Using Big Data To Solve Economic and Social Problems

Professor Raj Chetty

Head Section Leader Rebecca Toseland

Photo Credit: Florida Atlantic University

Forecasting Flu Outbreaks Using Google Search Data

- Data to be predicted: 1,152 observations from CDC on flu incidence
  - Weekly data from 9 regions of the U.S. from 2004-2007
- Data used for prediction: counts of Google search data
  - Weekly data on Google search counts for 50 million terms by state from 2004-2007

Google Flu Trends: Overfitting Problem

- This is an example of “wide data”
  - Many more variables than number of observations
  - Overfitting problem: can fit the data perfectly using 1,152 explanatory variables → cannot use traditional statistical methods like regression
- Solve this problem using *out-of-sample validation*
  - Idea: use separate samples to estimate the model and evaluate its predictive accuracy

Google Flu Trends: Methodology

- Construct predictive model in a series of steps:

1. Take each of the 50 million search queries Q *separately* and run a regression of CDC data on that term:

$$I(t) = \beta Q(t) + \varepsilon(t)$$

   - Calculate correlation between predictions from this model and true CDC data across 9 regions
   - Rank the 50 million terms based on this correlation and choose top 100
   - Includes terms like “cough” and “antibiotics” but also terms like “high school basketball” and “oscar nominations”

Google Flu Trends: Methodology

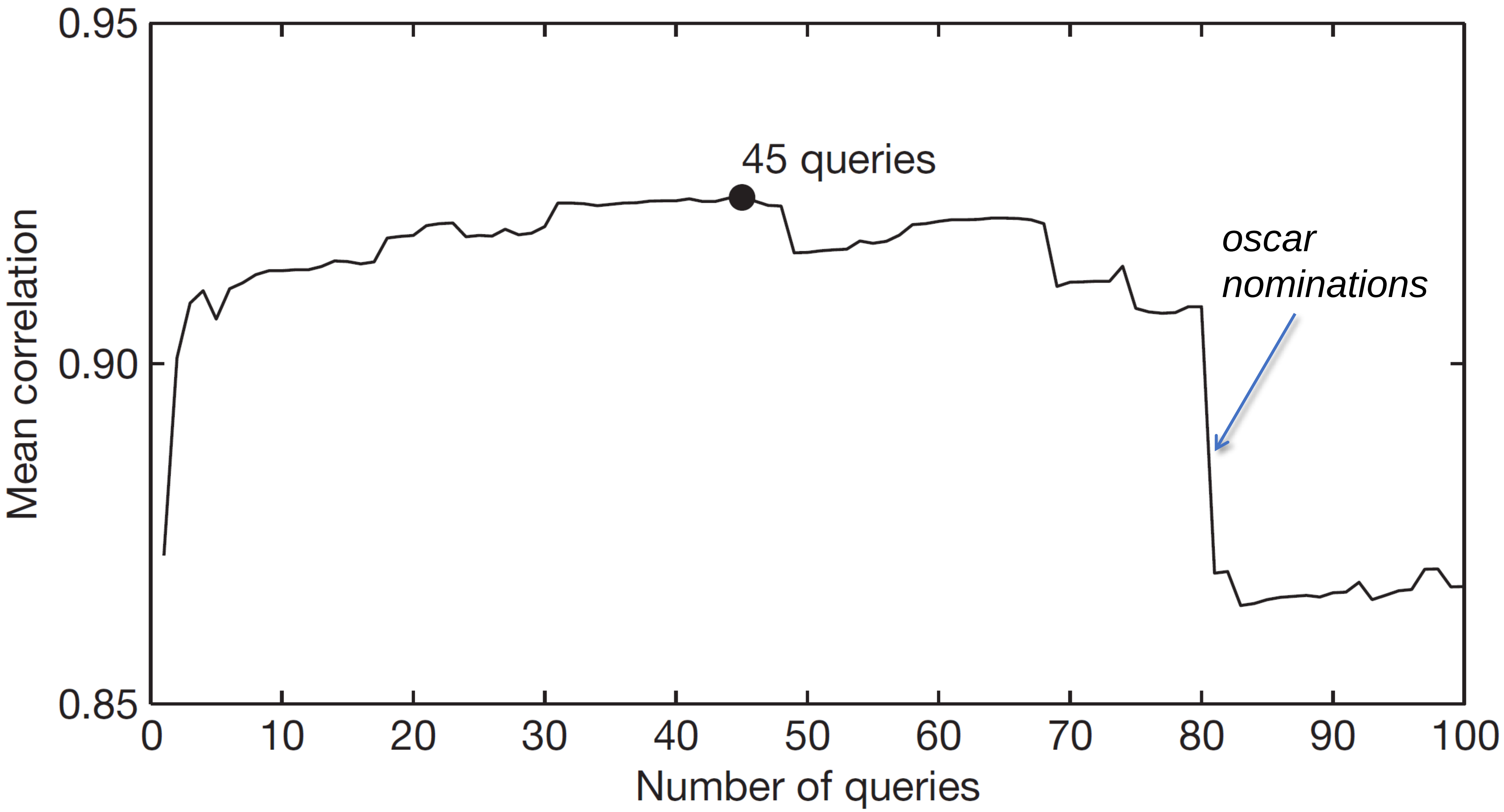
- Construct predictive model in a series of steps:

2. Using a *separate* set of data from later weeks to decide which of the top 100 terms to include in prediction model

   – Construct sum of search queries across top *n* terms

   – Evaluate how well this sum predicts regional and weekly variation in new sample, varying *n* from 1 to 100

Out of Sample Validation to Choose Optimal Number of Search Queries

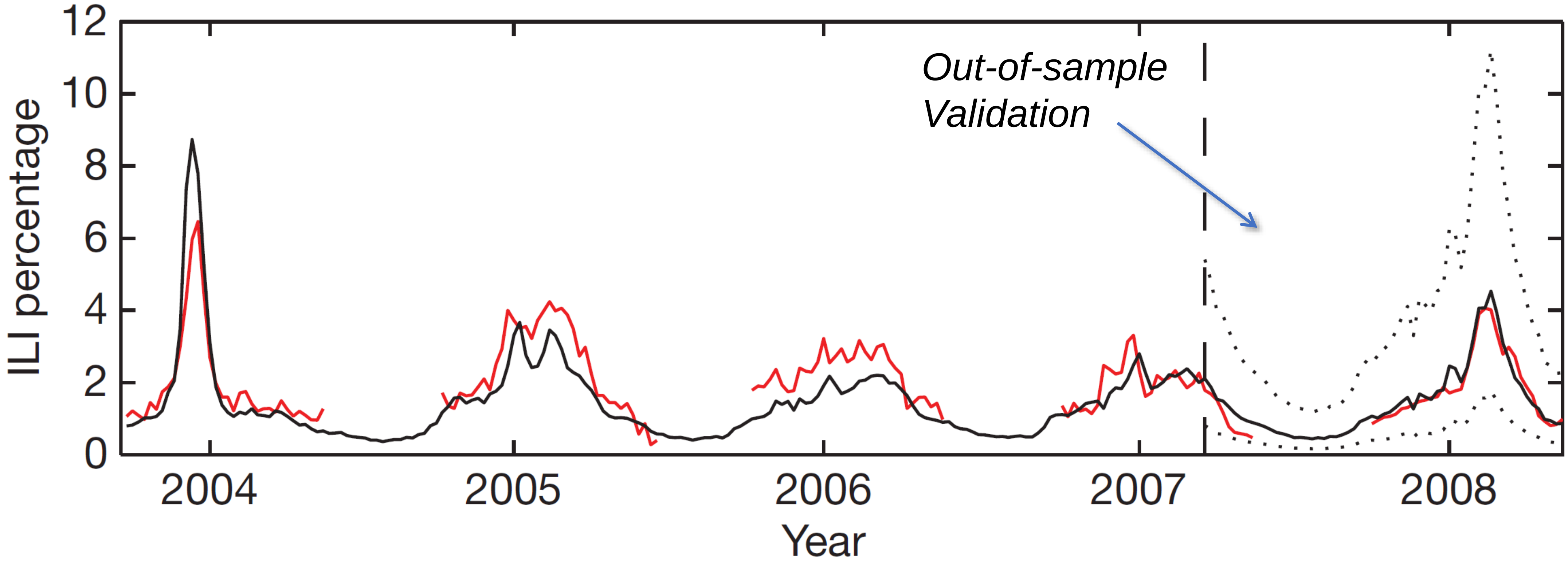


Google Flu Trends: Methodology

- Construct predictive model in a series of steps:

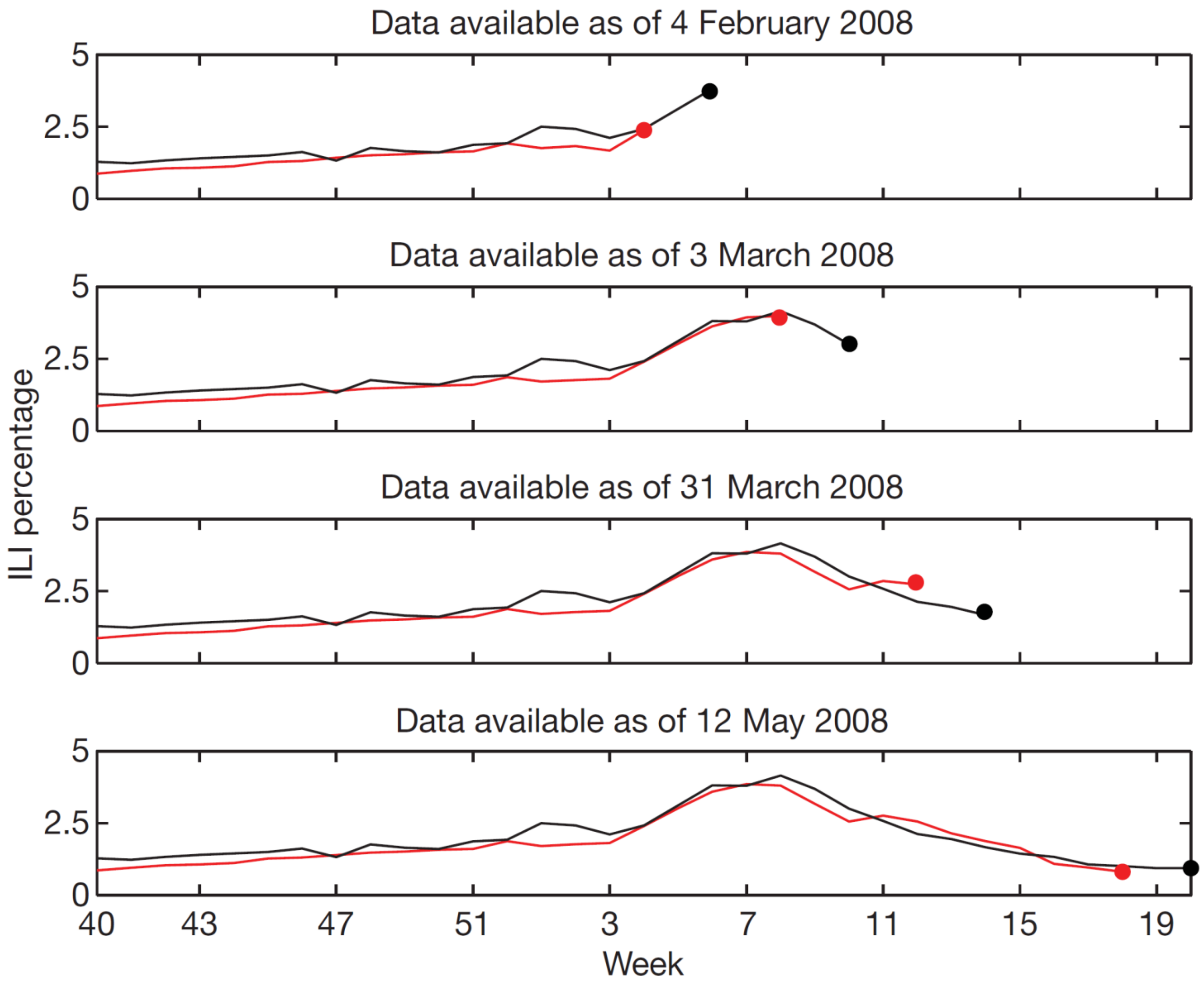
3. Finally, evaluate model fit and out of sample predictive accuracy using subsequent data that was not available when model was estimated

In-Sample and Out-of-Sample Fit of Prediction Model



Note: CDC official statistics in red; Google trends forecast in black

Out-of-Sample Model Validation Using Two-Week Lead Time



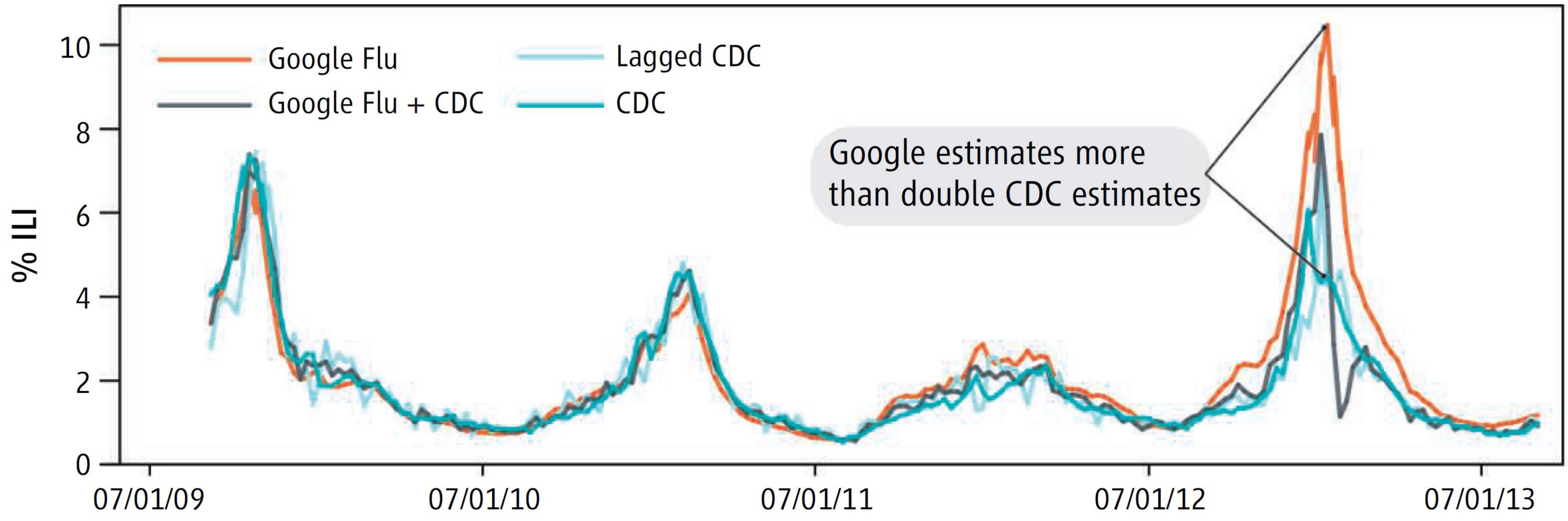
Note: CDC official statistics in red; Google trends forecast in black

Breakdown of Google Flu Trends Predictive Model

- Problem: predictive model began to break down in late 2012 and became very inaccurate in forecasting outbreaks of flu

- Lazer et al. (2014) document model's failure essentially by extending window used for out of sample to 2013

Out-of-Sample Fit of Prediction Model



Breakdown of Google Flu Trends Predictive Model

- Problem: predictive model started to break down over time and became very inaccurate

- Lazer et al. (2014) document this breakdown essentially by extending window used for out of sample to 2013

- Why did the model start to perform poorly?

  - Google search engine started to prompt users to search for additional diagnoses after entering a term like fever or cough

  - Autofill started to offer suggestions for search terms

  - Both of these factors changed nature of search queries; since model was not re-estimated, predictions changed

Broader Lessons from Google Flu Predictive Model

1. Big data has great potential for predictive modeling with applications to social problems

   – Ginsberg et al. (2009) became the basis for Google Correlate, a public tool to find searches that correlate with real-world data

Broader Lessons from Google Flu Predictive Model

1. Big data has great potential for predictive modeling with applications to social problems

2. But big data is not a substitute for ground truth

   – Good thing that CDC did not abandon its program to collect data on flu incidence from clinics after Ginsberg et al. (2009) was published

Broader Lessons from Google Flu Predictive Model

1. Big data has great potential for predictive modeling with applications to social problems

2. But big data is not a substitute for ground truth

3. Building good models requires both technical skill and careful judgement

   – Fitting black-box models is tempting, but models where mechanisms are sensible are more likely to yield stable predictions

   – When terms like “oscar nominations” show up, should be very cautious

   – Frontier of research in machine learning: developing tools to improve predictive accuracy in such settings

The Economics of Health Care

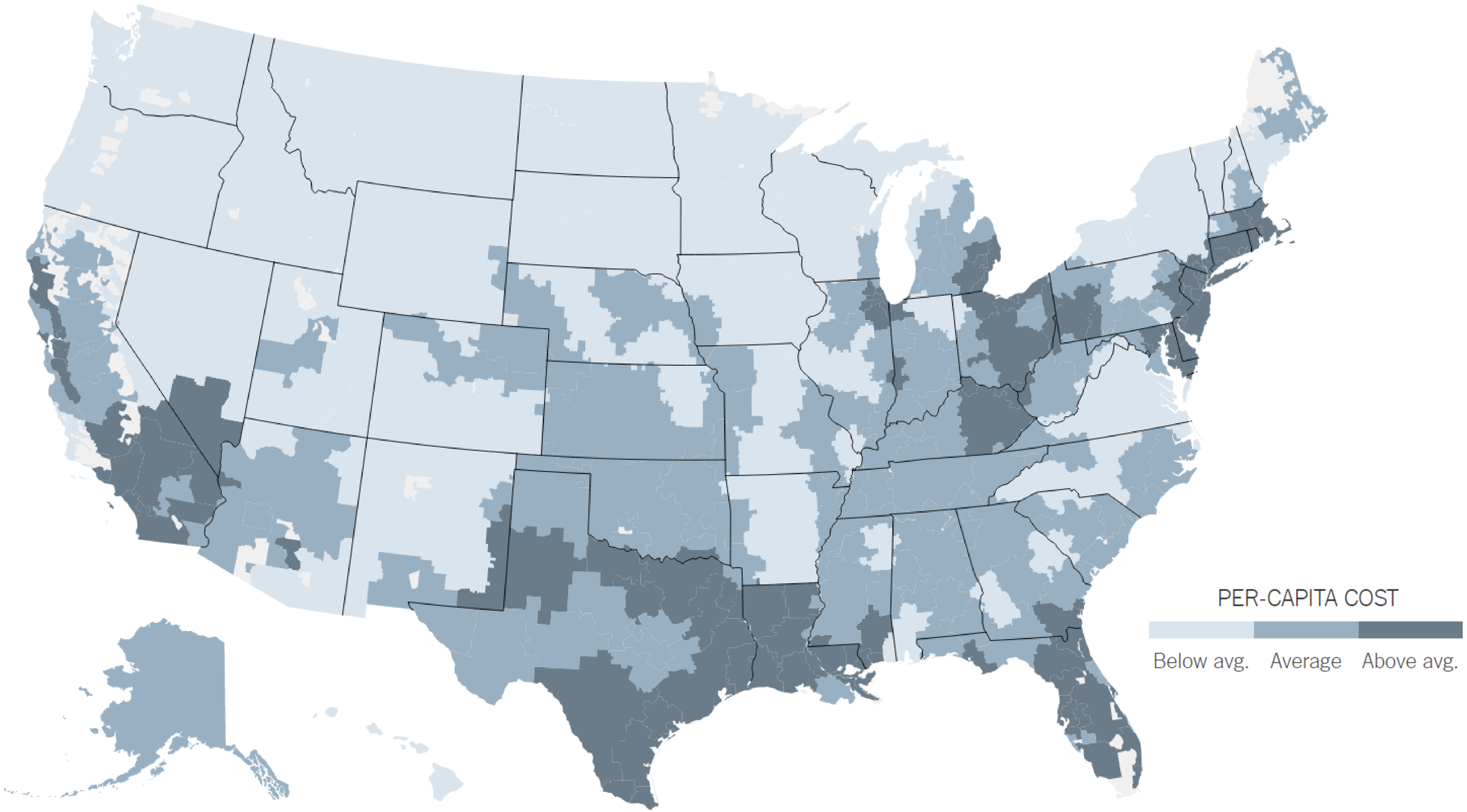
The Economics of Health Care

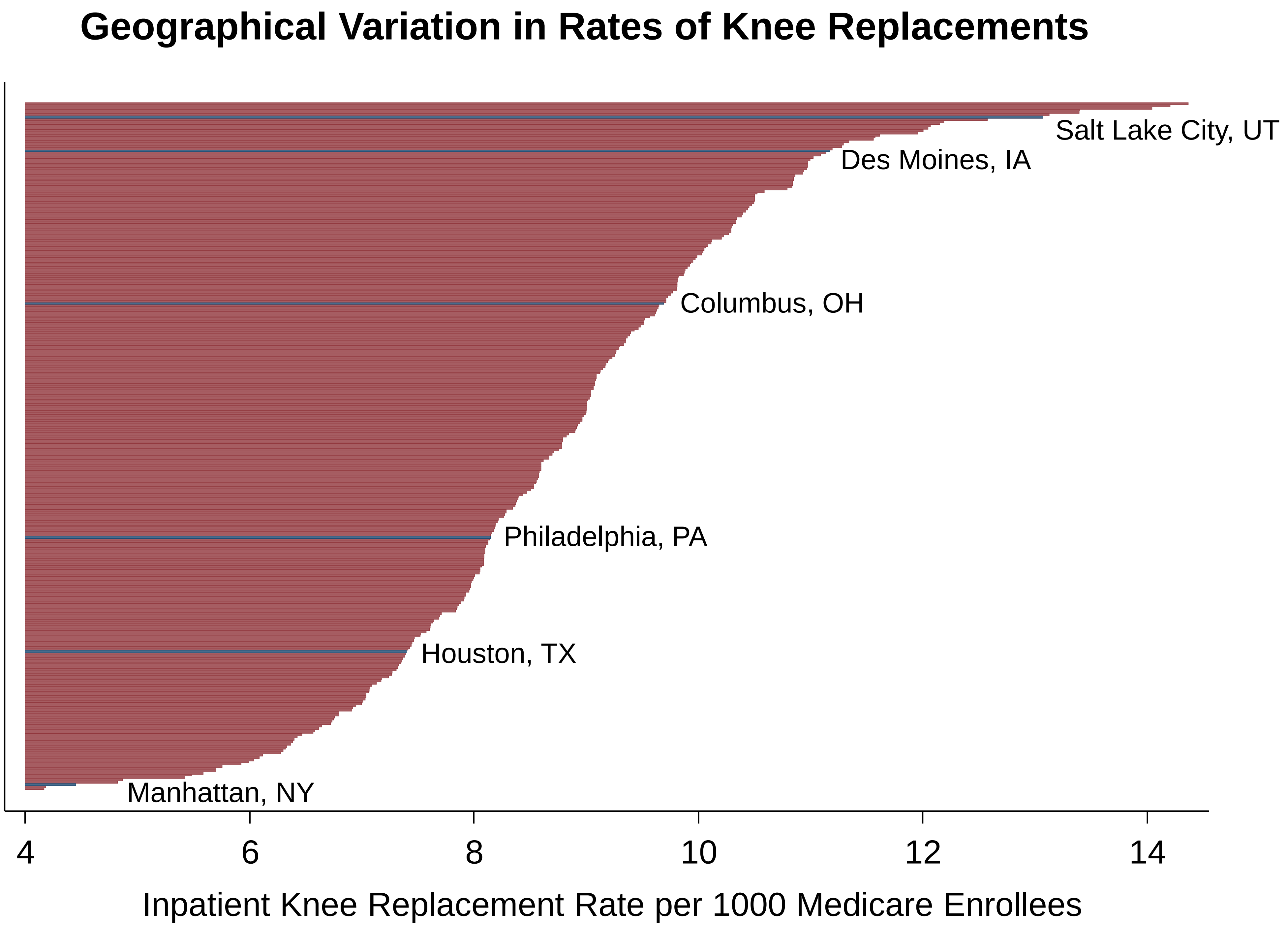
- Health economists focus on studying markets for health care
  - Why is health care so expensive in the United States?
  - Will expanding health insurance coverage improve health outcomes or just lead to more wasteful spending?
  - How can we provide health insurance to more Americans?

Dartmouth Atlas: Geographic Variation in Health Spending

- Dartmouth Atlas uses data from Medicare claims to calculate expenditures per adult in local areas
  - Adjust for differences in population demographics (race, sex, age)
- Substantial spatial variation in health care expenditures that is driven by variation in quantity of care
  - Medicare expenditures vary from $8,300 to $10,400 per person between 20th and 80th percentile across areas in the U.S.

Medicare spending per capita




Geographical Variation in Rates of Knee Replacements
Salt Lake City, UT
Des Moines, IA
Columbus, OH
Philadelphia, PA
Houston, TX
Manhattan, NY
4
6
8
10
12
14
Inpatient Knee Replacement Rate per 1000 Medicare Enrollees

Dartmouth Atlas: Geographic Variation in Health Spending

- Expenditures not correlated with health outcomes
  - Led to concern about "flat of the curve" medicine, particularly after a widely-read article by Atul Gawande in 2009
  - Physicians and hospitals compensated by government for non-essential procedures (e.g., MRIs) → concern about wasteful spending
  - Motivated efforts to reduce expenditures in areas such as McAllen, TX
  - But implications heavily debated: is there really wasteful spending or is it just that patient populations differ across places (selection effects)?

Geographic Variation: Private Health Insurers

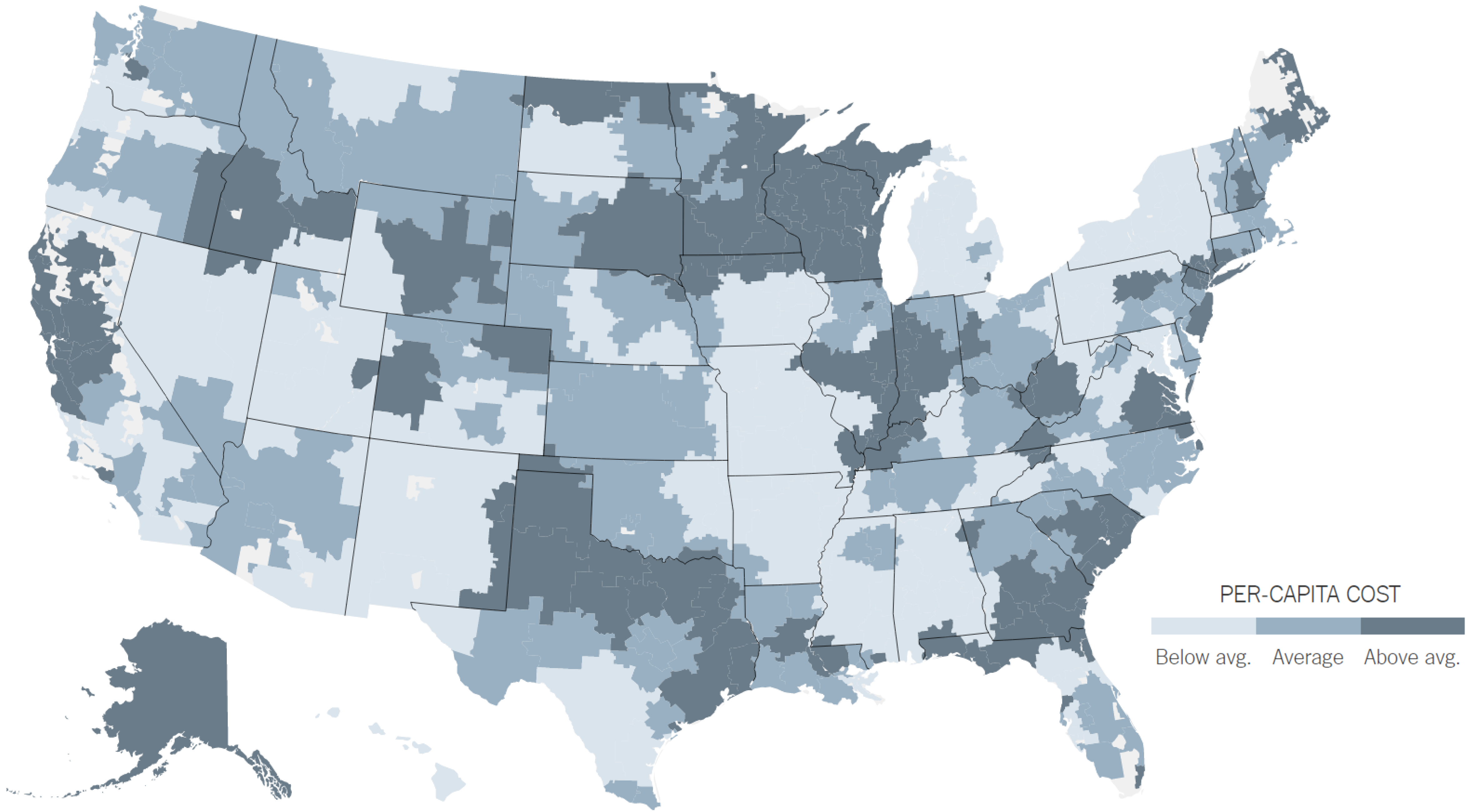
- Dartmouth Atlas only had data from Medicare, not from private insurance companies (below age 65)

- Cooper et al. (2015) show that there is substantial variation in private insurer expenditures as well

  - Expenditures vary from $3,000 to $3,900 between 20th and 80th percentile across areas

- But geographic pattern is very different for private health insurers

Private insurance spending per capita

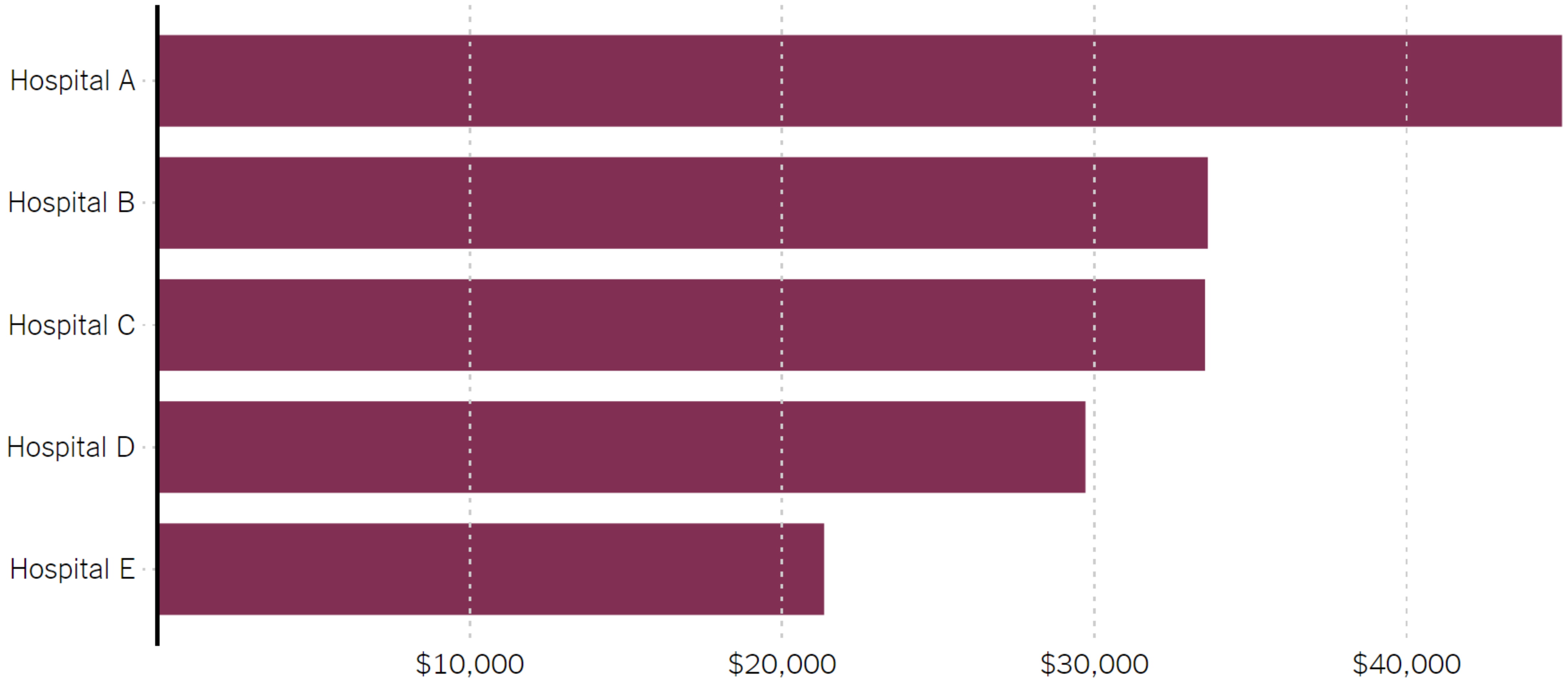


Geographic Variation: Private Health Insurers

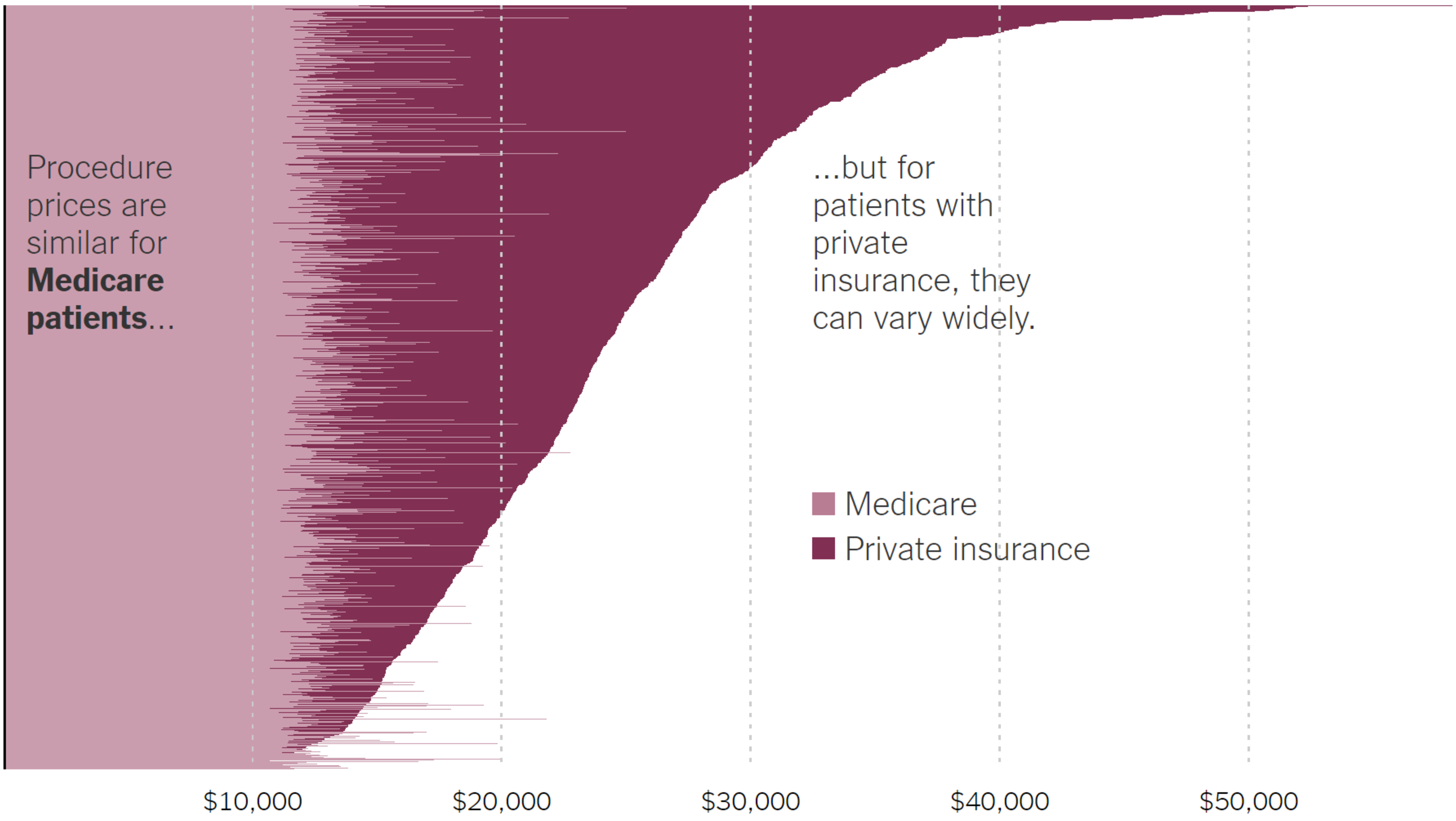
- Dartmouth Atlas only had data from Medicare, not from private insurance companies (below age 65)

- Cooper et al. (2015) show that a very different picture emerges for private health insurers
  - Correlation between private health insurance expenditures and Medicare expenditures is only 0.14 across areas
  - And most of the variation is due to *prices*, not quantities…

How much a knee replacement can cost in New York City



Price of Simple Knee Replacement Surgery In 937 Hospitals



Lessons from Geographic Variation on Efficiency of Markets for Health Care

- Health care markets function very differently from markets for other goods such as cars or cell phones

- Wide variation in prices and quantities for what appear to be similar services suggests that there may be considerable inefficiency

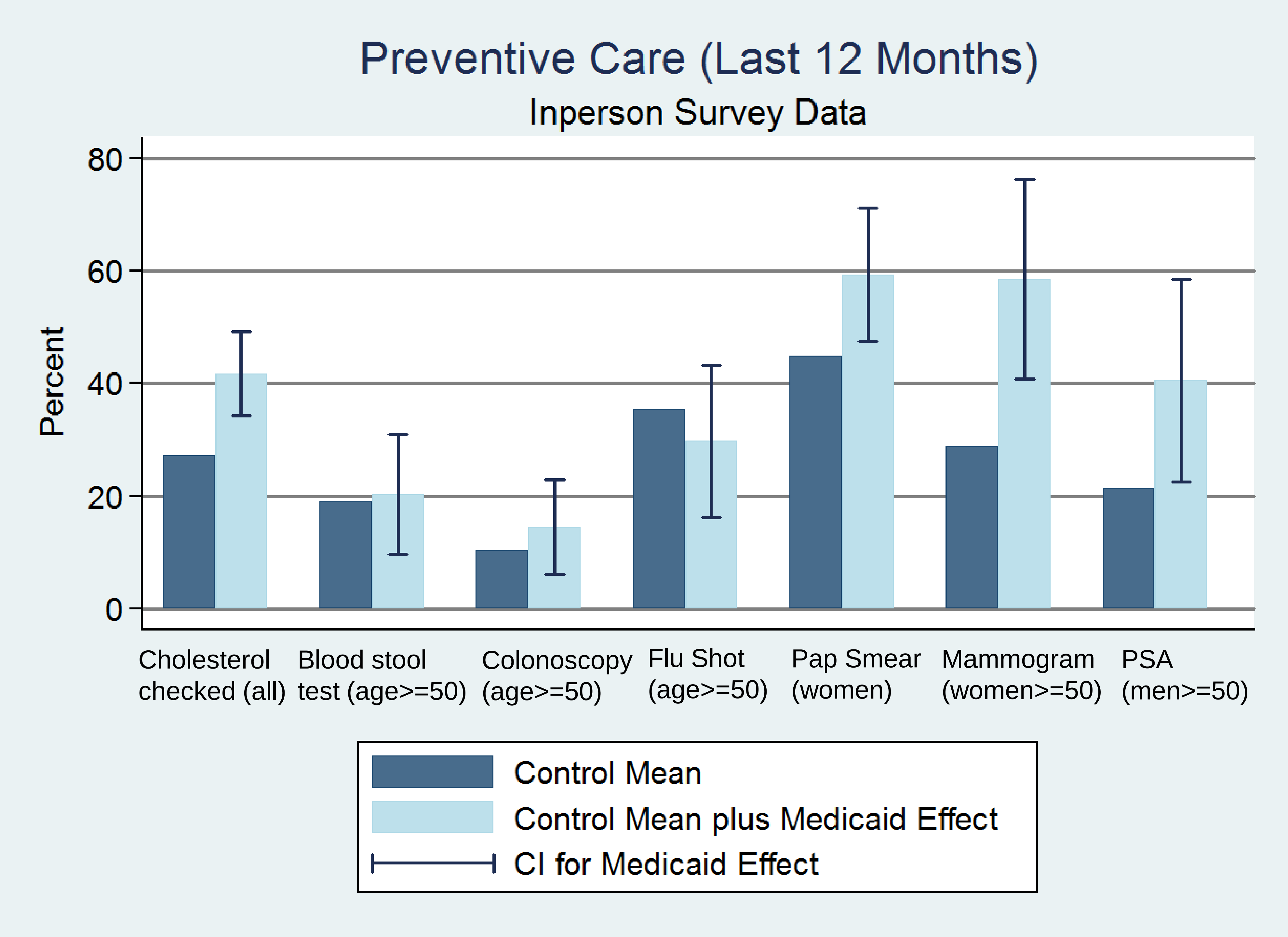
- Many factors at play, but one important and unique feature: third-party (insurance company or Medicare) payment
  - Customer is not paying the price → may be little incentive to find the cheapest price and little incentive to cut back on quantity

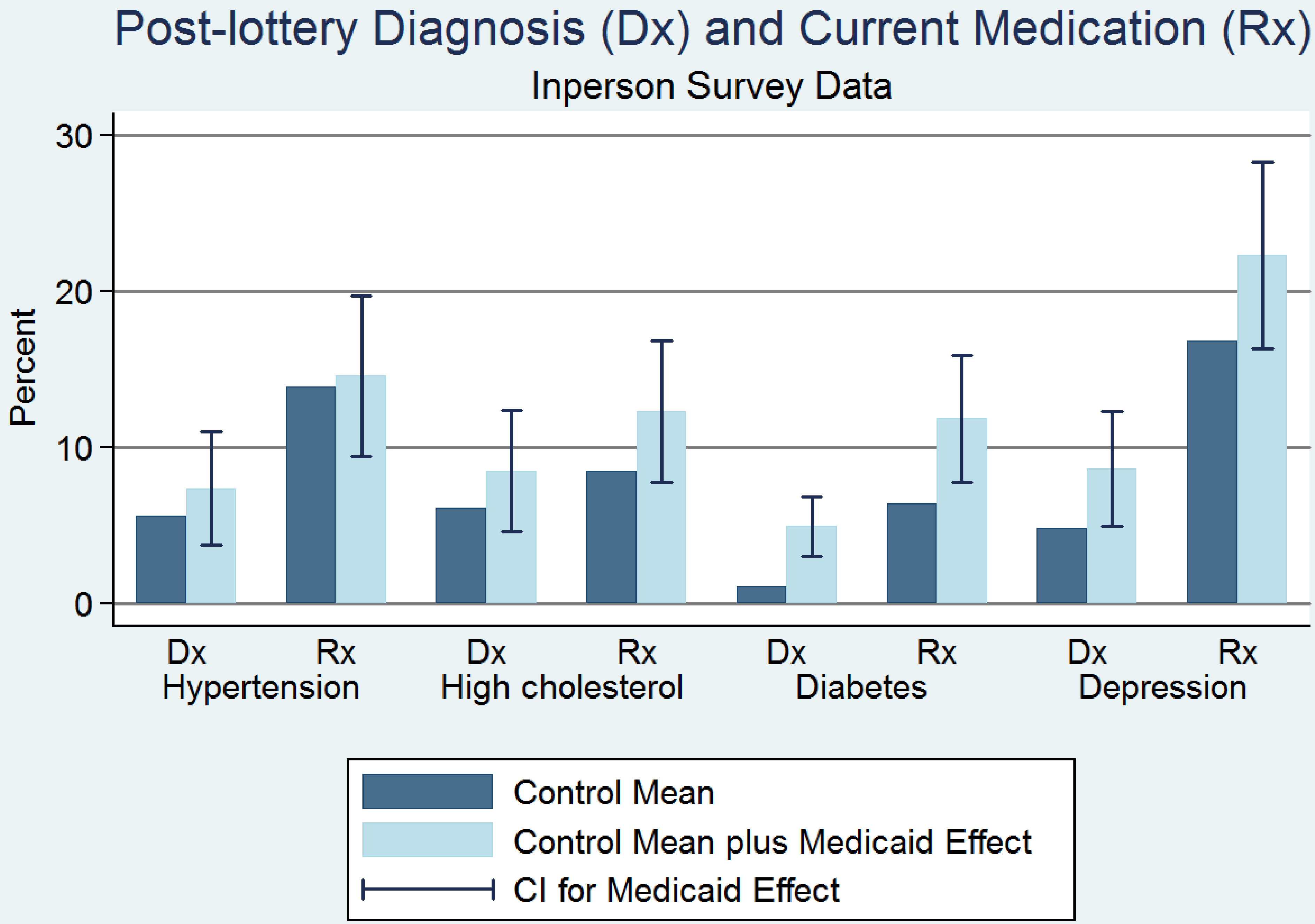
Insurance and Demand for Health Care

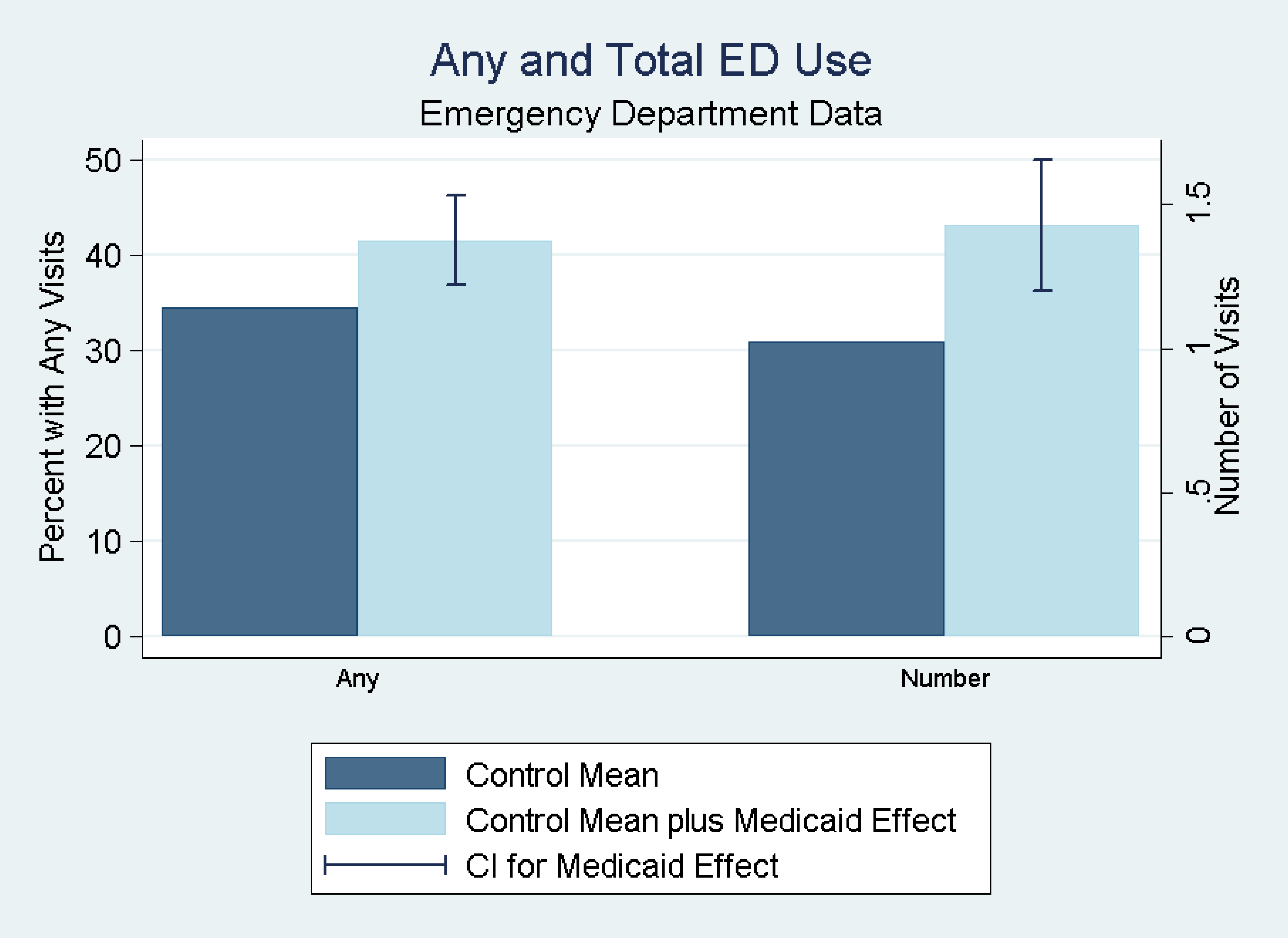
- What is the causal effect of insurance on demand for health care and health outcomes?
  - Does providing individuals' insurance actually encourage wasteful spending or does it improve health outcomes?
- Ideal experiment: randomly assign health insurance to some individuals and not others and compare outcomes
- This turns out to be a rare case where we actually have such an experiment

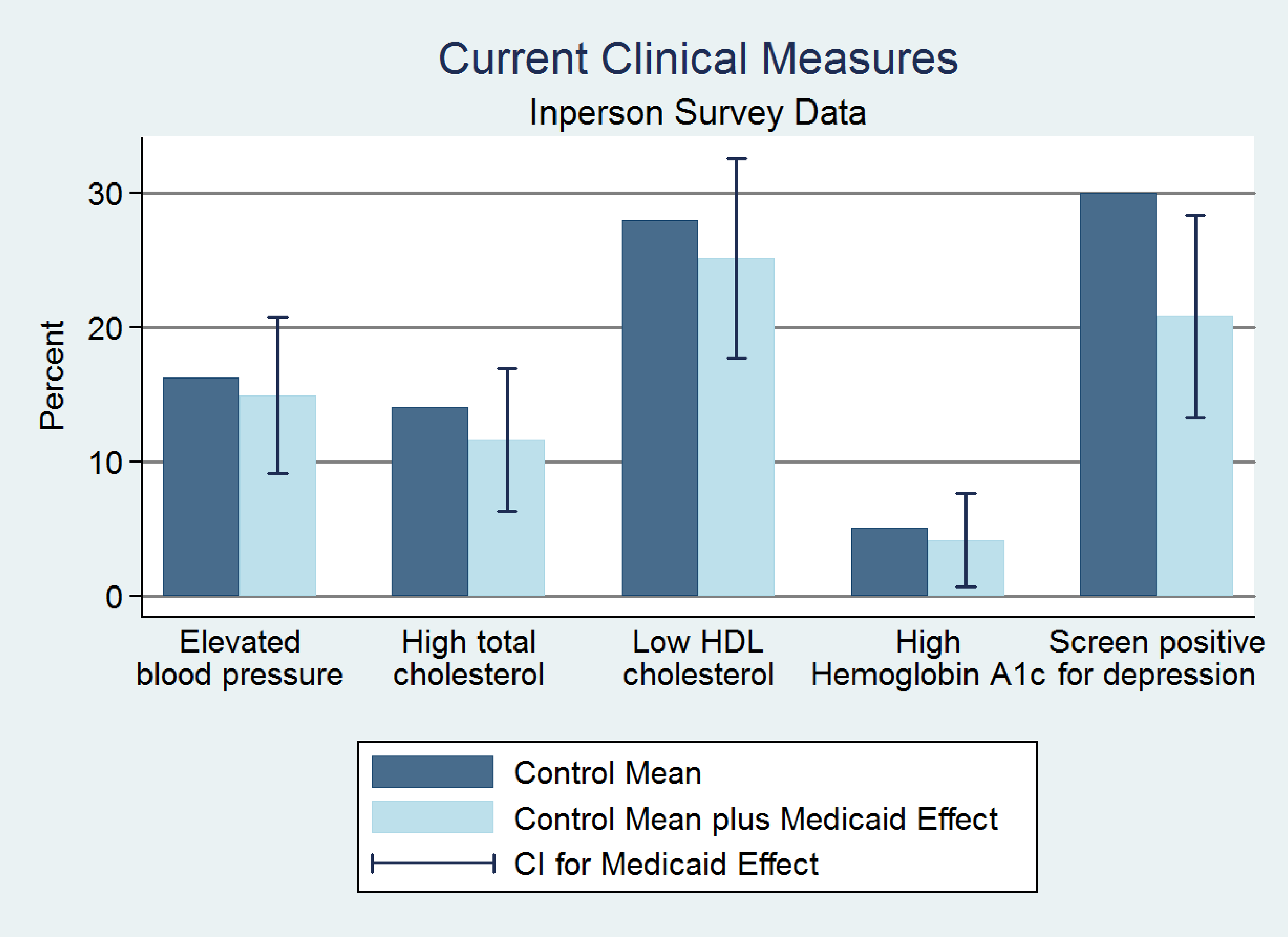
Oregon Health Insurance Experiment

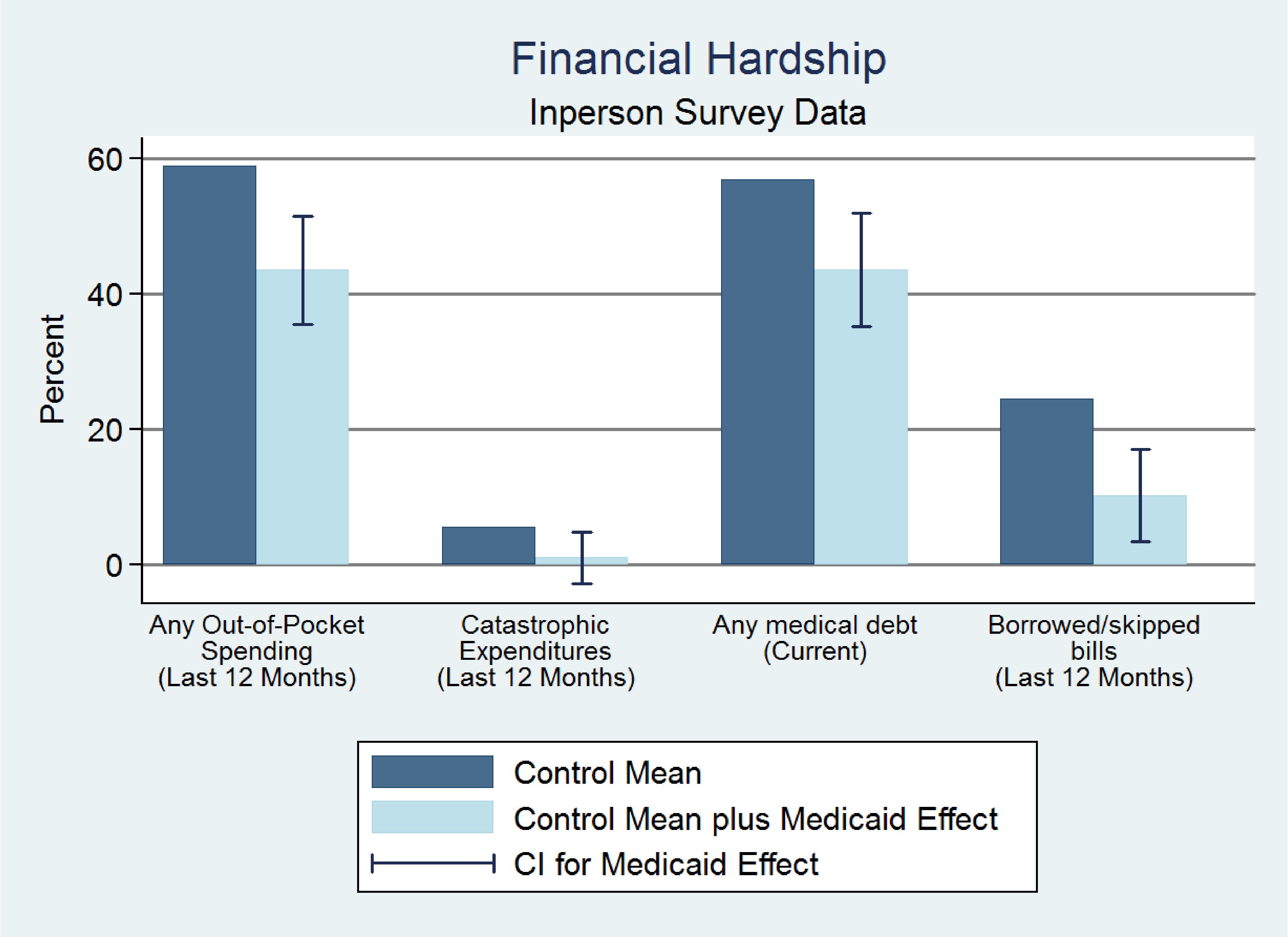
- In 2008, Oregon had capacity to expand Medicaid insurance coverage to individuals between ages 19-64
- Anticipated that budget would not cover all individuals who would want insurance → offered insurance through a randomized lottery
  - Treatment group: 30K individuals who received insurance
  - Control group: 45K individuals who did not
- Evaluate impacts using administrative data from Medicaid and hospitals as well as follow-up surveys
- Series of papers by Baicker, Finkelstein, and co-authors


Preventive Care (Last 12 Months)
Inperson Survey Data
Percent
80
60
40
20
0
Cholesterol checked (all)
Blood stool test (age>=50)
Colonoscopy (age>=50)
Flu Shot (age>=50)
Pap Smear (women)
Mammogram (women>=50)
PSA (men>=50)
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect


Post-lottery Diagnosis (Dx) and Current Medication (Rx)
Inperson Survey Data
Percent
0
10
20
30
Dx
Rx
Hypertension
Dx
Rx
High cholesterol
Dx
Rx
Diabetes
Dx
Rx
Depression
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect


Any and Total ED Use
Emergency Department Data
Percent with Any Visits
0
10
20
30
40
50
Number of Visits
0
.5
1
1.5
Any
Number
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect


Current Clinical Measures
Inperson Survey Data
Percent
0
10
20
30
Elevated blood pressure
High total cholesterol
Low HDL cholesterol
High Hemoglobin A1c
Screen positive for depression
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect


Financial Hardship
Inperson Survey Data
Percent
60
40
20
0
Any Out-of-Pocket Spending (Last 12 Months)
Catastrophic Expenditures (Last 12 Months)
Any medical debt (Current)
Borrowed/skipped bills (Last 12 Months)
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

Oregon Health Insurance Experiment: Lessons

- Insurance coverage increases utilization of health care moderately

- Insurance coverage improves self-reported health and reduces clinical depression

  – Insufficient statistical power to detect effects on physical measures of health

- Insurance coverage significantly reduces financial hardship

Oregon Health Insurance Experiment: Lessons

- Experimental data do not support view that insurance itself leads to substantial “wasteful spending” on health care

- Suggests that broader systemic differences may be more important, such as:

  – Differences in physicians’ practice styles across areas

  – Defensive medicine to protect against lawsuits

  – Monopoly power of hospitals → high prices in some areas [Cooper et al. 2015]

Long-Term Impacts of Health Insurance

- Oregon experiment evaluates *immediate* impact of health insurance

- As with earnings, plausible that health impacts show up with a delay

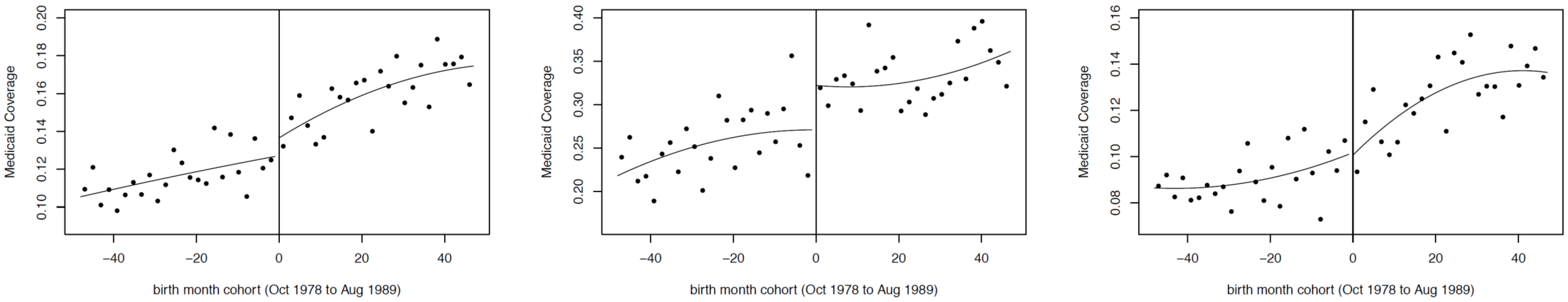
- Does providing Medicaid to children improve long-term outcomes and lower long-run costs (e.g., by reducing hospitalizations)?

Effects of Childhood Medicaid Coverage on Health Care Use and Outcomes in Adulthood

- Wherry and Meyer (2015) and Wherry et al. (2017) study these questions using a regression discontinuity design
  - Medicaid eligibility was expanded for children in low-income families born after September 30, 1983

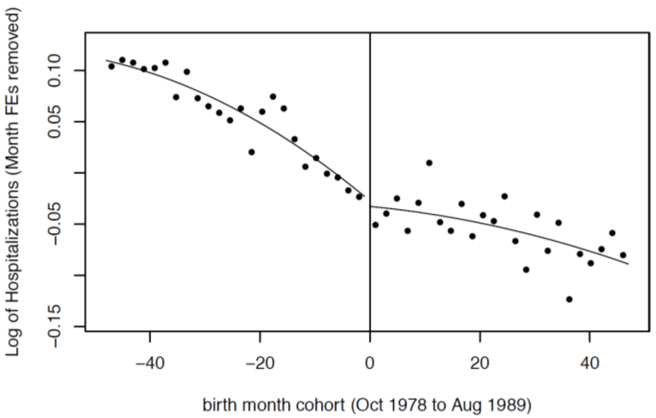
- Data: discharge-level hospital data and outpatient emergency department visits in California, Texas, New York, and other states
  - No data on income → compare black vs. white children instead

Fraction of Children with Medicaid Coverage Between the Ages of 8 and 13, by Birth Month

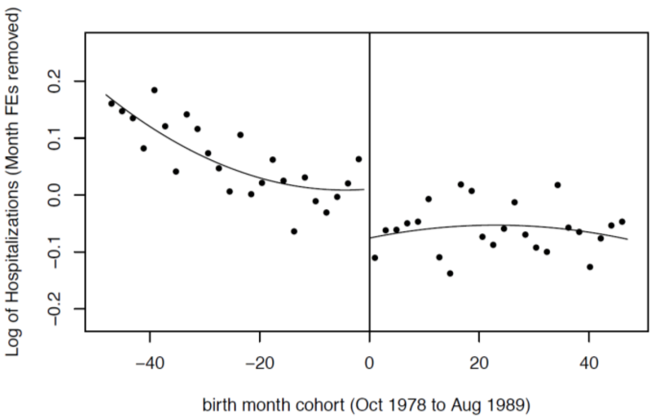


(a) All Races (b) Blacks (c) Non-Blacks

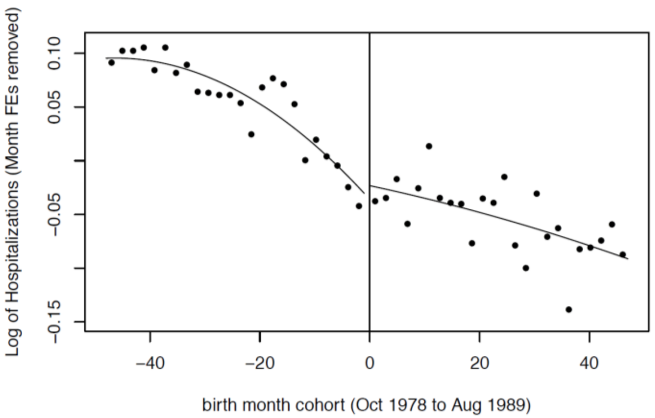
Hospitalizations in 2009 (mid 20s) by Month of Birth



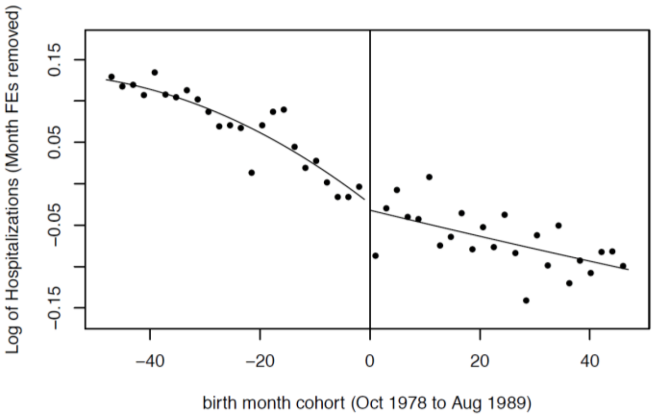
(a) All Hospitalizations, All Races



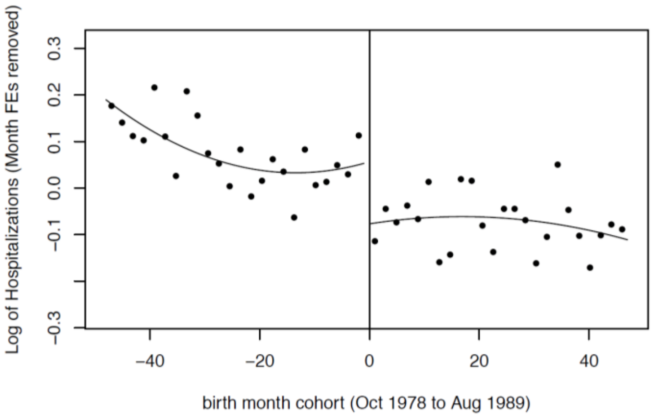
(b) All Hospitalizations, Blacks



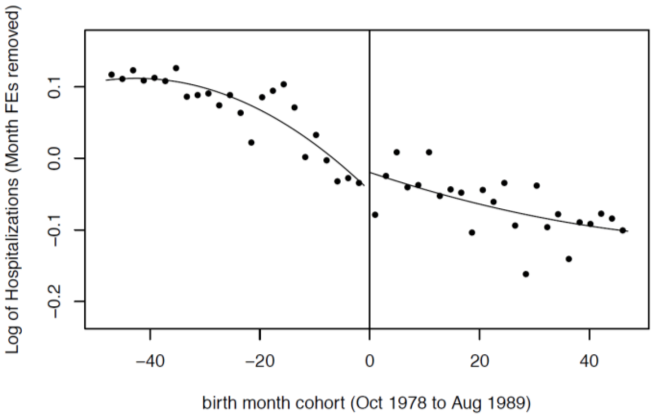
(c) All Hospitalizations, Non-Blacks



(d) Chronic Illness Hospitalizations, All Races

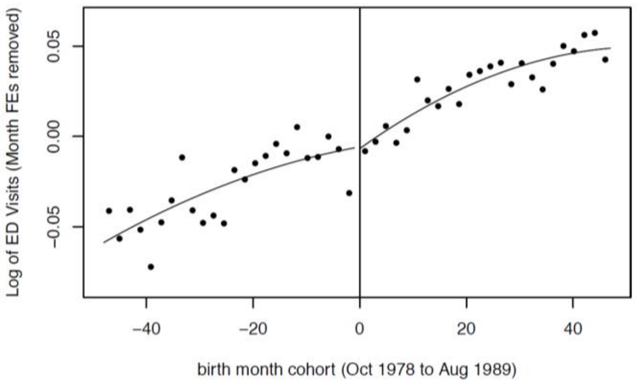


(e) Chronic Illness Hospitalizations, Blacks

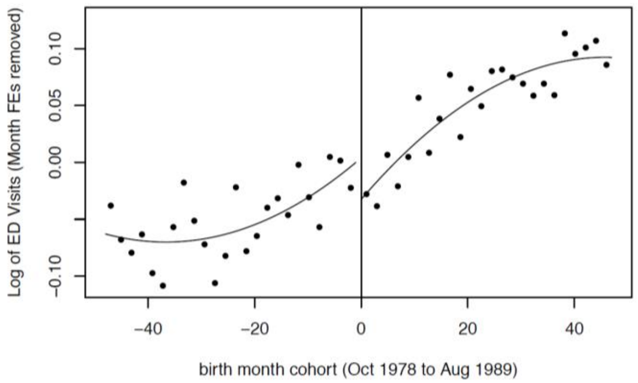


(f) Chronic Illness Hospitalizations, Non-Blacks

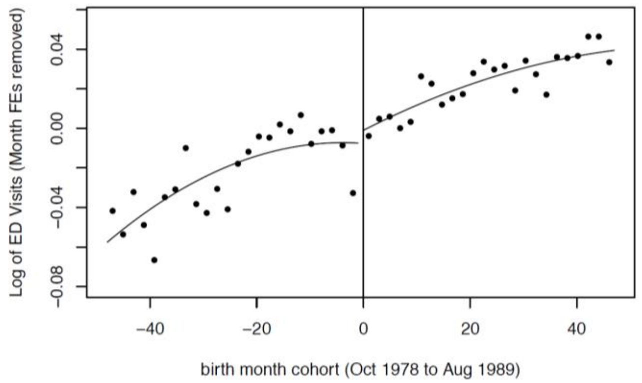
Emergency Department Visits in 2009 (mid 20s) by Month of Birth



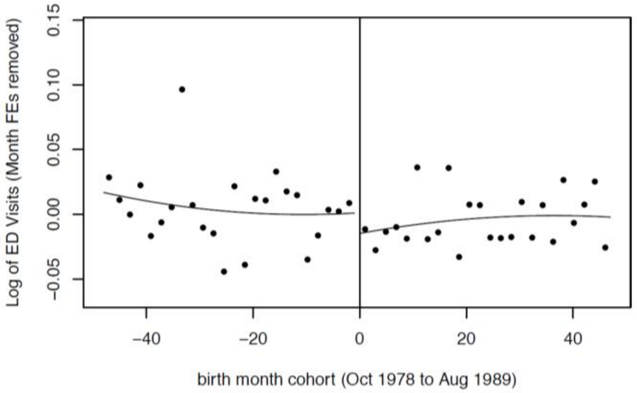
(a) All ED Visits, All Races



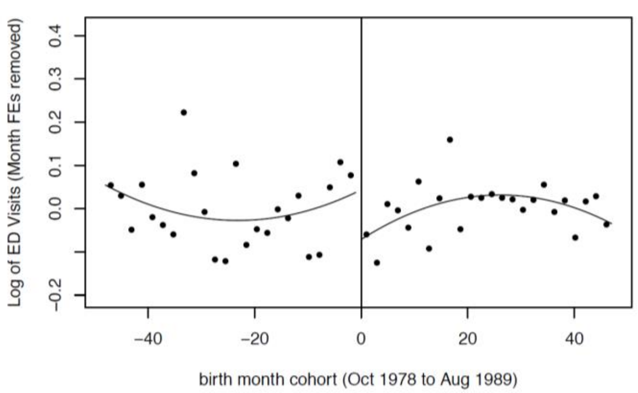
(b) All ED Visits, Blacks



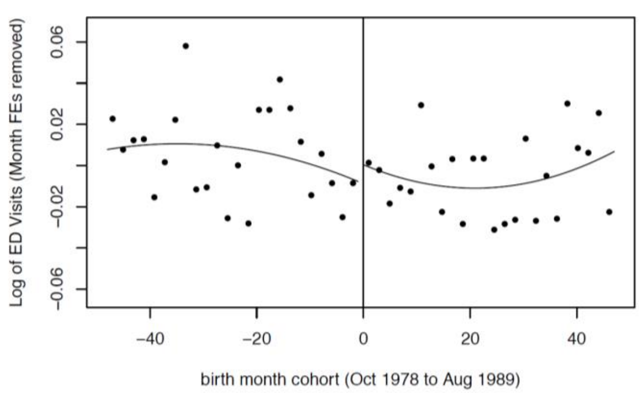
(c) All ED Visits, Non-Blacks



(d) Chronic Illness ED Visits, All Races

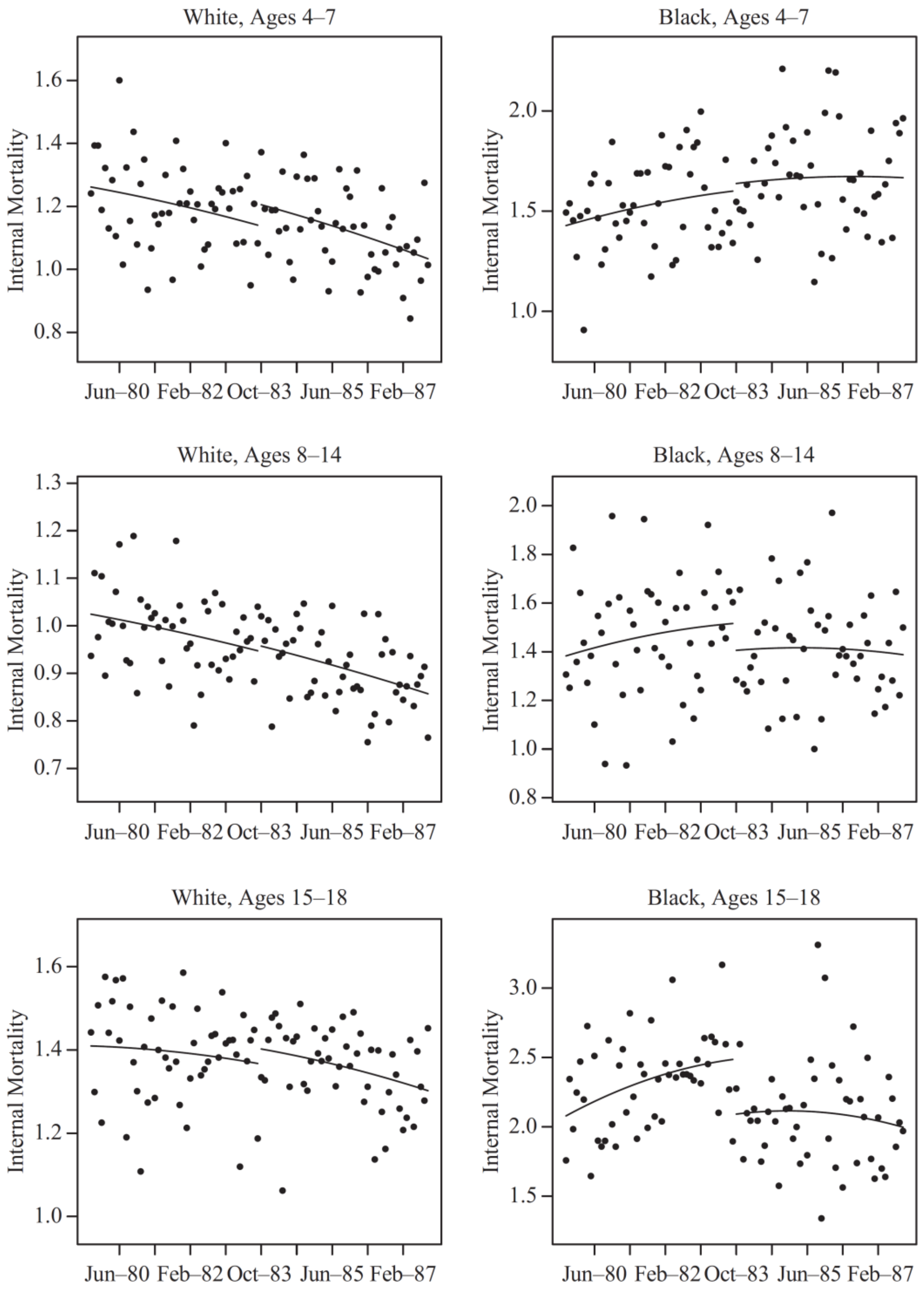


(e) Chronic Illness ED Visits, Blacks

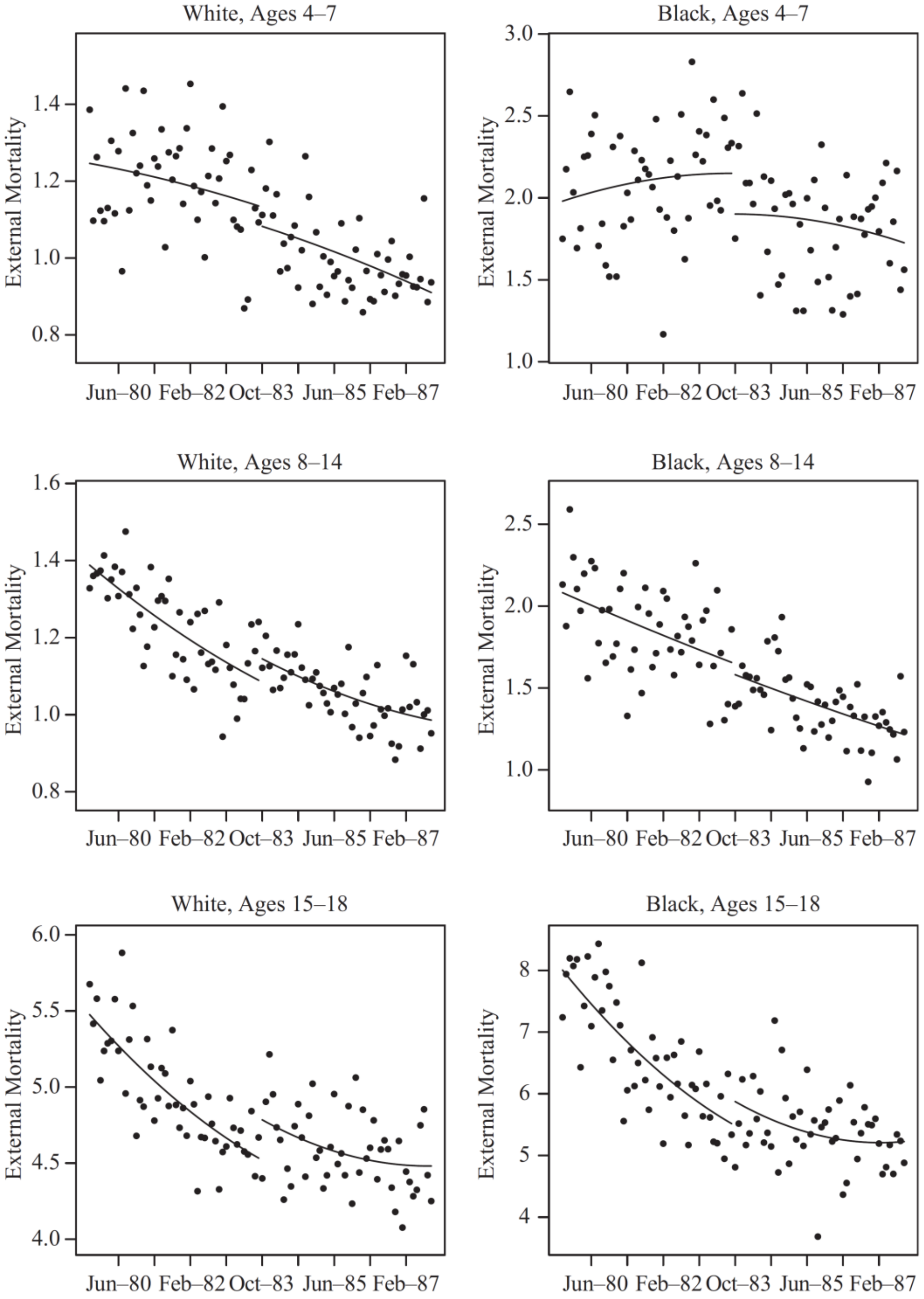


(f) Chronic Illness ED Visits, Non-Blacks

Mortality Rates by Month of Birth: Internal Causes



Mortality Rates by Month of Birth: External Causes



Government Intervention in Markets for Health Insurance

- Data show that insurance coverage leads to moderate increases in health care use and improvement in health outcomes

- Suggests that access to health insurance can be valuable for improving population health

- But does not necessarily follow that government needs to provide this insurance

  – Why can’t people buy it themselves in private markets, like they do other products like cars?

Summary: Health Care and Insurance in the U.S.

- Insurance matters for health outcomes and financial security

- Difficult to sustain markets for insurance without government insurance or direct government provision (single payer system)

- Insurance contributes modestly to higher costs

  - But reasons that health care costs are so high and so variable in the U.S. remain unclear

Summary: Health Care and Insurance in the U.S.

- Better data are likely to help in terms of answering this question and increasing accountability
  - Currently, prices are not even clear to patients and providers → little pressure to reduce or even monitor costs for any party