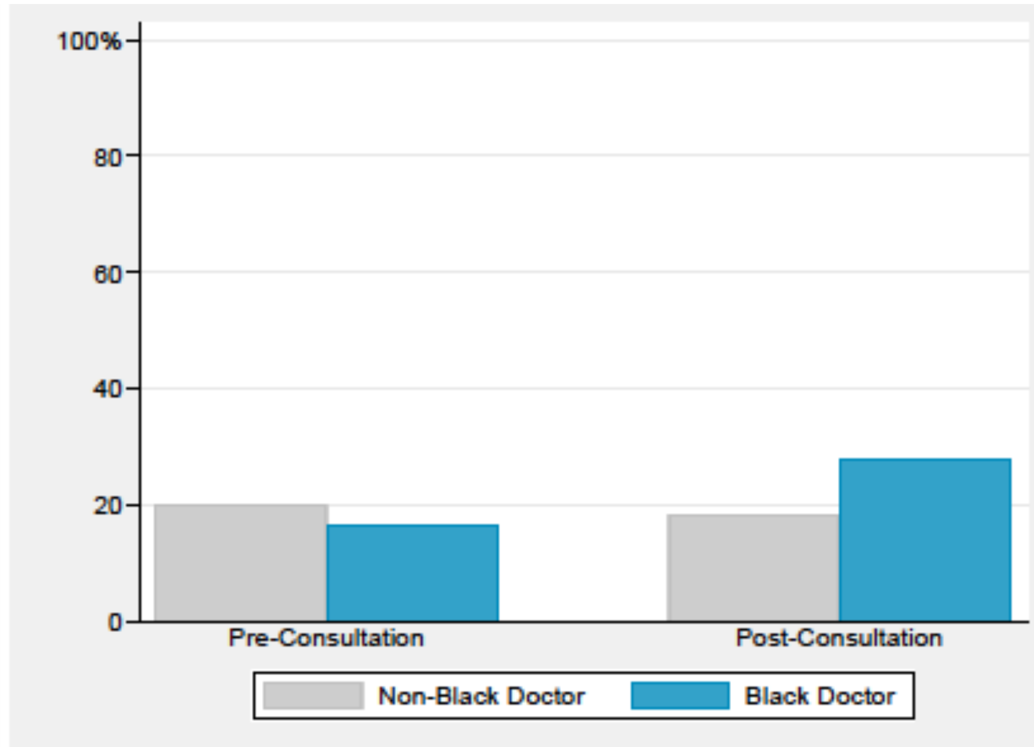
(c) Cholesterol



(d) Diabetes



(e) Flu Shot: With Incentive



(f) Flu Shot: Without Incentive

Cabral and Dillender (2021)

- Applicants for workman’s compensation are randomly assigned to doctors who perform examinations for the system.
- Female patients assigned to a female doctor were more likely to be rated as disabled, and receive 8.5% higher benefits.
- For male, patients, the gender of the physician did not matter.
- Women also showed a preference for female doctors for their continuing treatment.
- A Qualtrics survey (of the general population) reports that many women feel that their concerns are ignored or dismissed by male doctors.

Will spend the bulk of the lecture on:

- Varying skills
- Training
- Cognitive limitations

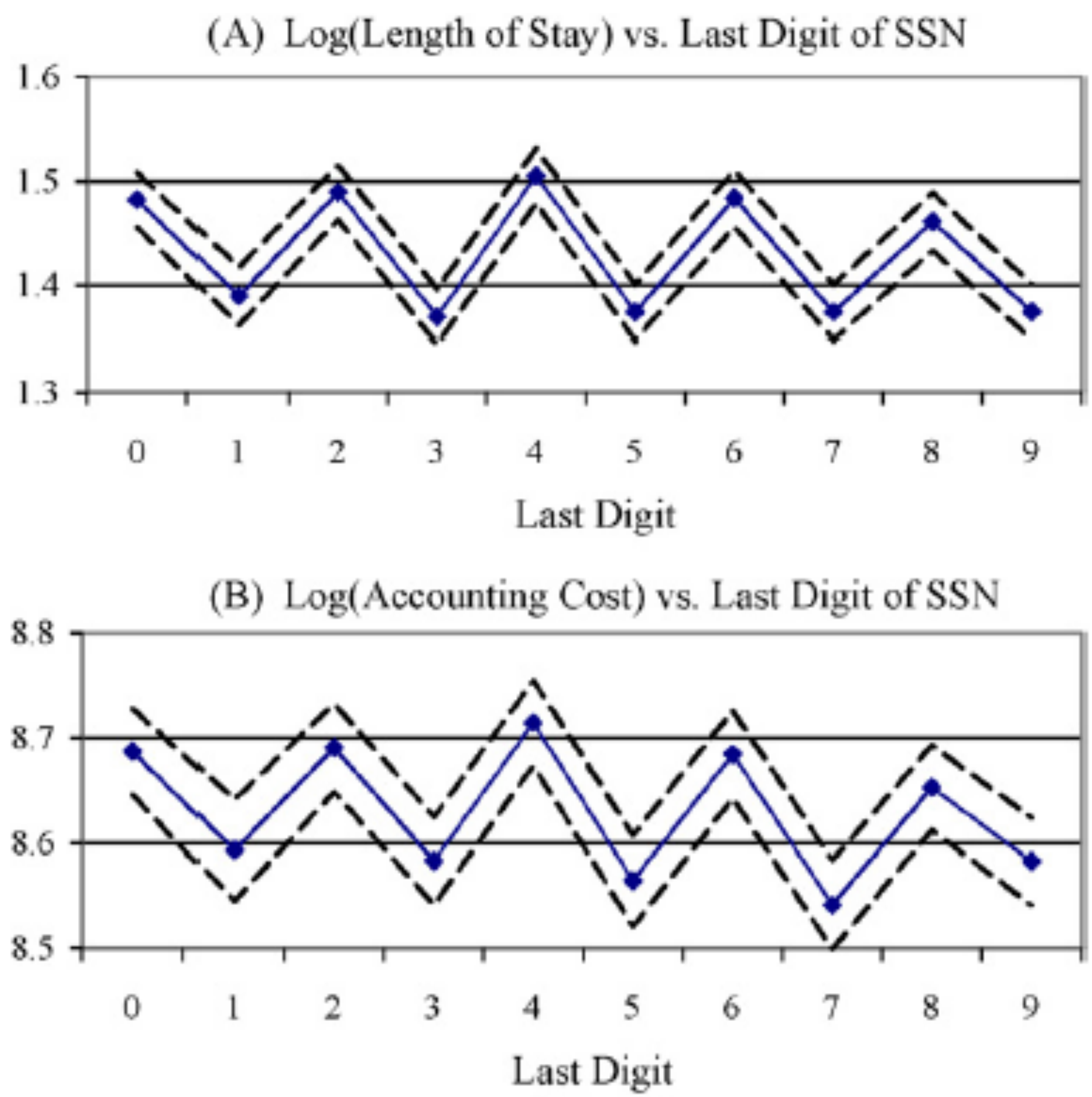
Those as we will see it can be difficult to separate these effects, especially skills vs. training given that doctors are selected into medical schools and residency programs.

Random Assignment to Doctors Rarely Occurs

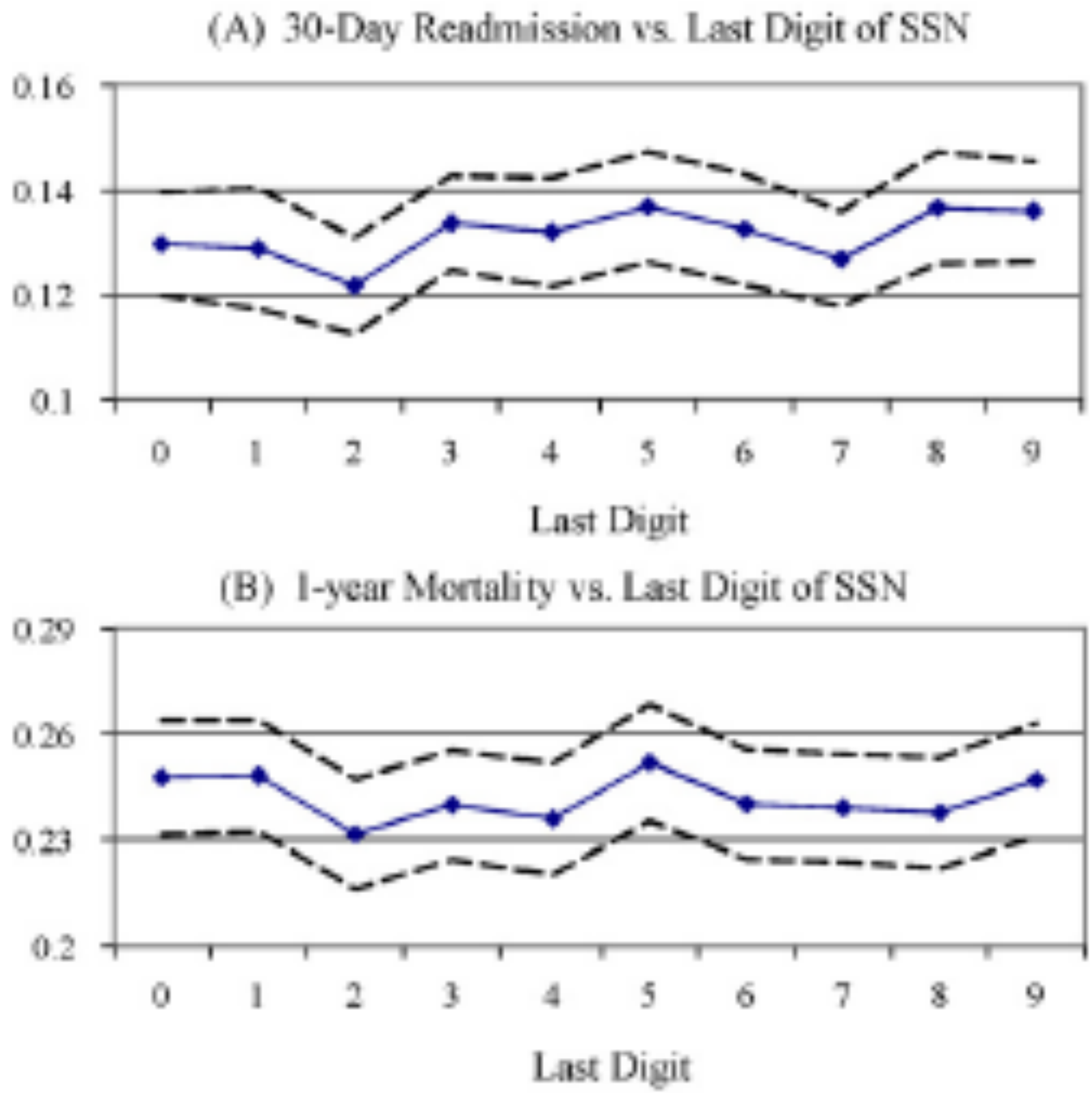
- Doyle, Ewer and Wagner (2010) looks at hospital patients within one hospital treated by randomly assigned teams from higher and lower ranked hospitals.

	Program A	Program B
Medical College Admissions Test (MCAT) Ranking	Top 5	Top 50
NIH Funding Ranking	Top 5	Top 80
US News Honor Roll (overall)	Top 10	Not listed
% with MD from Top 10 Medical School (US News rankings)	30%	3%
% with MD from Top 25 Medical School (US News rankings)	50%	9%
% with MD from Top 10 Medical School (NIH Funding rankings)	25%	2%
% with MD from Top 25 Medical School (NIH Funding rankings)	40%	8%
% Foreign Medical School	10%	20%
American Board of Internal Medicine	99% (95th percentile)	85% (20th percentile)
American Board of Surgery	85% (75th percentile)	60% (20th percentile)

Patients Randomized by Last Digit of SSN Odd or Even



The B-team takes longer and orders more tests, but arrives at the same health outcomes



Currie and Zhang (2022)

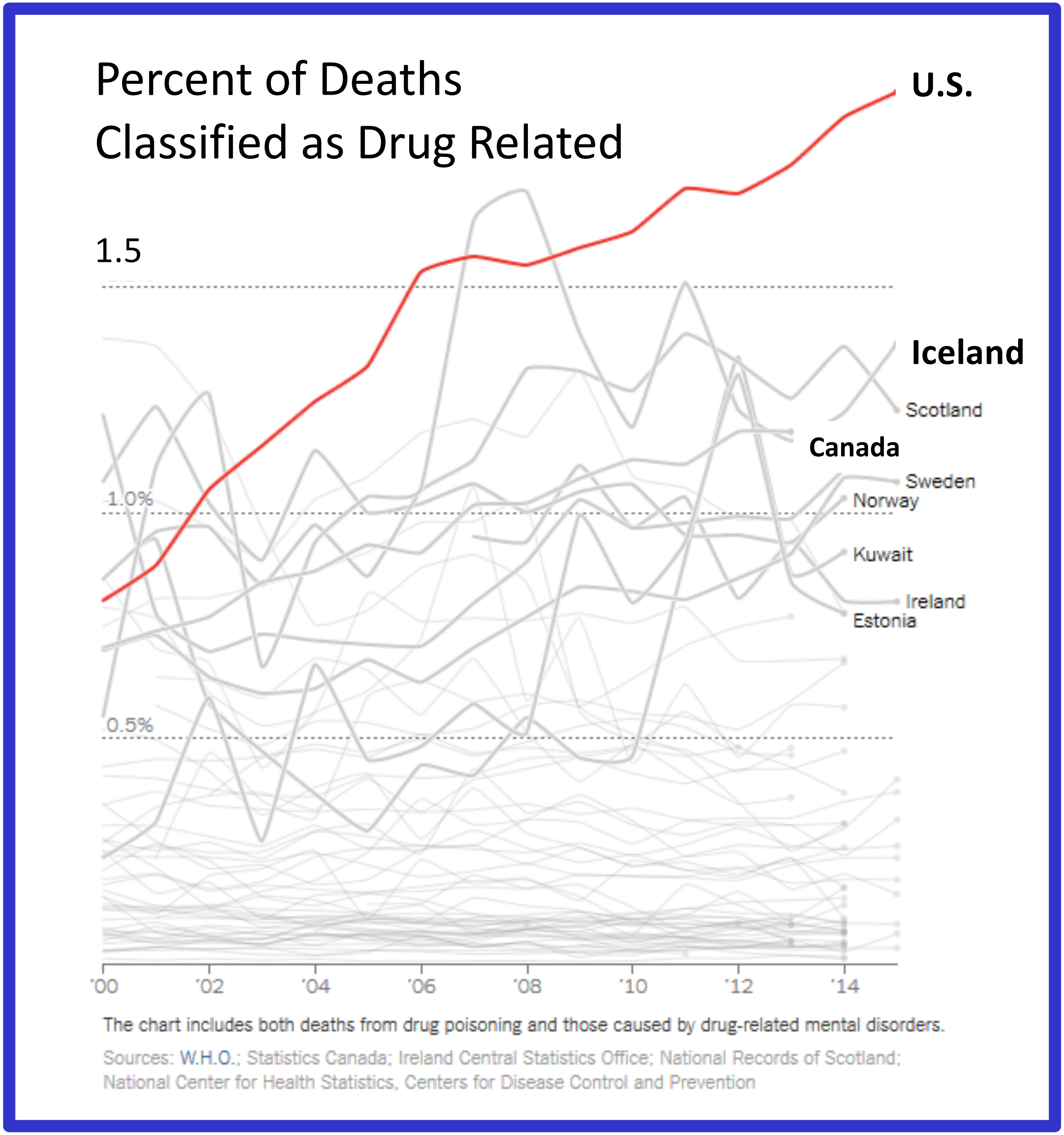
- Also take advantage of random assignment within the Veteran's administration system, this time to primary care physicians.
- Construct measures of physician effectiveness by looking at patient outcomes for cardiovascular disease, mental health, and preventable hospitalizations 3 years after assignment. For each patient, use a "leave-out" measure based on the doctor's other patients.
- Like Doyle et al. (2010), find that more effective doctors do more with less – fewer visits, fewer tests, but in this case, better outcomes.

Is the difference due to training or initial skill levels?

- Doyle (2020) shows that specialized training matters for patient outcomes.
- When heart failure patients enter the Emergency Department when more cardiologists are available, they are more likely to be treated directly by a cardiologist, have more invasive procedures, and are more likely to survive over the following year.
- However, cardiologists are not randomly selected from the pool of medical school graduates, so this does not rule out the hypothesis that they may have higher skill even in the absence of special training.

Schnell and Currie (2018)

- Examine all opioid prescriptions in the U.S. from 2006-2014.
- Have a prescriber ID
- Show that there is much variation in prescribing practices within counties, specialties, and even within practice addresses.
- 48% of opioids prescribed by GPs though they get little to no training about opioids.
- Focus on physician training as a possible explanation for variation in practice style.

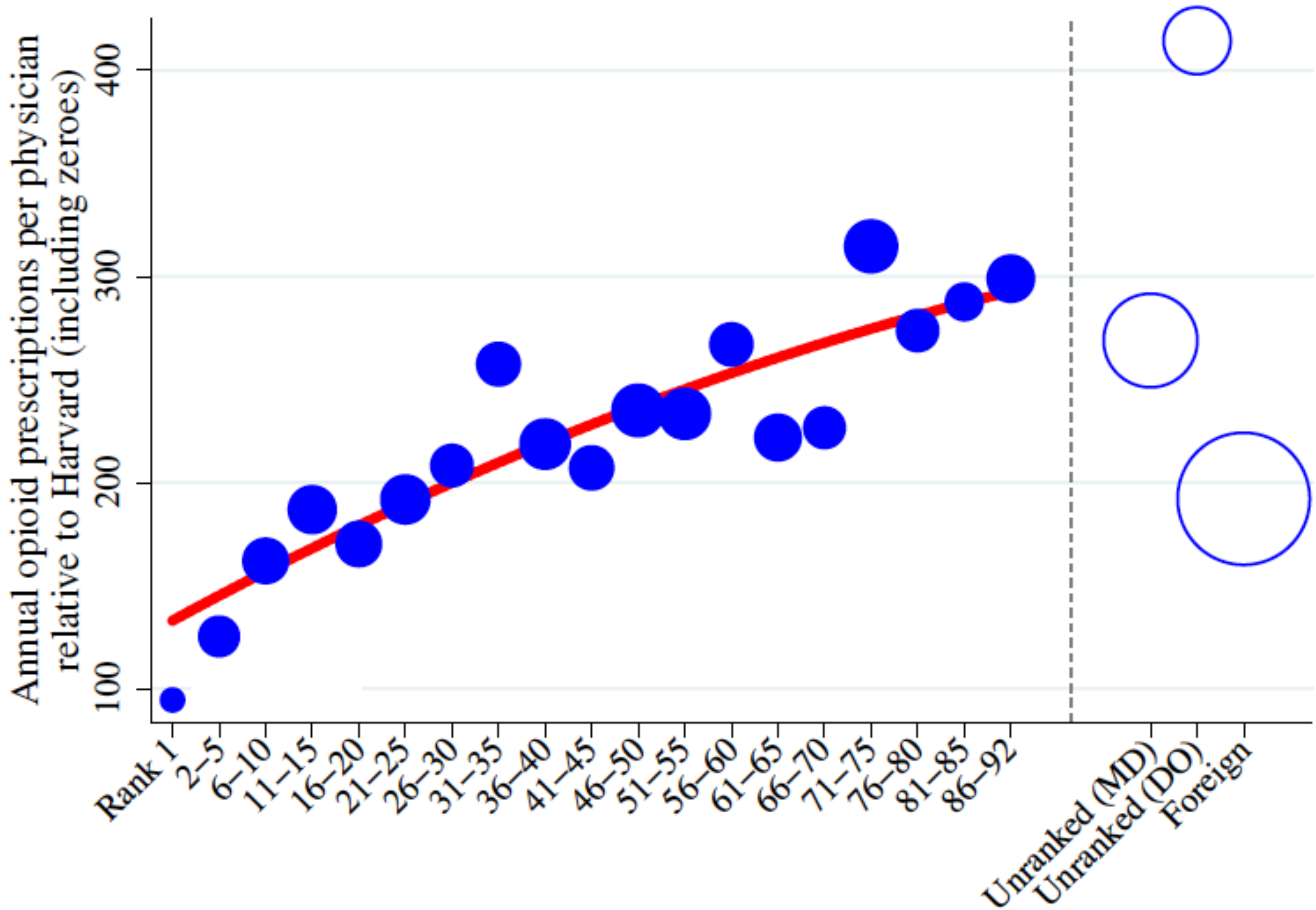


Source: New York Times

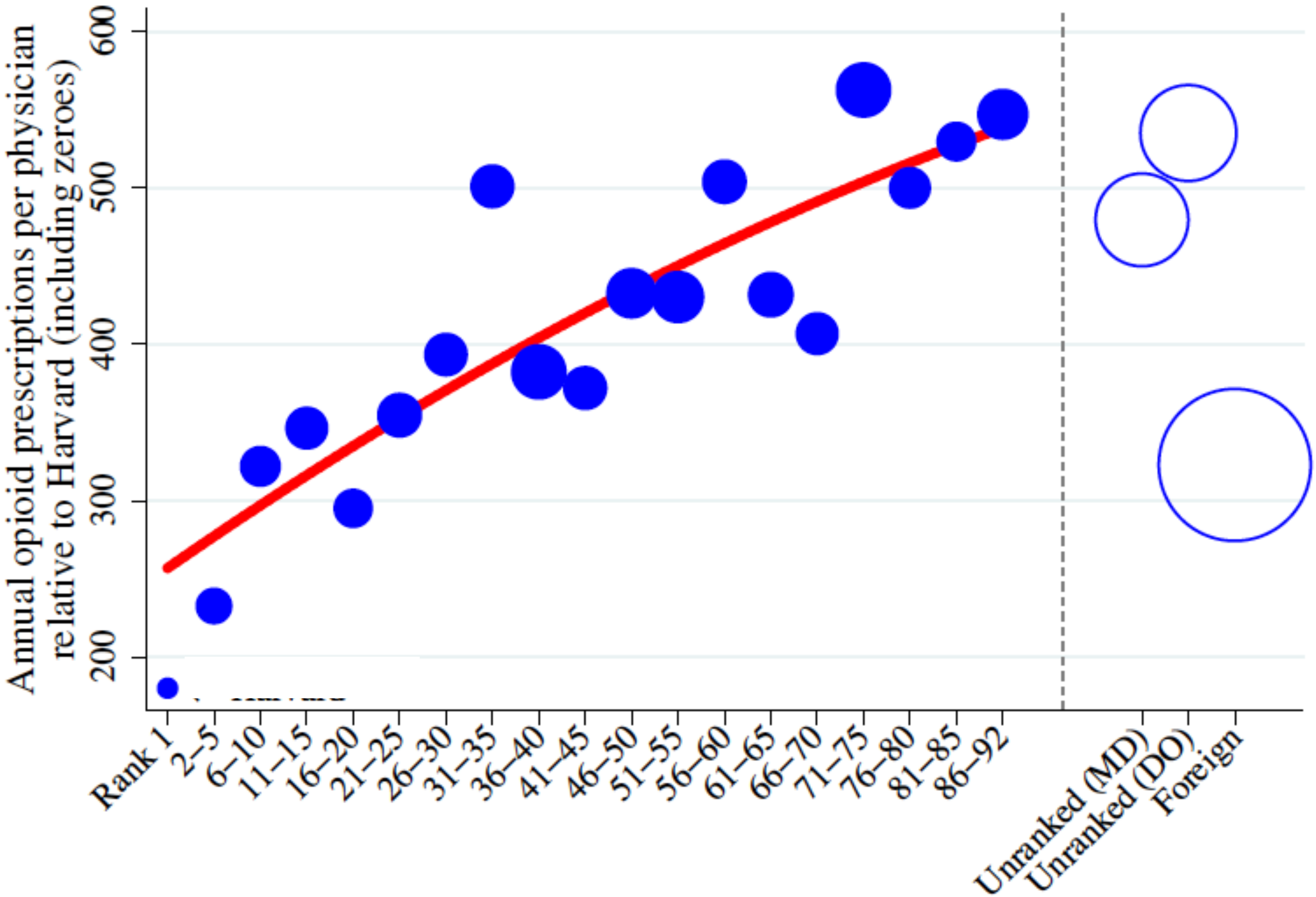
Opioid Prescriptions Per Capita and Deaths Involving Drugs

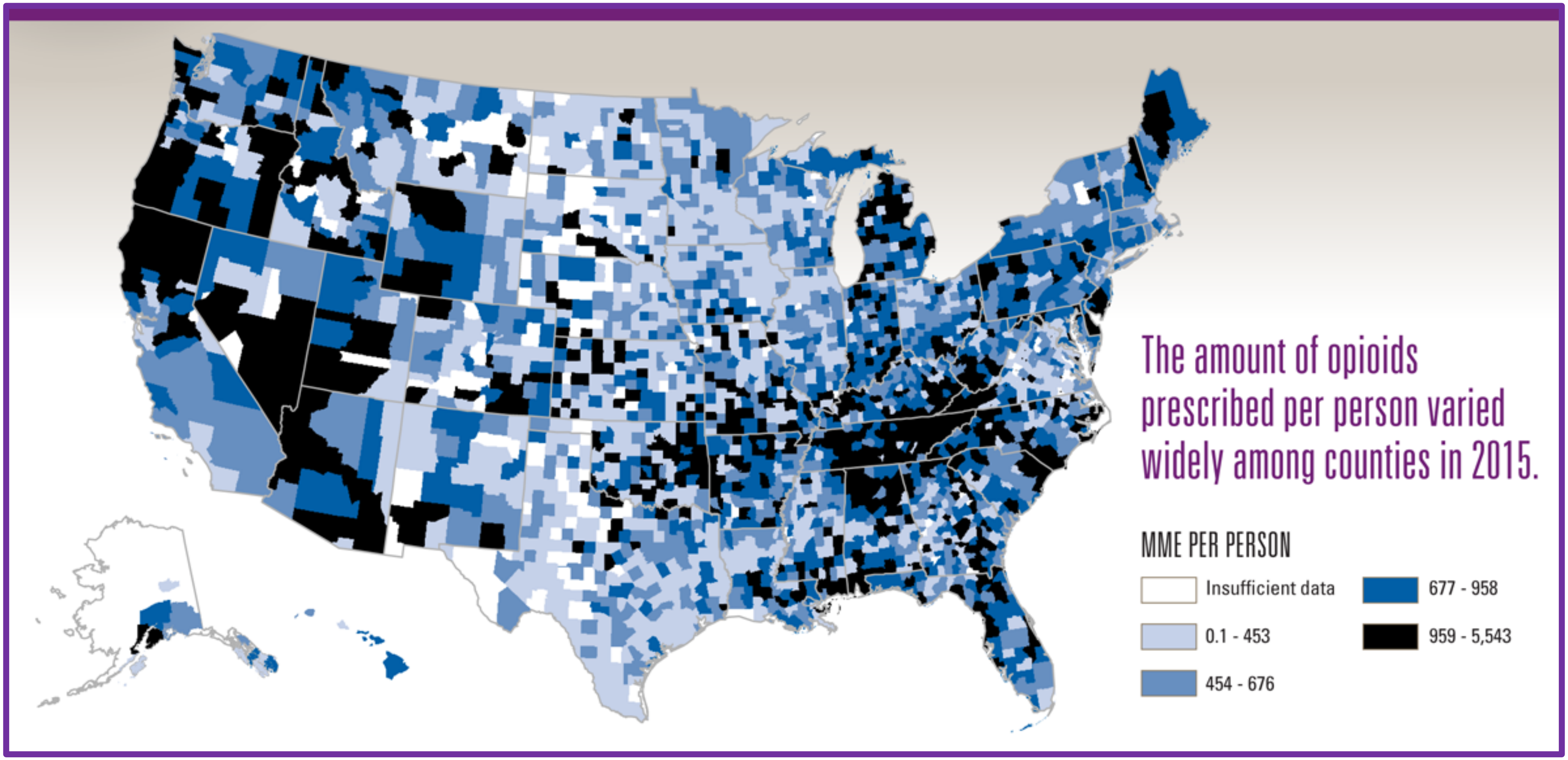
	ln(Deaths involving drugs per capita)	
	(1)	(2)
ln(Opioid prescriptions per capita)	0.286***	0.150***
	(0.026)	(0.040)
County FEs	No	Yes
N (county-years)	22,801	22,801
R^2	0.118	0.706

Opioid Prescriptions by Medical School Rank: All Physicians



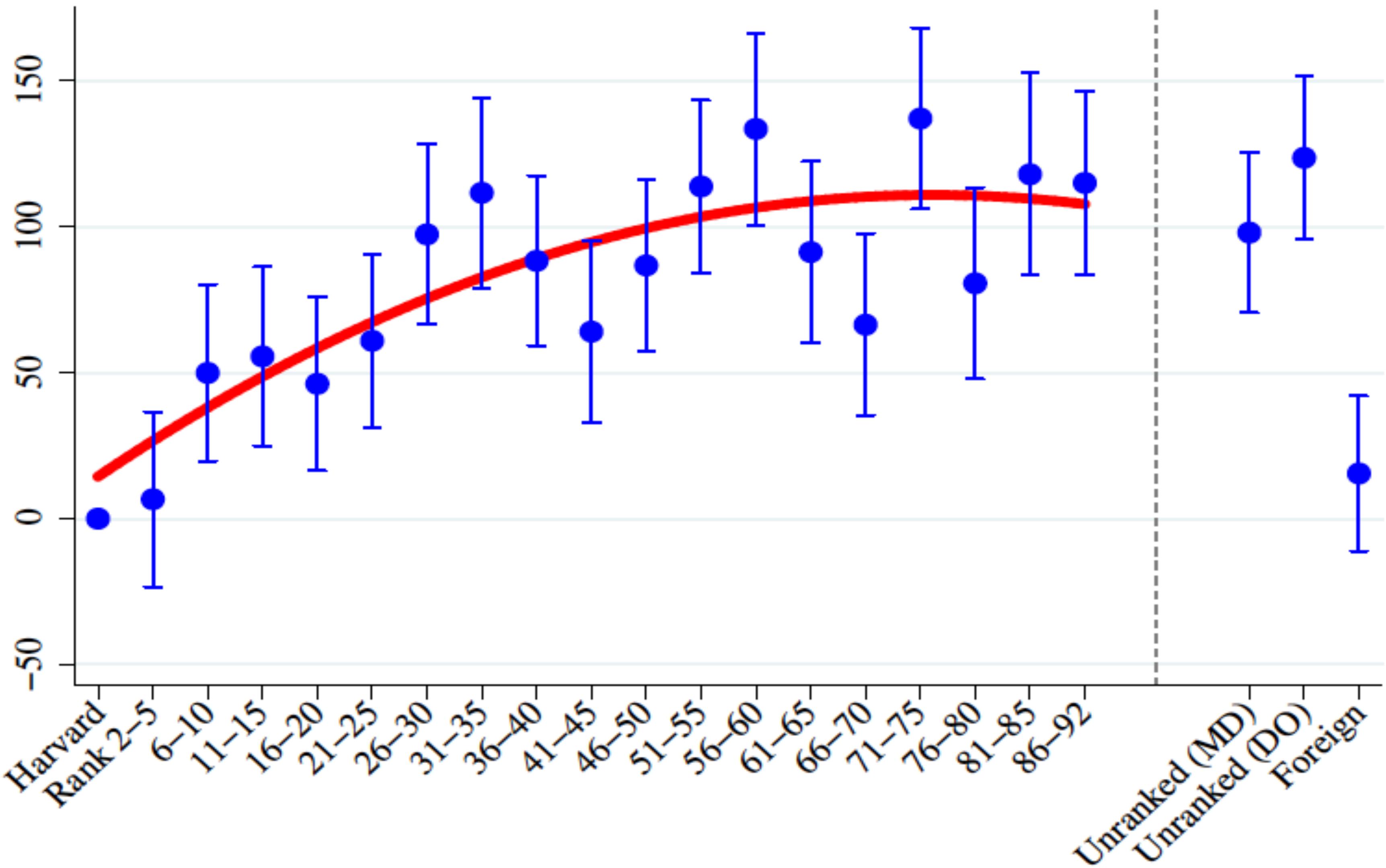
Opioid Prescriptions by Medical School Rank: General Practitioners





Source: CDC.gov and New York Times

Opioid Prescriptions by Medical School Rank: General Practitioners, Controlling for County



Annual Opioid Prescriptions on Medical School Rank (all physicians)

	(1)	(2)	(3)	(4)	(5)	(6)
Medical rank	4.147***	2.995***	2.644***	2.784***	2.418***	1.441***
	(0.309)	(0.307)	(0.301)	(0.292)	(0.292)	(0.257)
$(\text{Medical rank})^2$	-0.011***	-0.003	-0.015***	-0.021***	-0.018***	-0.014***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	202.380***	321.419***	-8.3e+03***	295.736***	354.644***	362.420***
	(5.818)	(6.521)	(1297.679)	(5.712)	(6.264)	(5.713)
Specialty FEs	No	Yes	No	No	Yes	Yes
County demographics	No	No	Yes	No	No	No
County FEs	No	No	No	Yes	Yes	No
Practice address FEs	No	No	No	No	No	Yes
N (physician-years)	832,005	832,005	832,005	832,005	832,005	832,005
R^2	0.014	0.029	0.096	0.174	0.178	0.636

Medical School Rank and Prescribing by Specialty

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Specialty:	General practice	Orthopaedic surgery	Emergency medicine	Pain medicine	Phy. med. & rehab.	Ob./gyn.	Anesthe-siology	General surgery
Medical rank	2.418***	1.920**	-0.631	-3.814	-4.467	0.658***	0.788	0.650***
	(0.292)	(0.846)	(0.368)	(9.275)	(4.110)	(0.181)	(0.865)	(0.244)
(Medical rank)^2	-0.018***	-0.018*	0.013***	0.038	0.067	-0.005**	-0.003	-0.003
	(0.003)	(0.009)	(0.004)	(0.101)	(0.044)	(0.002)	(0.010)	(0.003)
Constant	354.644***	537.923***	70.700***	1507.539***	301.616	87.348***	-26.520	126.145***
	(6.264)	(20.956)	(20.580)	(207.711)	(328.576)	(22.719)	(23.612)	(20.778)
N (phys-years)	832,005	155,547	187,785	15,318	33,462	213,282	162,225	158,913
R^2	0.178	0.195	0.245	0.356	0.288	0.210	0.118	0.235

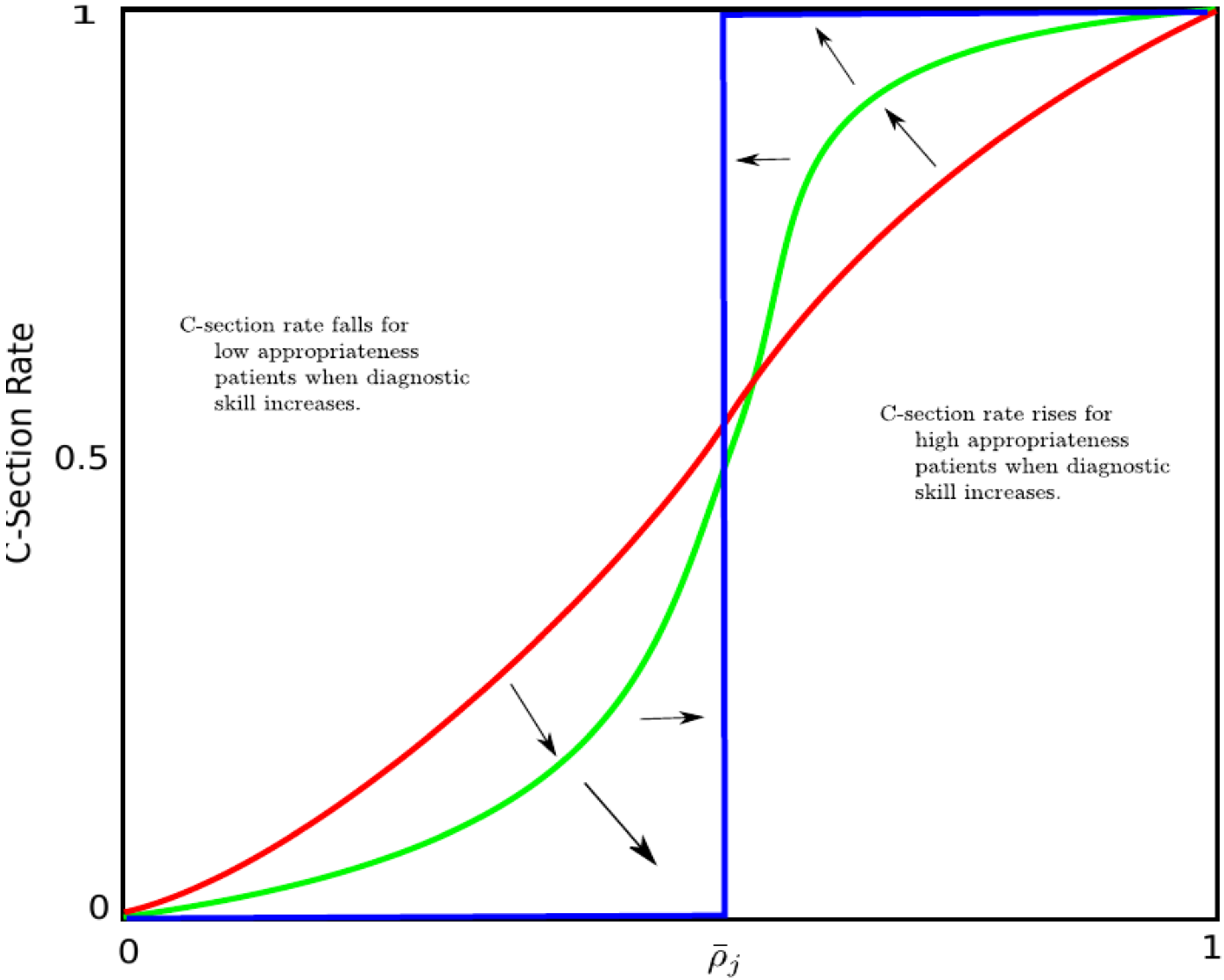
Physicians Trained at the Lowest Ranked Schools:

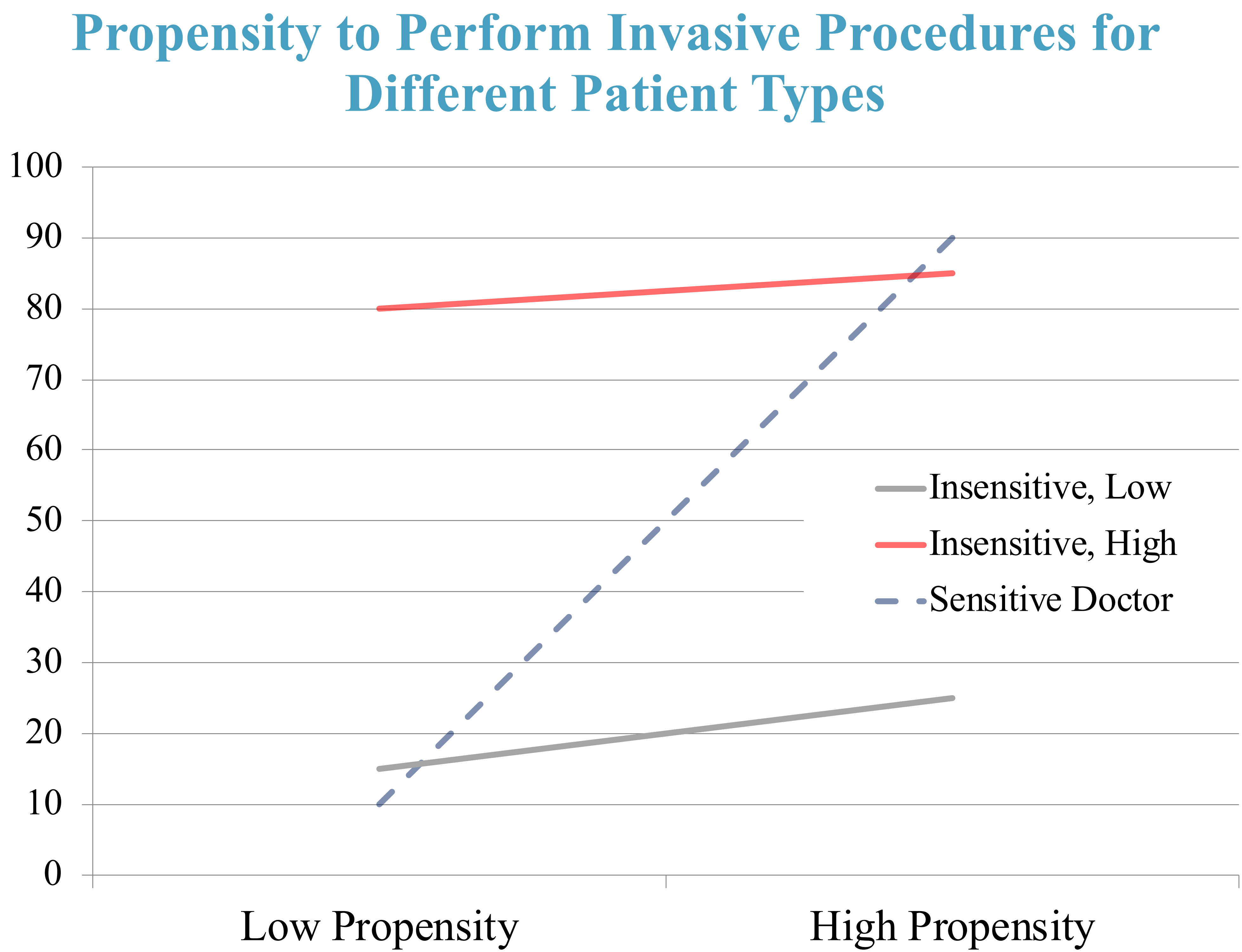
- Are more likely to prescribe any opioids.
- Prescribe 3x more opioids per year than those trained at the highest ranked schools.
- GPs responsible for 48% of all opioids prescribed. If all GPs practiced like those at the top schools, we would have had 56.5% fewer opioids prescriptions and 8.5% fewer deaths over the 2006-2014 period.
- These relationships hold within specialties and within practice addresses.
- But school rank doesn’t matter in specialties that receive additional training in use of opioids.

Currie and MacLeod (2017) and Currie, MacLeod and Van Parys (2016) conceptualize skill as the correct matching of patients to procedures

- First use administrative hospital data to rank patients with many different clinical characteristics using single index of their *a priori* appropriateness for a procedure.
- In this single-index problem, the doctor's problem is to determine a cutoff value above which patients get the procedure and below which they do not.

Consider the Relationship Between Patient Condition and a 0/1 Procedure Choice




Propensity to Perform Invasive Procedures for Different Patient Types
100
90
80
70
60
50
40
30
20
10
0
Insensitive, Low
Insensitive, High
Sensitive Doctor
Low Propensity
High Propensity

Steps to Assess Doctor Skill at Matching

- To calculate an index (risk score) using administrative hospital data and and rank patients by their appropriateness for an intensive procedure.
- Then estimate doctor specific regressions, to see how each doctor responds to the index.
- The constant measures mean propensity to do the procedure, while the slope measures the doctor's responsiveness to the patient's condition.

Application 1: C-section
(Currie and MacLeod, 2017)

- One of the most common surgical procedures.
- There are believed to be too many: In the U.S. 40% vs. 15% WHO recommendation.
- Many policies to reduce C-section have been discussed.
- It is possible to identify women who are good candidates for C-section prior to delivery.
- We can observe outcomes for both mothers and children in administrative hospital data.

Data

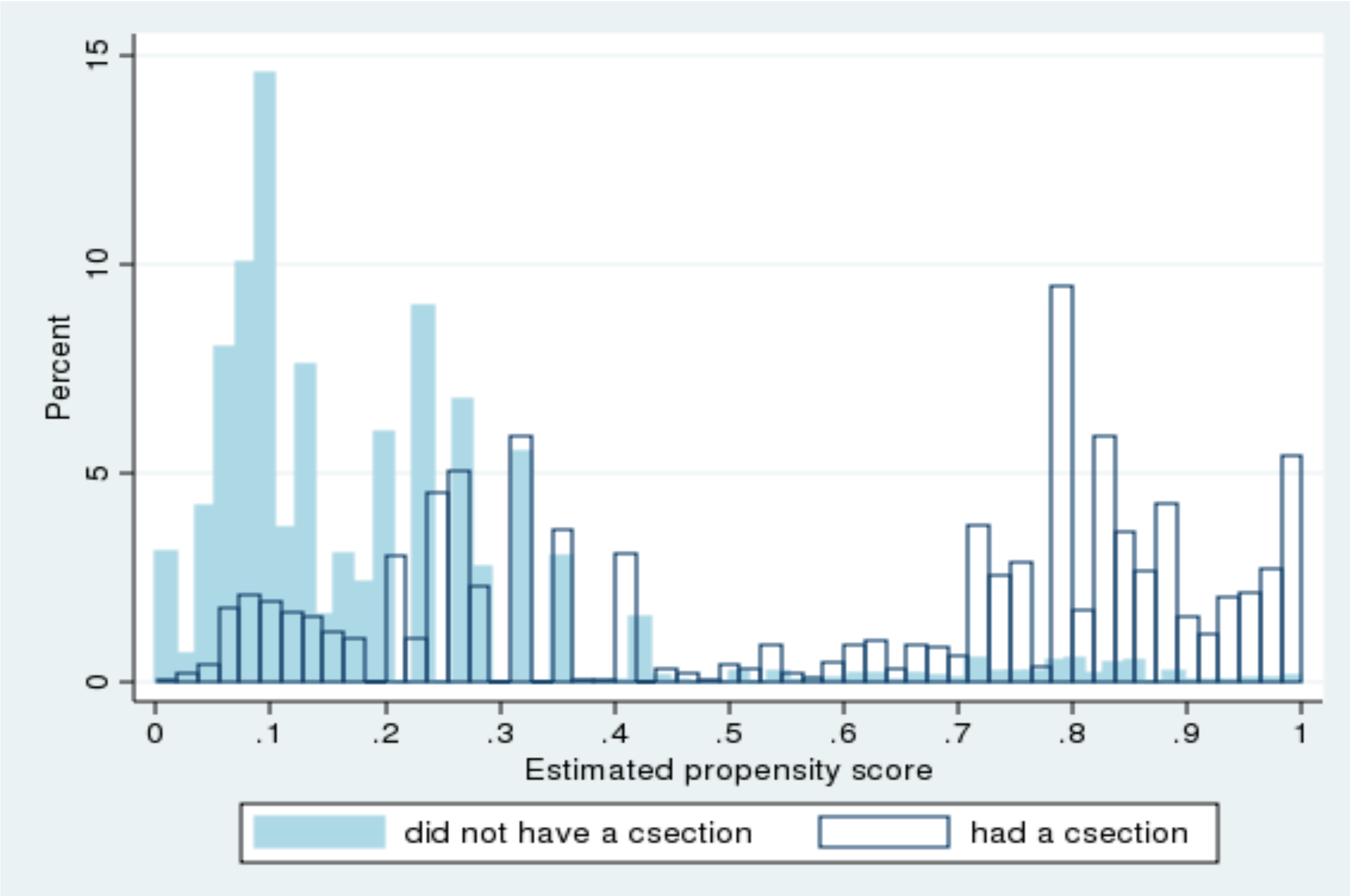
1,000,000 Electronic Birth Records for New Jersey, 1997-2006. Includes:

- risk factors for C-section (e.g. previous C-section, breech, medical conditions)
- delivery method
- maternal and child outcomes
- codes for physician and hospital
- demographics and residential address

Estimating Medical Risk for C-section

- Estimate a logit for C-section using all 10 years of data (could use machine learning).
- Medical risk factors include: age, previous C-section, parity, multiple birth, risk factors for the pregnancy (including placenta previa, breech birth, hypertension, diabetes etc.)
- Model explains ~1/3 of variation in outcomes.

This single model index is highly predictive

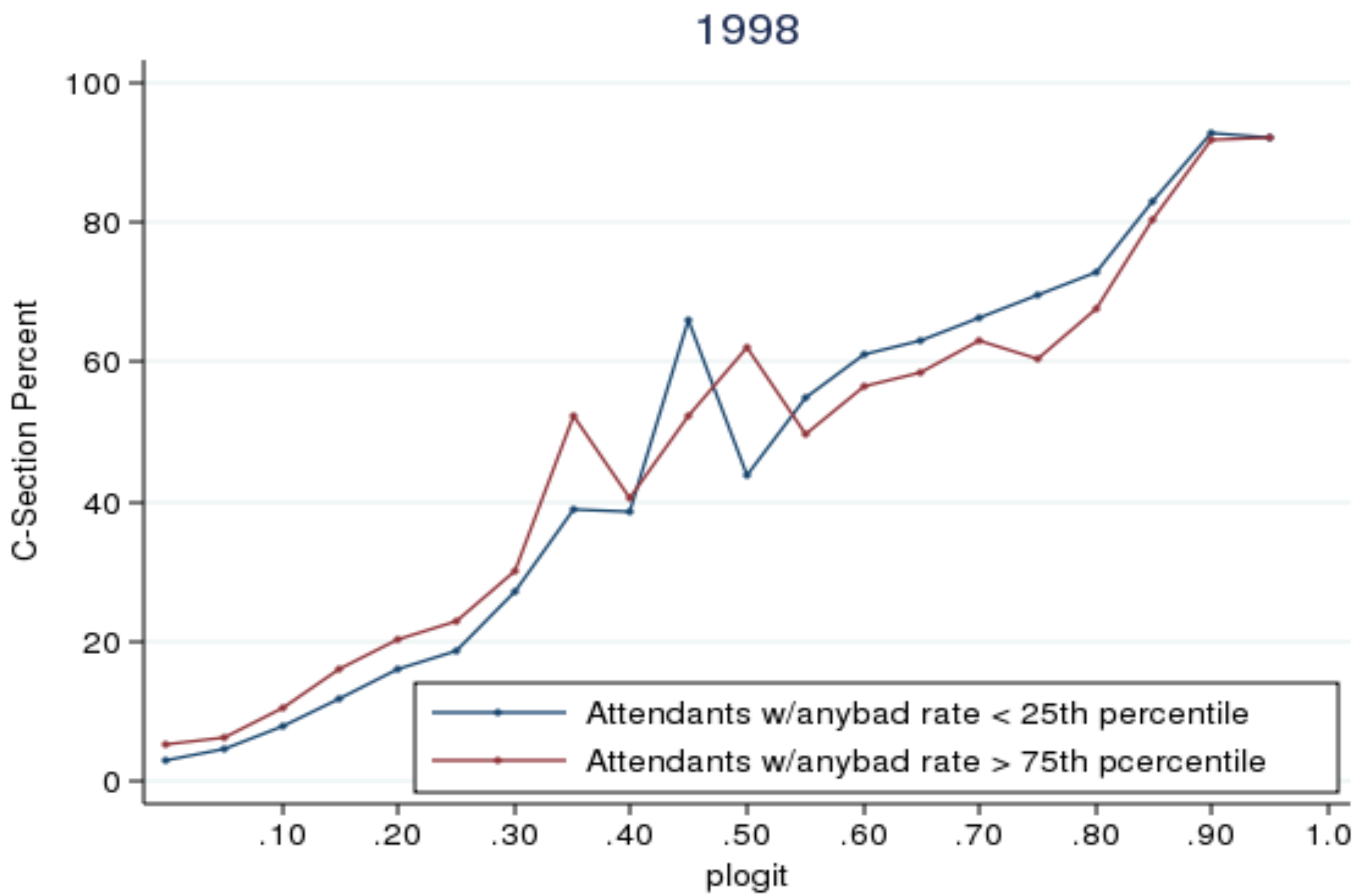


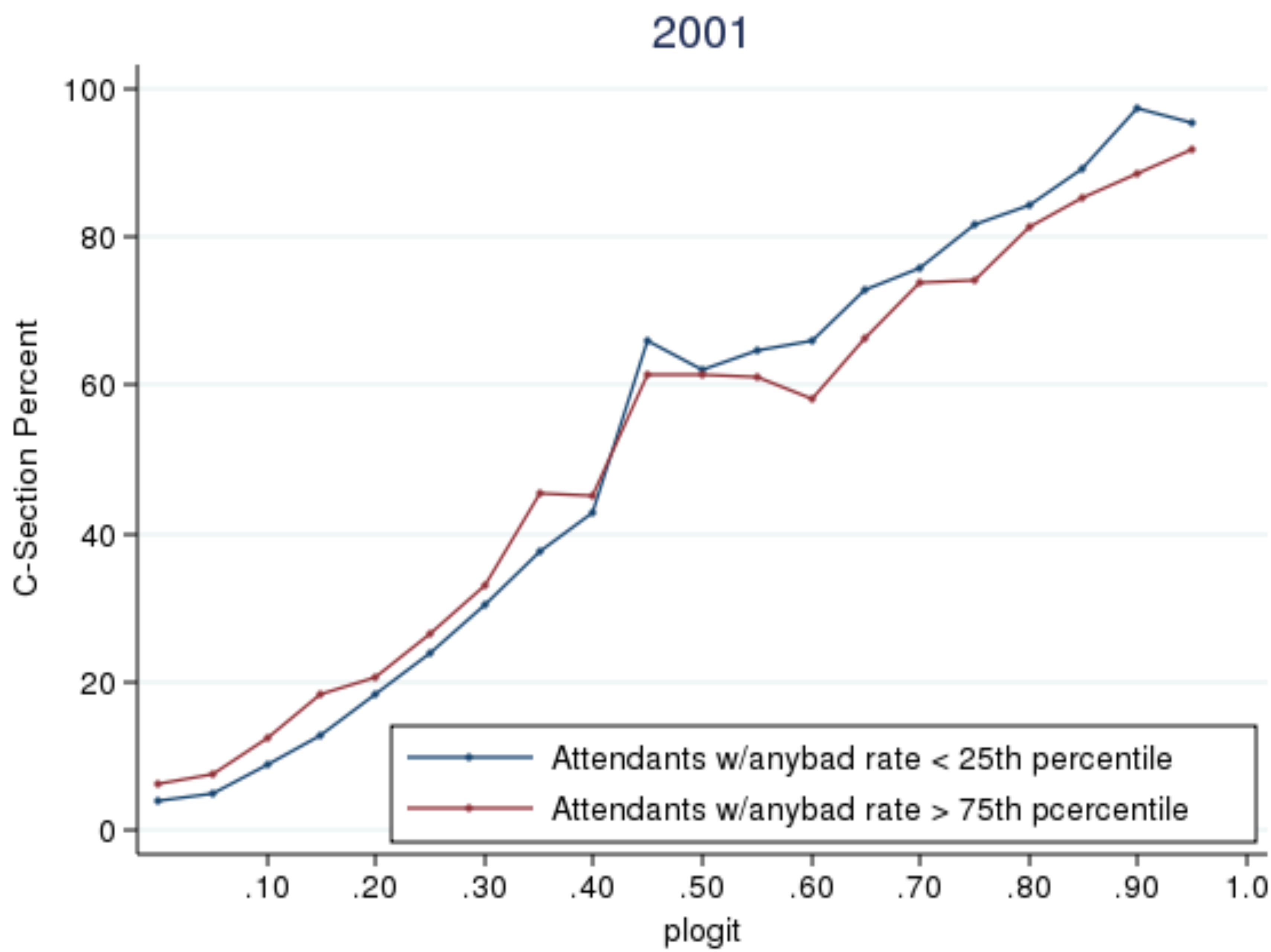
The Ranking of risk for C-section is robust:

- to using only early years of the data
- to using only the top quartile of doctors in terms of outcomes
- to omitting some variables such as previous C-section from the model

We interpret this stability as evidence that the a priori ranking of patients in terms of risk for C-section is not controversial. What may be controversial is where the risk cutoff should be.

C-section vs. Medical Risk for "Good" and "Bad" outcomes doctors




2001
100
80
60
40
20
0
C-Section Percent
.10
.20
.30
.40
.50
.60
.70
.80
.90
1.0
plogit
Attendants w/anybad rate < 25th percentile
Attendants w/anybad rate > 75th pcercentile

Measuring Decision Making and Procedural Skill

- For each doctor, regress probability of C-section on the patient propensity index for their sample of cases.
- The slope coefficient is an indicator of decision-making skill.
- Procedural skill can be proxied using the rate of bad outcomes for patients with a high propensity for C-section (top quartile) less the rate of bad outcomes for patients with low propensity (bottom quartile).
- Normalize using z-scores.

Mean doctor characteristics by mother's c-section risk

C-section Risk:	All	Low Risk	High Risk
<u>Doctor Characteristics</u>			
# Deliveries per doctor	1019	1030	1009
	(650)	(675)	(626)
Decision Making	0.000	-0.032	0.030
	(1.000)	(1.013)	(0.987)
Procedural Skill Differential	0.000	-0.016	0.014
	(1.000)	(1.026)	(0.974)
Market Price Differential	4.711	4.687	4.734
($1000s)	(1.606)	(1.590)	(1.621)
Share High Risk	0.122	0.116	0.127

Patient Characteristics by C-section Risk

C-section Risk:	All	Low Risk	High Risk
<u>Mother & Child Characteristics</u>			
African American	0.158	0.185	0.132
Hispanic	0.210	0.244	0.179
Married	0.713	0.645	0.776
High School Dropout	0.128	0.177	0.082
Teen mom	0.030	0.052	0.009
Mom Age 35 or More	0.238	0.221	0.254
Smoked	0.081	0.090	0.073
Child Male	0.513	0.514	0.513
Child First Born	0.398	0.200	0.584
Medicaid	0.206	0.260	0.155
# of Observations	968748	469170	499578

Mean Outcomes by C-section Risk

C-section Risk:	All	Low Risk	High Risk
<u>Outcomes</u>			
C-Section Rate	0.331	0.103	0.545
Any Bad Outcome	0.127	0.111	0.143
Bad Maternal Outcome	0.055	0.037	0.073
Bleeding, Fever, Seizures	0.039	0.024	0.053
Late Maternal Complications	0.019	0.014	0.024
Bad Child Outcome	0.080	0.080	0.081
Fetal Distress	0.071	0.073	0.069
Birth Injury	0.003	0.003	0.003
Neonatal death	0.004	0.003	0.006

Empirical Model

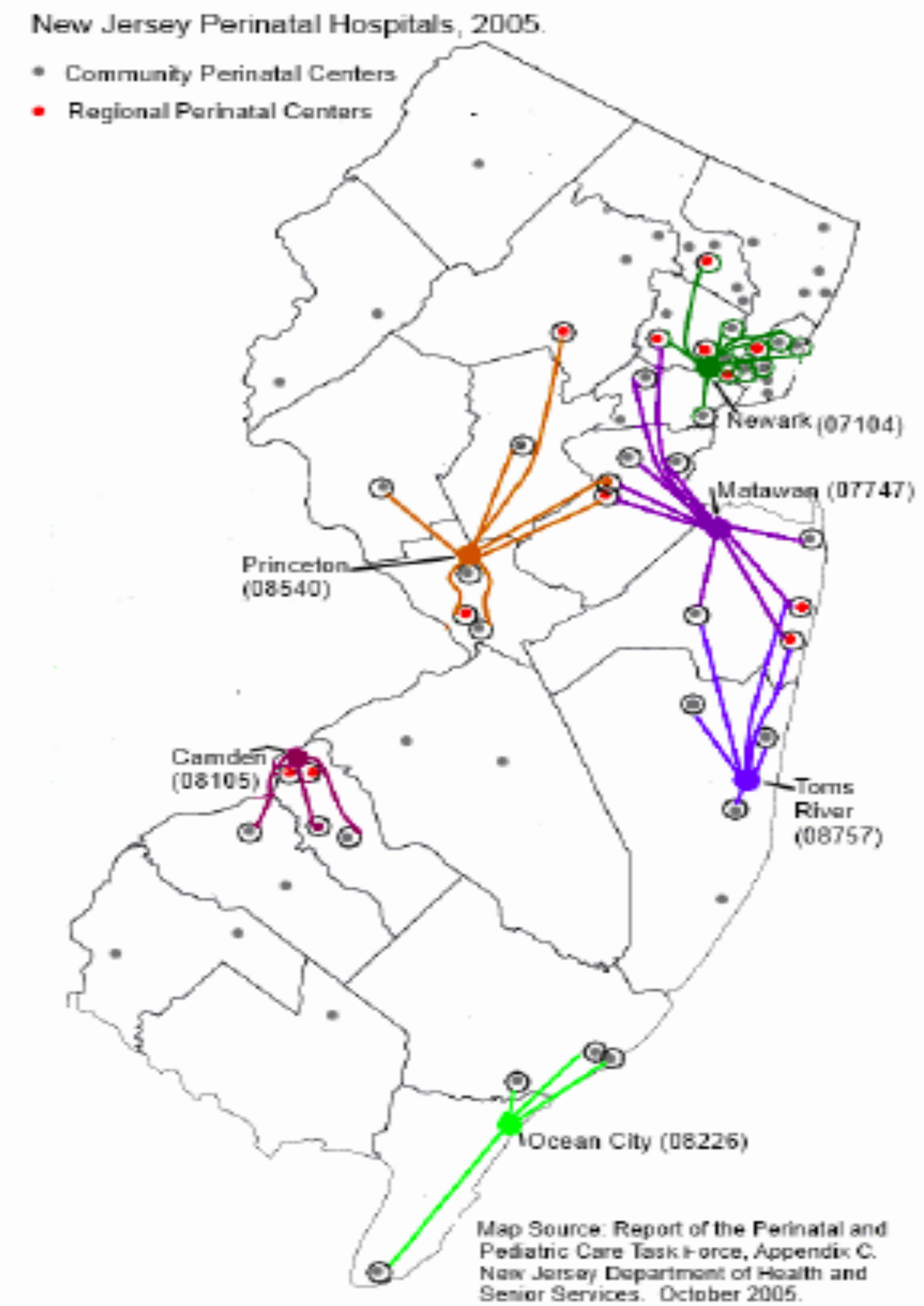
$$Outcome_{ijt}=f(Decision\ making\ skill_j,\ Surgical\ skill_j,\ P^C{}_j\text{-}P^N{}_j,\ X_i,\ hospital\ f.e.,\ day\ f.e.,\ month\ f.e.,\ year\ f.e\ ,\ zip\ f.e.)$$

Where surgical skill is proxied using $s^C{}_j\text{-}s^N{}_j$, j indexes the doctor, i indicates the patient, and X_i includes maternal age, race, insurance, education, marital status, and child gender and birth order.

Dealing with Selection and Measurement Error

- Patients may select doctors, and doctor decision making skill is estimated and measured with error.
- Instrument individual doctor measures with market-level measures.
- Markets defined as all hospitals within 5 miles of a woman's residence plus any other hospital that at least 3 women from her zip code used in the delivery year.
- Market measures are constructed by taking a weighted average of the physician measures in the market, where the number of deliveries is the weight.

An Illustration of NJ Hospital Markets in 2005



First stage Regressions

	Doctor Decision Making			Doctor Surgical Skill		
	All	Low	High	All	Low	High
Market Decision Making	0.353	0.356	0.347	-0.026	-0.024	-0.028
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Market Surgical	-0.014	-0.009	-0.019	0.284	0.290	0.276
	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
R-squared	0.165	0.179	0.152	0.098	0.105	0.090

Effects on C-section risk and Any Bad outcome, by patient risk category

C-section Risk:	Low p(csect)<.084	Medium p(csect)>=.084 p(csect)<=.439	High p(csect)>.439
Dep. Var: C-section			
Decision Making	-0.015	-0.013	0.043
	(0.004)	(0.009)	(0.006)
Procedural Skill	0.014	0.022	0.034
	(0.007)	(0.012)	(0.012)
Dep. Var: Any Bad Outcome			
Decision Making	-0.009	-0.018	-0.010
	(0.007)	(0.008)	(0.003)
Procedural Skill	-0.043	-0.058	-0.078
	(0.006)	(0.008)	(0.005)
# Observations	251948	472955	243845

Effects on Maternal and Infant Outcomes

	Low Risk	Medium Risk	High Risk
Dep. Var: Bad Maternal Outcome			
Decision Making	-0.004	-0.008	0.003
	(0.002)	(0.004)	(0.004)
Procedural Skill	-0.017	-0.033	-0.060
	(0.006)	(0.009)	(0.008)
Dep. Var: Bad Infant Outcome			
Decision Making	-0.006	-0.011	-0.013
	(0.006)	(0.010)	(0.004)
Procedural Skill	-0.029	-0.034	-0.025
	(0.007)	(0.011)	(0.007)
# Observations	251948	472955	243845

Effects on Maternal and Infant Outcomes

	Low Risk	Medium Risk	High Risk
Dep. Var: Bad Maternal Outcome			
Decision Making	-0.004	-0.008	0.003
	(0.002)	(0.004)	(0.004)
Procedural Skill	-0.017	-0.033	-0.060
	(0.006)	(0.009)	(0.008)
Dep. Var: Bad Infant Outcome			
Decision Making	-0.006	-0.011	-0.013
	(0.006)	(0.010)	(0.004)
Procedural Skill	-0.029	-0.034	-0.025
	(0.007)	(0.011)	(0.007)
# Observations	251948	472955	243845

How Large are These Effects?

- A one s.d. increase in decision making (matching) skill reduces C-section among the low risk by ~15% and increases it among the high risk by ~10%.
- Reduces the risk of bad outcomes for low-risk mothers, and high-risk infants by ~12%.
- A one s.d. increase in procedural skill increases C-sections especially for the low risk and reduces the risk of bad outcomes by up to 50%, especially for the high risk.

Application 2: Heart Attacks in Florida (Currie, MacLeod, Van Parys, 2016)

- Examine all patients hospitalized from the ER for Acute Myocardio Infarction (AMI) in FL 1992-2011
- The decision is whether to do an invasive procedure (angioplasty) vs. treat with drugs.
- Measure decision making skill in the same way (i.e. as a “slope” in a regression of procedure choice on an index of the patient’s appropriateness for the invasive procedure.)
- Also look at the “intercept,” i.e. the doctor’s mean propensity to use the invasive procedure.

The Joint Commission

Accountability Measure List

Note: New accountability measures for 2014 have been highlighted

Year Designated	Measure Set	Measure ID#	Measure Name
Heart Attack Care			
2010	AMI-1	14229	Aspirin at Arrival
2010	AMI-2	14230	Aspirin Prescribed at Discharge
2010	AMI-3	14231	ACEI or ARB for LVSD
2010	AMI-5	14232	Beta-Blocker Prescribed at Discharge
2010	AMI-7a	14236	Fibrinolytic Therapy Received Within 30 Minutes of Hospital Arrival
2010	AMI-8a	14235	Primary PCI Received Within 90 Minutes of Hospital Arrival
2011	AMI-10	14237	Statin Prescribed at Discharge

In this checklist, the doctor is “good” if:

- The heart attack patient received an aspirin on arrival at the hospital.
- A Percutaneous Coronary Intervention (PCI) was received within 90 minutes of hospital arrival *given that a PCI was performed*
- Fibrinolytic therapy is performed within 30 minutes *given that it is performed*.

- **But there is no discussion in the guideline of whether PCI or fibrinolytics is the right procedure for the patient**.

Our approach:

- Use doctors in hospitals with cardiology teaching programs to estimate the model of appropriateness for surgery (i.e. to define the “standard of care”) and allow it to change each year.
- The model of appropriateness for surgery considers patient diagnoses, age, and co-morbidities.
- Estimate doctor specific slopes and intercepts for every 3 years of practice.
- We are able to explain about 12% of the variation in procedure choice.

Appropriateness for Surgery:	All	Low	High
# Observations	658,553	217,323	223,853
Female	0.40	0.53	0.27
Age	69.91	80.69	59.65
White	0.79	0.83	0.76
Black	0.08	0.07	0.10
Hispanic	0.10	0.08	0.11
Medicaid	0.04	0.02	0.06
Medicare	0.66	0.88	0.38
Private Insurance	0.21	0.07	0.39
Self Pay or Other	0.09	0.03	0.17
Morbidity Index	0.45	-1.33	2.02
#Diagnoses	8.20	8.98	7.16
Arrhythmia	0.26	0.32	0.20
Hypertension	0.43	0.33	0.56
CHF	0.32	0.51	0.11
CVD	0.07	0.14	0.01
COPD	0.16	0.20	0.09
Cancer	0.06	0.10	0.02
Diabetes	0.21	0.18	0.22
Kidney Disease	0.15	0.28	0.03

Mean patient characteristics by appropriateness for surgery

Procedure Use and Outcomes by Appropriateness for Surgery

Appropriateness for Surgery:	All	Low	High
Any invasive procedure	0.59	0.28	0.86
Length of Stay	6.81	7.68	5.18
Total Hospital Costs	19380	16601	20099
Medical Devices Costs	2702	1819	3466
Cardiology Costs	3617	1754	5453
Operating Room Costs	1025	687	1135
Hospital-Acquired Conditions	0.14	0.22	0.06
Discharged to Home	0.65	0.47	0.81
Died in the Hospital	0.10	0.17	0.04
N	658,553	217,323	223,853

Empirical Model

$Y_{ijt} = \phi_1 * LowResponsiveness_{ijt} + \phi_2 * LowAggressiveness_{ijt} + \phi_3 * HighAggressiveness_{ijt} + \Pi Z_j + \Omega X_i + \delta_h + \lambda_t + \varepsilon_{ijt}$,
where *i* indexes the patient, *j* indexes the physician, and *t* indexes the year-quarter.

LowResponsiveness corresponds to an estimated physician slope that is significantly less than one (a slope significantly greater than one is rare).

LowAggressiveness corresponds to an estimated physician intercept that is significantly less than zero,

HighAggressiveness corresponds to an estimated physician intercept that is significantly greater than zero.

Physician Responsiveness, Aggressiveness and Costs

Appropriateness for Invasive Procedure:	High			Low		
Outcome:	Any Invasive Procedure	Total Costs	Length of Stay	Any Invasive Procedure	Total Costs	Length of Stay
Low Responsiveness	-0.08***	-0.07***	-0.10	0.08***	0.05***	0.16*
(Beta<1)	(0.00)	(0.01)	(0.07)	(0.00)	(0.01)	(0.07)
Low Aggressiveness	-0.09***	-0.11***	-0.30***	-0.11***	-0.08***	-0.01
(Alpha<0)	(0.01)	(0.01)	(0.07)	(0.00)	(0.01)	(0.09)
High Aggressiveness	0.05***	0.09***	0.22**	0.17***	0.13***	0.29***
(Alpha>0)	(0.00)	(0.01)	(0.08)	(0.00)	(0.01)	(0.09)
Hospital*Year FE	Y	Y	Y	Y	Y	Y
Patient Appropriateness Index, Age, Gender	Y	Y	Y	Y	Y	Y
Patient Comorbidities	Y	Y	Y	Y	Y	Y
Physician Characteristics	Y	Y	Y	Y	Y	Y
N	223853	223853	223853	217323	217321	217323
R^2	0.14	0.38	0.11	0.16	0.38	0.11

Physician Responsiveness, Aggressiveness, and Patient Outcomes

Appropriateness for Invasive Procedure:	High			Low		
Outcome:	Hosp. Aquired Infection	Died in Hospital	Not Discharged Home	Hosp. Aquired Infection	Died in Hospital	Not Discharged Home
Low Responsiveness	0.007***	0.009***	0.025***	-0.010***	-0.011***	-0.008***
(Beta<1)	(0.002)	(0.001)	(0.003)	(0.003)	(0.003)	(0.004)
Low Aggressiveness	0.010***	0.009***	0.019***	0.014***	0.013***	0.024***
(Alpha<0)	(0.002)	(0.001)	(0.003)	(0.003)	(0.002)	(0.003)
High Aggressiveness	-0.003*	-0.005***	-0.013***	-0.011***	-0.019***	-0.021***
(Alpha>0)	(0.001)	(0.001)	(0.002)	(0.003)	(0.002)	(0.003)
Hospital*Year FE	Y	Y	Y	Y	Y	Y
Patient Appropriateness Index, Age, Gender	Y	Y	Y	Y	Y	Y
Patient Comorbidities	Y	Y	Y	Y	Y	Y
Doctor Characteristics	Y	Y	Y	Y	Y	Y
N	223853	223853	223853	217323	217323	217323
R^2	0.05	0.06	0.29	0.08	0.08	0.12

How Large are These Effects?

- If the doctor is significantly less responsive than the standard (cardiologists in teaching hospitals), then fewer high risk, and more low risk patients get an invasive procedure.
- As a result, a high appropriateness patient has a ~50% *higher* probability of dying in hospital, while a low appropriateness patient has a ~25% *lower* probability of dying in hospital.
- Costs are higher for the low risk and lower for the high risk, so that an increase in responsiveness is cost neutral.

This example suggests that:

- Not only can we identify physicians who are operate outside the norm in terms of responsiveness or aggressiveness
- But we can also determine whether the norm itself leads to good patient outcomes.
- In this case, "low risk" patients would have benefitted from more aggressive procedures so *de fact*o rationing of procedures to "high risk" patients harmed the low risk.
- Doctors used a simple heuristic (age) to allocate procedures in a way that was not efficient.

Physician Characteristics and Practice Style

Physician Characteristics	Responsiveness	Aggressiveness
Responsiveness (t-1)	0.1655***	0.0387***
	(0.0025)	(0.0018)
Aggressiveness (t-1)	0.0167***	0.4791***
	(0.0020)	(0.0018)
6-9 years experience	0.0165**	0.0438***
	(0.0051)	(0.0045)
	coefficients surpressed	
18-21 years experience	-0.0129*	-0.0290***
	(0.0055)	(0.0048)
	coefficients surpressed	
>30 years experience	-0.1218***	-0.1150***
	(0.0127)	(0.0079)
US Medical School	-0.0192***	0.0081**
	(0.0033)	(0.0027)
Top-20 Medical School	0.0603***	0.0189***
	(0.0041)	(0.0035)
Female Physician	0.0353***	-0.0528***
	(0.0063)	(0.0057)
Spanish-Speaking Physician	-0.0318***	0.0136***
	(0.0040)	(0.0032)
Hospital*Year FE	Y	Y
Patient Appropriateness Index	Y	Y
Patient Characteristics	Y	Y
Patient Comorbidities	Y	Y
N	543192	543192
R^2	0.18	0.43

Physician practice style (for heart attacks) depends on:

- Past practice style
- Cohort
- Training
  - U.S. trained doctors are less responsive and more aggressive.
  - Doctors trained at top 20 places are both more responsive and more aggressive.
- Gender and ethnicity

Our results imply that:

- Even when much procedure use is unnecessary, reducing procedure use across the board may harm high risk patients given imperfect matching of patients and procedures.
- Policy recommendations such as reducing reimbursement for invasive procedures will have this effect.
- Improving the matching of procedures to patients could improve outcomes for both high and low risk patients, even if overall rates of procedure use remain unchanged.

Implications for Policy?

- What makes evaluating physician decision making difficult is that the doctor *always* has access to more information about a patient’s condition than the researcher.
- We cannot second guess a decision in any individual case. But we can identify doctor practice styles and norms that lead to better outcomes. Skilled matching of patients and procedures generates better outcomes.
- *These results raise the question of whether, and under what circumstances, constraints on physician decision making (e.g. guidelines) improve outcomes?*

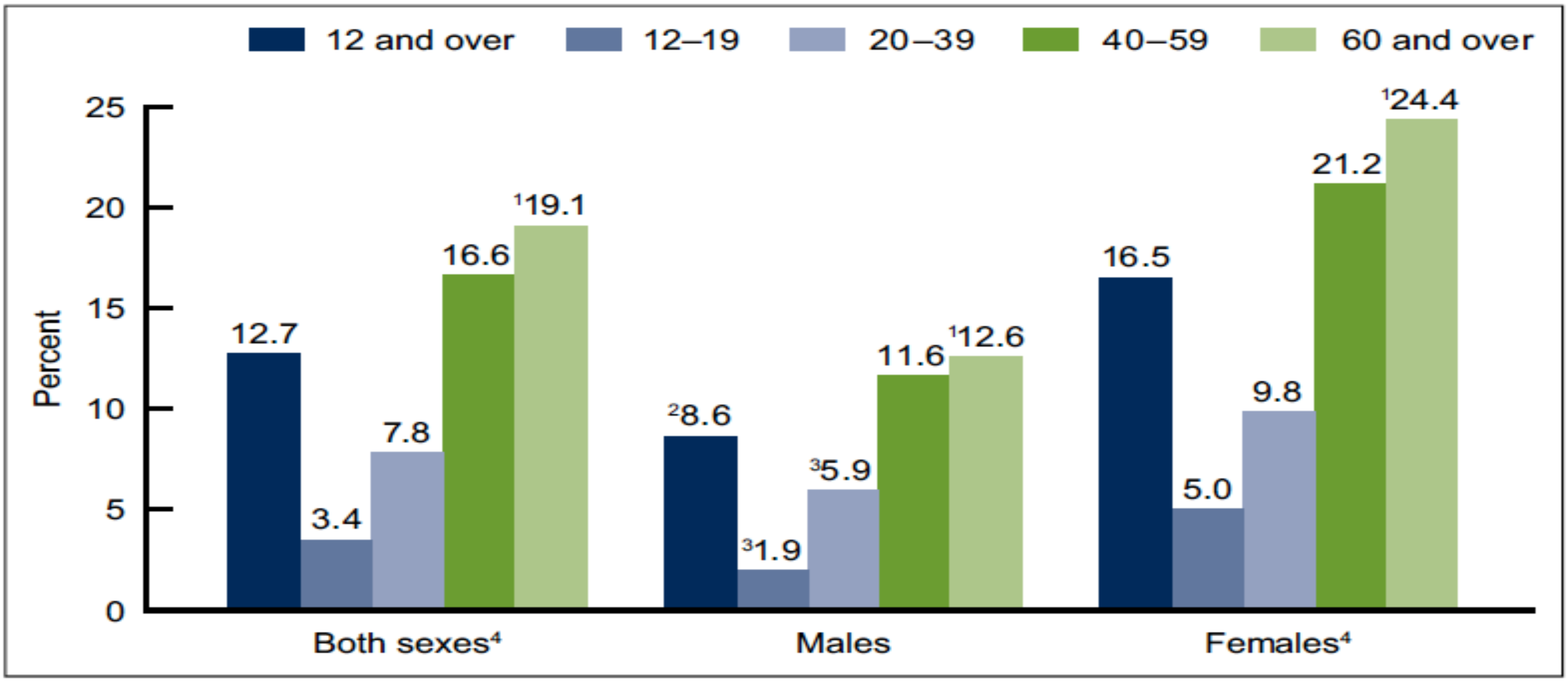
Currie and MacLeod (2020)

- Consider guidelines – they will limit a physician's ability to experiment which may slow learning and prevent the best match between patient and treatment.
- But if followed, they could also prevent worst treatments.
- So there may be a tradeoff between preventing bad outcomes and allowing physician learning.
- In turn, physician learning capability may depend on the physician's human capital, i.e. their ability and previous training.
- So the payoff to experimentation may vary across physicians.

Why Anti-Depressants?

One of the largest and fastest growing classes of drugs.
12.7% of Americans over 12 takes an anti-depressant.

Figure 1. Percentage of persons aged 12 and over who took antidepressant medication in the past month, by age and sex: United States, 2011–2014



Muted financial incentives to favor one drug over another

- Doctors don’t sell psychiatric drugs.
- Most anti-depressants available as generics.
- Variation in costs is largely absorbed by insurers (or government), not by either doctors or patients (the decision makers).
- Data availability:
  - All prescriptions for anti-depressants filled at retail pharmacies between 2006-2014.
  - Includes a doctor identifier, as well as patient age category, gender, and postal code of residence.
  - Merge to American Medical Association provider information to get provider’s specialty, initial medical school and practice address.

Two measures of practice style: Concentration, and adherence to guidelines

- Concentration measured using Shannon's entropy score (following Theil, 1967) (alternatively, could use a Herfindahl index)
- $\Phi = -\Sigma_{k \in D}\, p_{kt} log(p_{kt})/log(n_k)$

  $= \Sigma_{k \in D}\, p_{kt} log(1/p_{kt})/log(n_k),$

  $= \phi\,(p_{kt})/\log(n_k)$

- where k indexes drugs, n_k is the number of drugs that are ever available over the sample period, and p_{kt} is the share of patients who are taking drug k at time t.

Exploring the Interaction of Experimentation and Skill

- We assume that psychiatrists have more skill in drug treatment for mental health than General Practitioners and other family doctors.
- Since most patients in our data are not treated for very long, we look at the first drug chosen.
- We associate higher entropy in these drugs with a doctor’s greater propensity to experiment.

We also examine violation of guidelines for drug transitions

- UK National Institute for Health Care Excellence: Start with an SSRI. If an anti-depressant doesn't work, try another class of drugs.
- Canada: Even within classes, some drugs are more effective than others. If one drug doesn't work, try a more effective one.
- US American Psychiatric Association. Most patients can be treated using: SSRIs, SNRIs, mirtazapine, and bupropion. List excludes two drugs that make up 17.4% of market share in 2014.
- Drug "cocktails" are not generally recommended, and guidelines express concerns about "poly-pharmacy."

Patient-level claims data from a large public health insurer

- 10% random sample of all 99 million members aged 18-64 as of Jan. 2013 who had any claims from Jan. 2013-Dec. 2017.
- Select all members ever prescribed antidepressants over the sample period (n=723,818).
- For each member, generate a panel with a record for each month and year, including whether they are taking any anti-depressant drug, what drug, who prescribed it, and health care costs from inpatient, outpatient, and pharmacy claims. We also know the member's age, gender, and county.

Merging Claims and IQVIA

- Merge in physician entropy for each year calculated from the IQVIA data base using physician exact names and states; we can match ~74.0 percent of the doctors in the BCBS sample.
- Entropy score is annual for each physician. But within patient, we use the average physician entropy score across all months the patient saw that physician.
- Hence, in our data, entropy scores change when the patient changes physician.
- We look at effects of changes at t-1 at outcomes at time t.

Following guidelines would lower entropy (data for 2013)

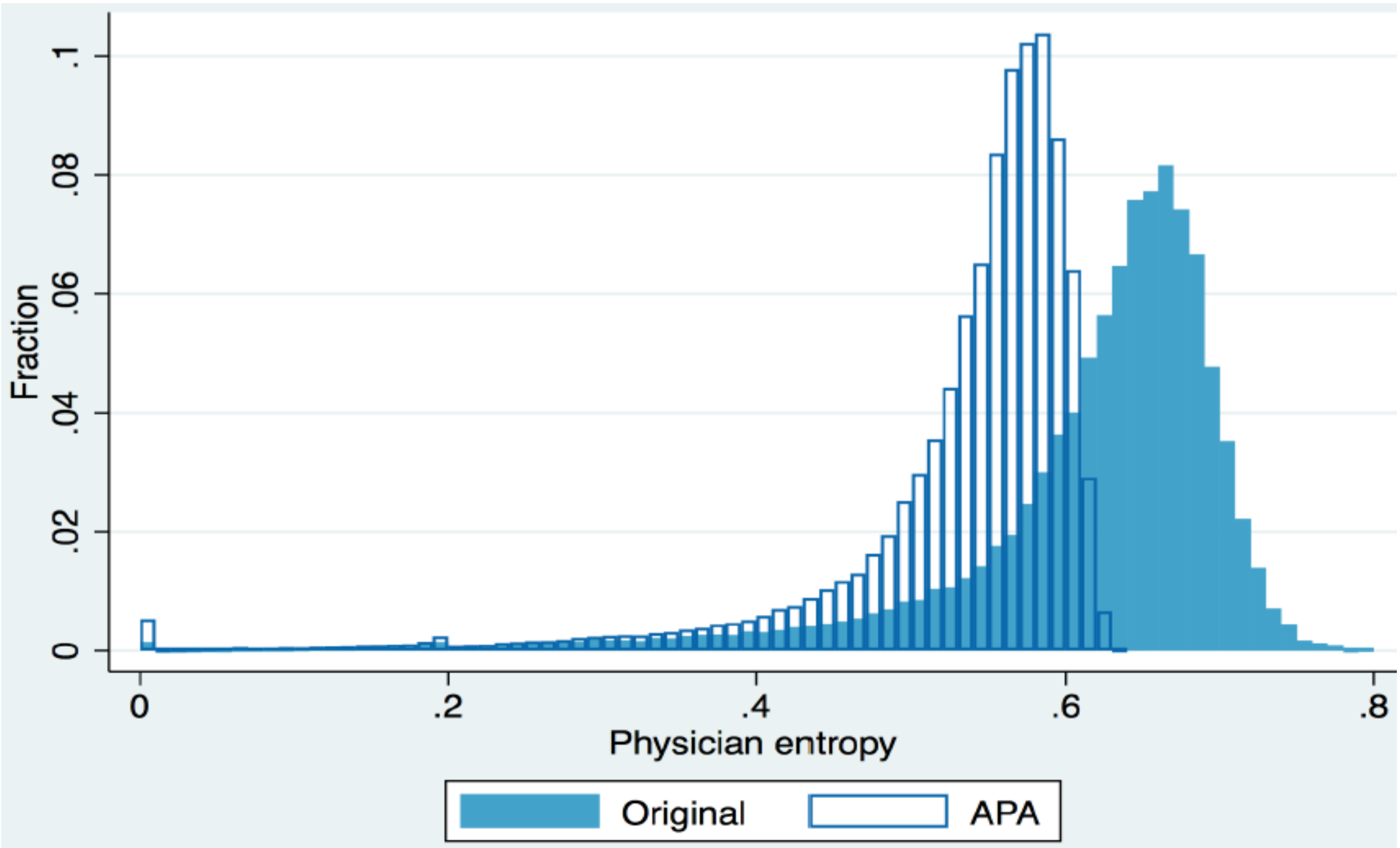


Table 3: Summary of BCBS Patient Data by Patient Outpatient Provider

Patient Type:	All Patients	Ever Saw Psychistrist	Never Saw Psychiatrist
# members	450,802	82,810	367,992
# member-months	5,409,124	1,117,032	4,292,092
# months/member	11.999	13.489	11.664
# months antidepressants/member	8.303	9.499	8.034
# changes in entropy/member	1.331	1.621	1.265
# member-month with nonmissing drug transitions	4716167	976768	3739399
Drug transitions from t-2 to t-1 that violate each guideline		**%**	
UK	0.102	0.117	0.098
Canada	2.406	2.177	2.466
US	3.601	4.623	3.334
Cocktail	4.491	8.917	3.335

Modelling the Relationship Between Entropy and Outcomes

$$Y_{ijt} = a_0 + b_1\phi_{jt-1} + b_2 x_i + b_3 county_i + b_4 y_t + e_{ijt}, \tag{15}$$

or alternatively:

$$Y_{ijt} = a_i + b_1\phi_{jt-1} + b_2 y_t + e_{ijt}, \tag{16}$$

where Y is one of the outcomes discussed above, x are the observable patient characteristics (age category and gender), *county* indicates county fixed effects, and y indicates year fixed effects. T

Patient Outcomes at t on Provider Entropy at t-1

Outcome:	ln(total cost)	ln(non-drug costs)	ER or Hospital	ER/Hosp for Mental Health
Patient FE	yes	yes	yes	yes
Panel B: Patients who ever saw a psychiatrist as an outpatient				
Entropy (t-1)	0.044	-0.013	-0.013	-0.010
	(0.073)	(0.073)	(0.004)	(0.003)
Constant	5.500	3.661	0.046	0.024
	(0.049)	(0.049)	(0.003)	(0.002)
Mean Dep. Variable	4.557	3.024	0.032	0.017
Adj. R2	0.424	0.353	0.110	0.098
# Obs. (millions)	1.117	1.117	1.117	1.117
# Members	82,801	82,801	82,801	82,801
Panel C: Patients who never saw a psychiatrist as an outpatient				
Entropy (t-1)	-0.180	-0.222	0.000	0.003
	(0.045)	(0.045)	(0.002)	(0.001)
Constant	4.907	3.036	0.024	0.006
	(0.028)	(0.029)	(0.001)	(0.001)
Mean Dep. Variable	3.825	2.378	0.021	0.008
Adj. R2	0.374	0.282	0.086	0.060
# Obs. (millions)	4.292	4.292	4.292	4.292
# Members	367,992	367,992	367,992	367,992

Patient Outcomes at t when prescribing between t-2 and t-1 Violated Guidelines

Outcome:	ln(total cost)	ln(non-drug costs)	ER or Hospital	ER/Hosp for Mental Health
Patient FE	yes	yes	yes	yes
Panel B: Patients who ever saw a psychiatrist as an outpatient				
Violation UK Guidelines	0.249	0.315	-0.001	-0.001
	(0.066)	(0.078)	(0.006)	(0.005)
Violation US Guidelines	0.241	0.271	0.005	0.003
	(0.013)	(0.015)	(0.001)	(0.001)
Violation Can. Guidelines	0.513	0.416	0.002	0.002
	(0.017)	(0.020)	(0.001)	(0.001)
Cocktail	0.581	0.444	0.005	0.002
	(0.014)	(0.015)	(0.001)	(0.001)
Mean Dep. Variable	4.573	3.017	0.032	0.017
Adj. R2	0.391	0.362	0.111	0.101
# Obs.	976768	976768	976768	976768
# Members	82,810	82,810	82,810	82,810
Panel C: Patients who never saw a psychiatrist as an outpatient				
Violation UK Guidelines	0.200	0.183	0.004	0.000
	(0.039)	(0.047)	(0.003)	(0.002)
Violation US Guidelines	0.298	0.299	0.003	0.001
	(0.008)	(0.009)	(0.001)	(0.0003)
Violation Can. Guidelines	0.483	0.418	0.005	0.002
	(0.009)	(0.010)	(0.001)	(0.0004)
Cocktail	0.468	0.307	0.004	0.002
	(0.012)	(0.013)	(0.001)	(0.0004)
Mean Dep. Variable	3.863	2.383	0.021	0.008
Adj. R2	0.380	0.286	0.086	0.060
# Obs. (millions)	3.739	3.739	3.739	3.739
# Members	367,992	367,992	367,992	367,992

Summary

- There are persistent differences in practice style within locations which matter for patient outcomes.
- We highlight experimentation with drug treatment as one aspect of practice style that can be captured by the entropy measure.
- Our model indicates that experimentation will be more valuable when physicians have greater diagnostic skill, and we find evidence that this is the case.
- At the same time, loose guidelines are shown to be useful in restricting practice style among all physicians, in the sense that violations of guidelines lead to worse outcomes.

Conclusions

- Physician behavior is complex and there are many possible reasons that medically similar patients receive different care.
- Traditional economic models treat physicians as profit-maximizing firms. They emphasize patient demand as well as constraints on physician behavior, such as time constraints.
- More recently, other factors are being considered. These include:
  - Peer effects
  - Discrimination/communications/homophily
  - Training
  - Skill/Cognitive limitations
- Guidelines may improve outcomes by limiting really bad decision making but may also come at the cost of reducing experimentation and learning.
- Physician decision making is a rich area for future research. And growing availability of data will facilitate that.