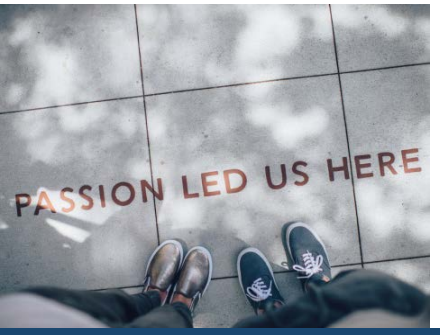


Using Big Data to Solve Economic and Social Problems

Professor Raj Chetty
Head Section Leader: Gregory Bruich, Ph.D.

Spring 2019

HARVARD UNIVERSITY

OPPORTUNITY INSIGHTS

Improving Health Outcomes: Overview

- This part of the class illustrates how big data is helping us learn how to improve health, in three segments:

1. Descriptive analysis of health outcomes in U.S. population
   [method: survival analysis]

2. Economics applications: impacts of food stamps and health insurance
   [method: regression discontinuities]

3. Epidemiology application: using big data to forecast pandemics
   [method: predictive modeling

The Economics of Health Insurance and Health Care

The Economics of Health Insurance and Health Care

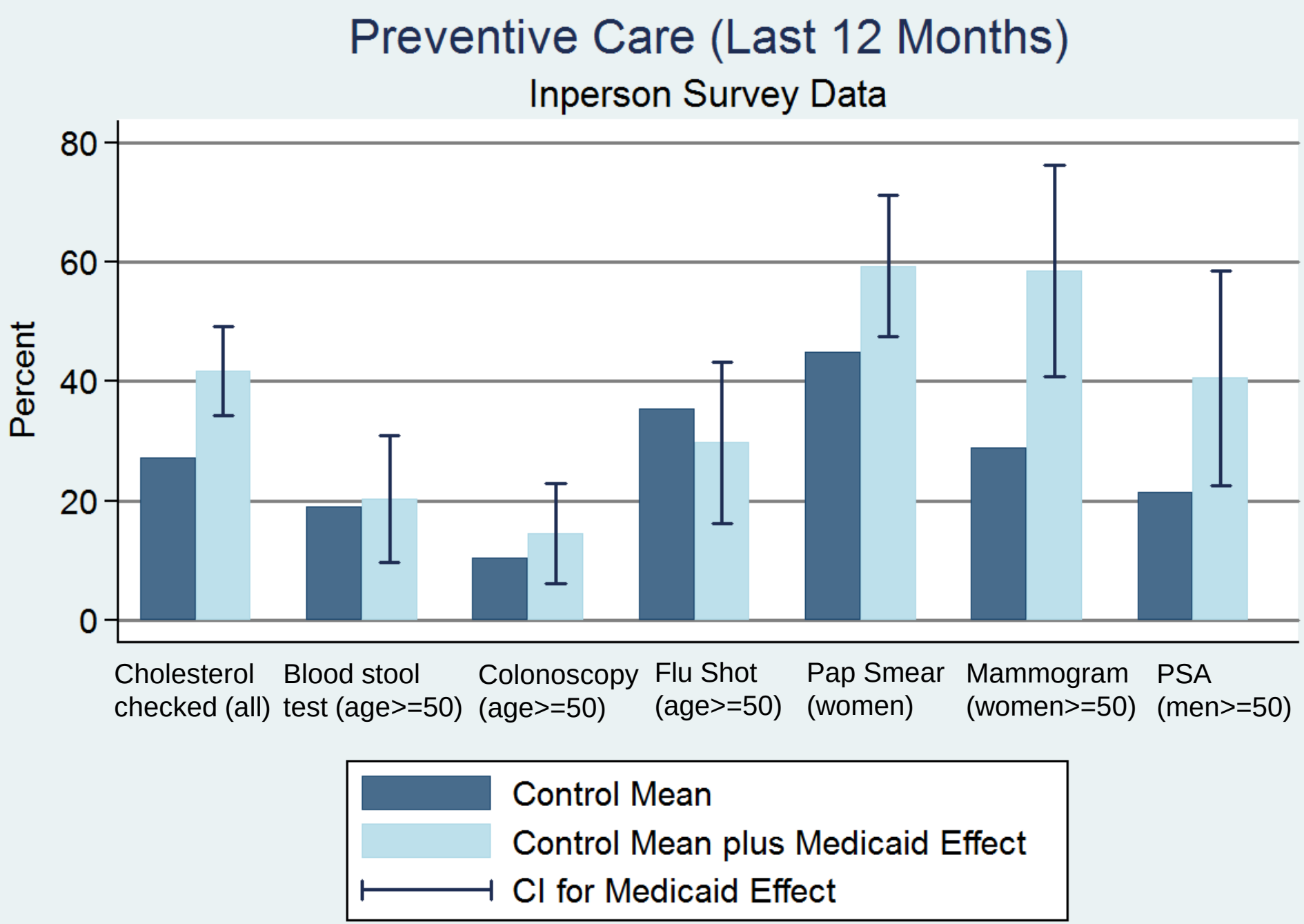
- Health economists focus on studying markets for health care
  - Will expanding health insurance coverage improve health outcomes and reduce health inequality?
  - If so, how can we provide health insurance to more Americans?

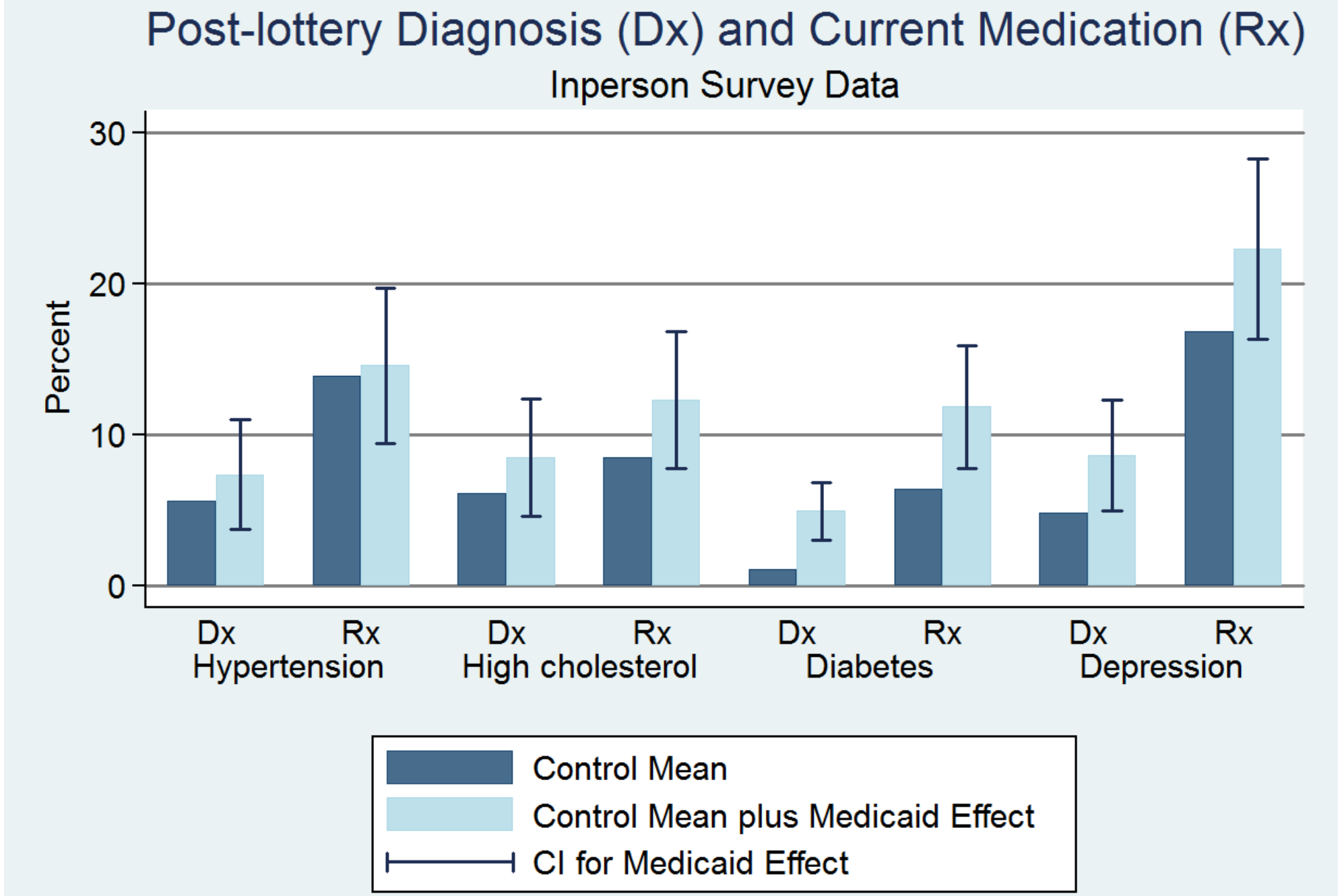
Insurance and Demand for Health Care

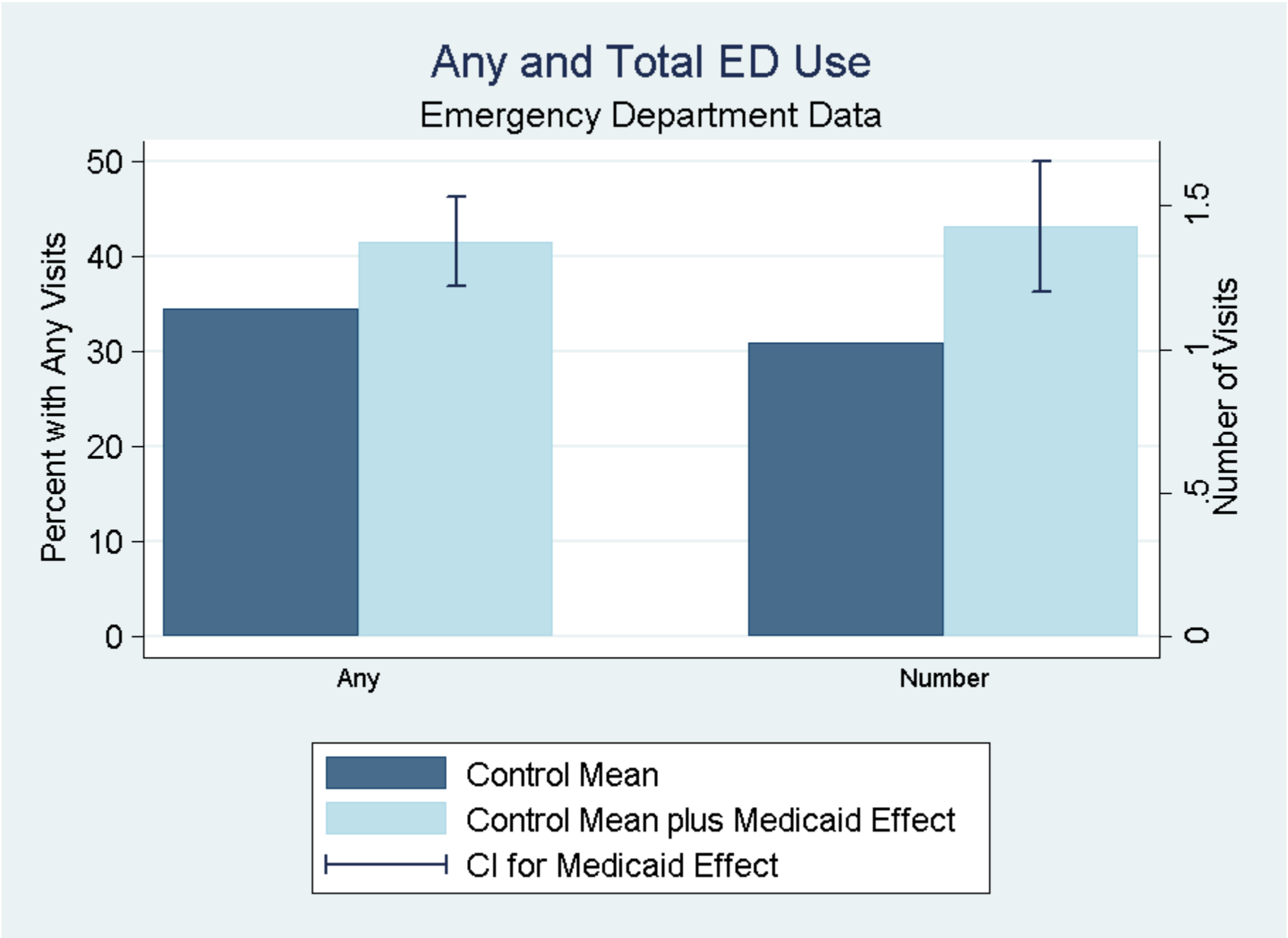
- What is the causal effect of insurance on demand for health care and health outcomes?
  - Does providing individuals' insurance actually encourage wasteful spending or does it improve health outcomes?
- Ideal experiment: randomly assign health insurance to some individuals and not others and compare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

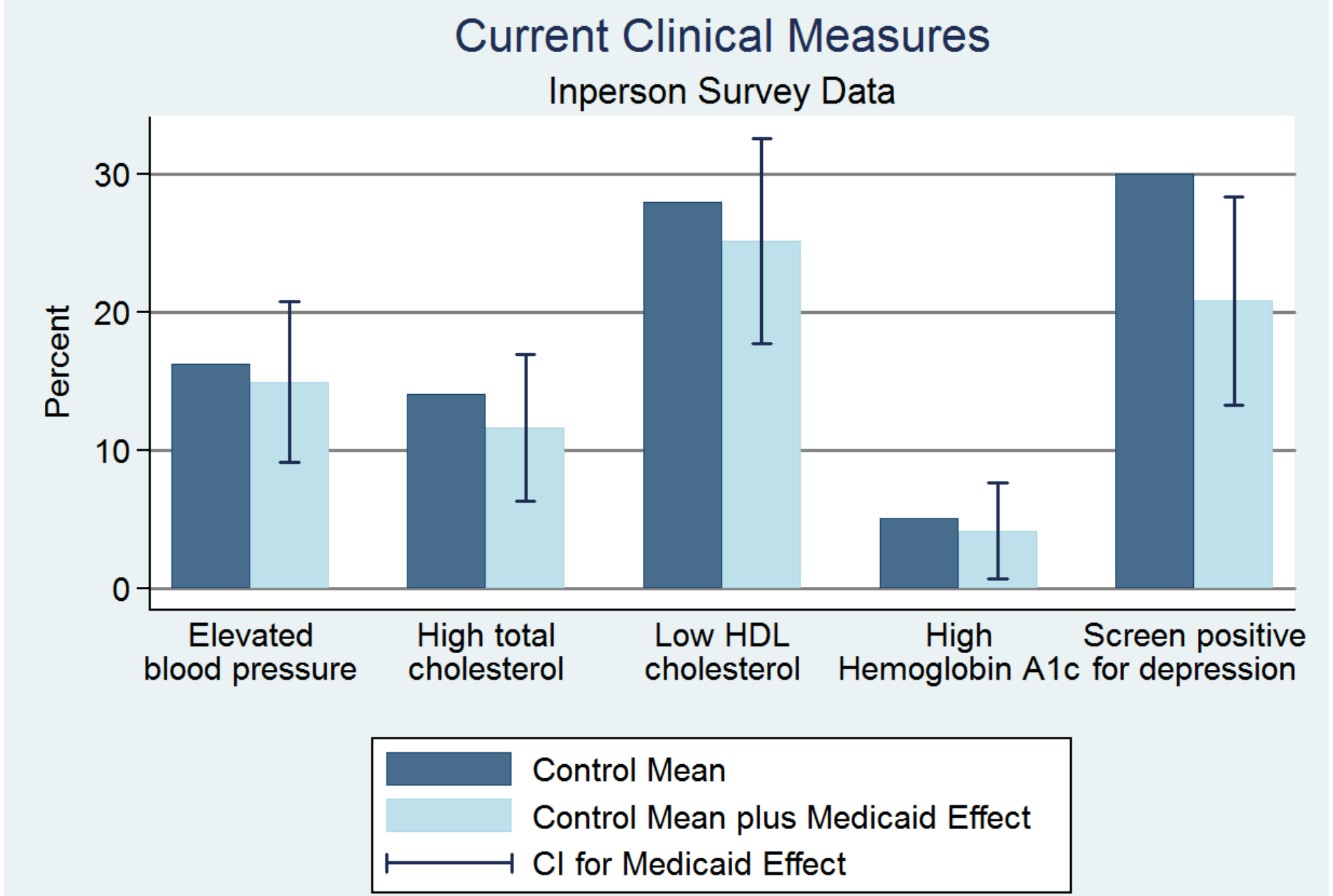
Oregon Health Insurance Experiment

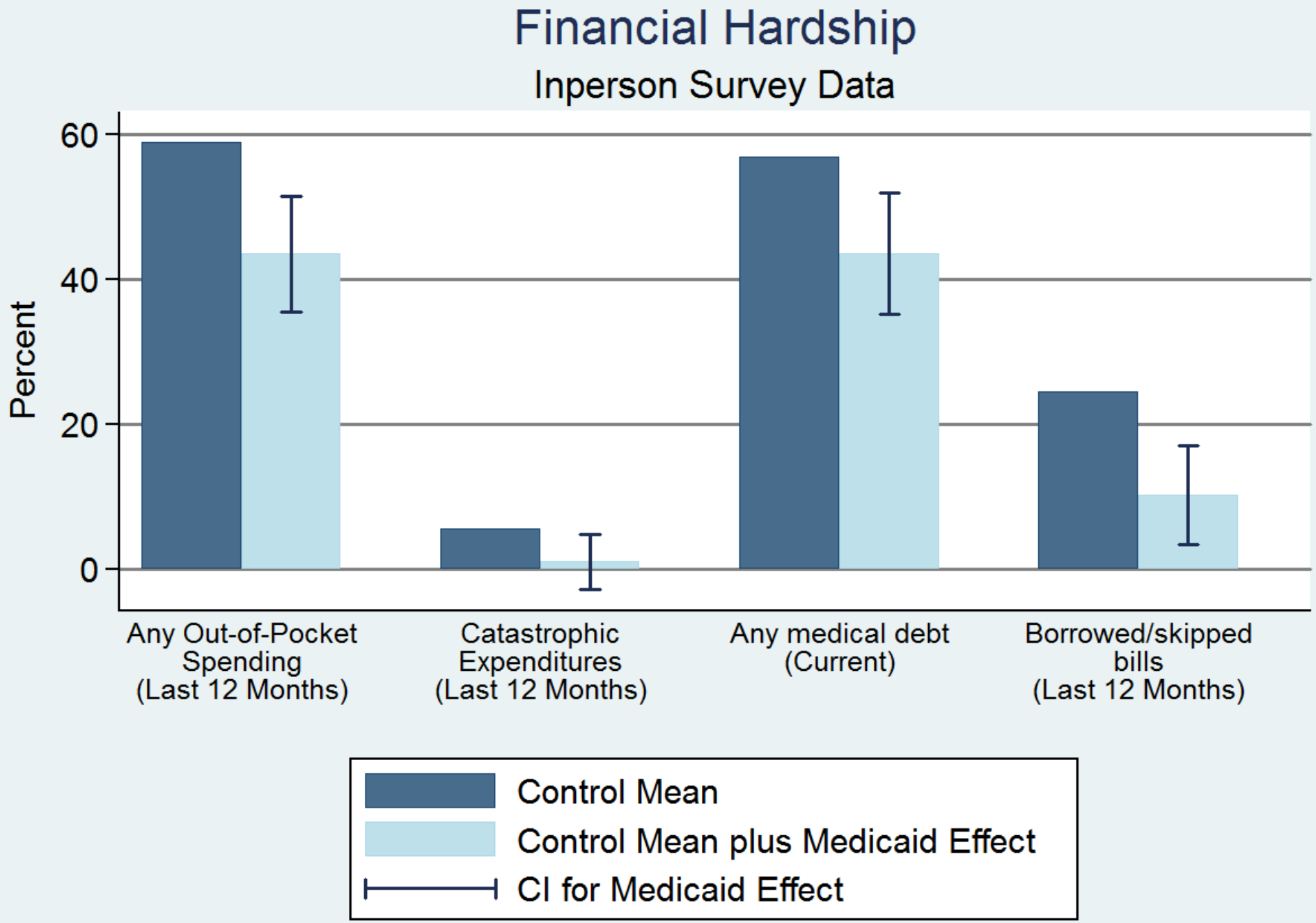
- In 2008, Oregon had capacity to expand Medicaid insurance coverage to individuals between ages 19-64
- Anticipated that budget would not cover all individuals who would want insurance → offered insurance through a randomized lottery
  - Treatment group: 30K individuals who received insurance
  - Control 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
30
20
10
0
Dx
Rx
Hypertension
High cholesterol
Diabetes
Depression
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect


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


Current Clinical Measures
Inperson Survey Data
Percent
30
20
10
0
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 significantly reduces financial hardship
- Insurance coverage improves self-reported health and reduces clinical depression
  - Insufficient statistical power to detect effects on physical measures of health
  - But probably insufficient to explain large differences in health outcomes by income

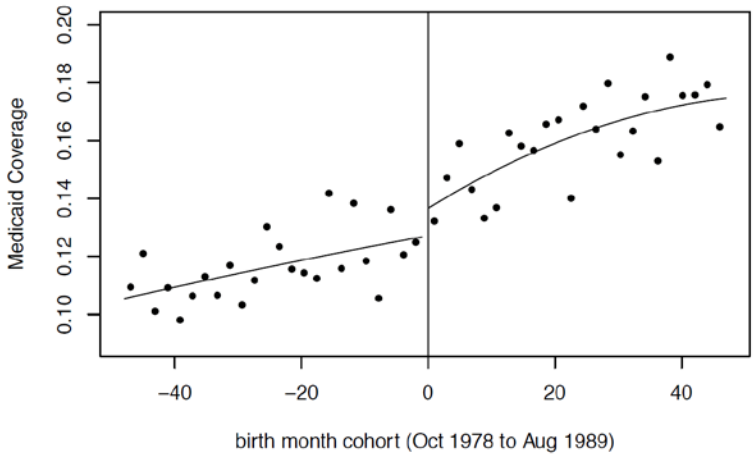
Long-Term Impacts of Health Insurance

- 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)?

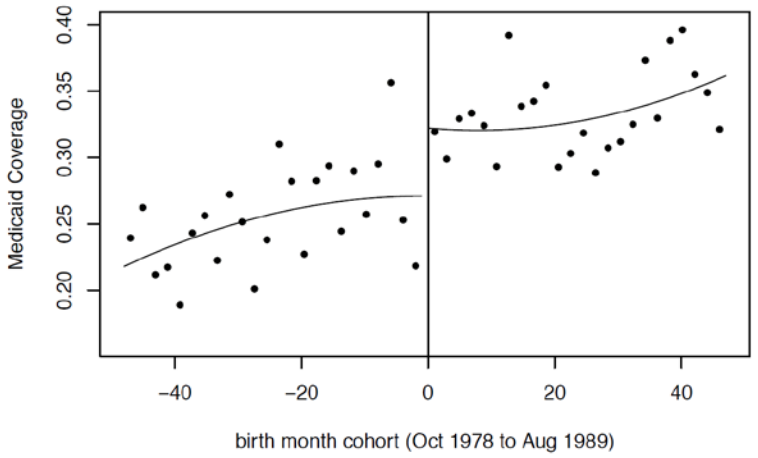
Effects of Childhood Medicaid Coverage on Health Care Use and Outcomes in Adulthood

- 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

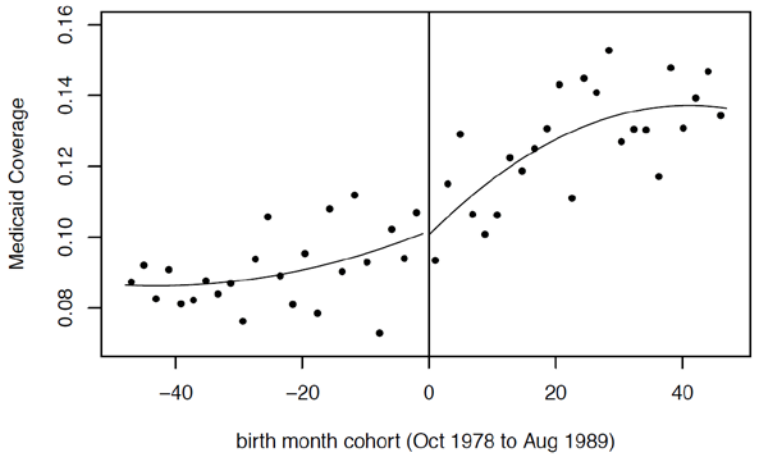
Fraction of Children with Medicaid Coverage Between the Ages of 8 and 13, by Birth Month



(a) All Races

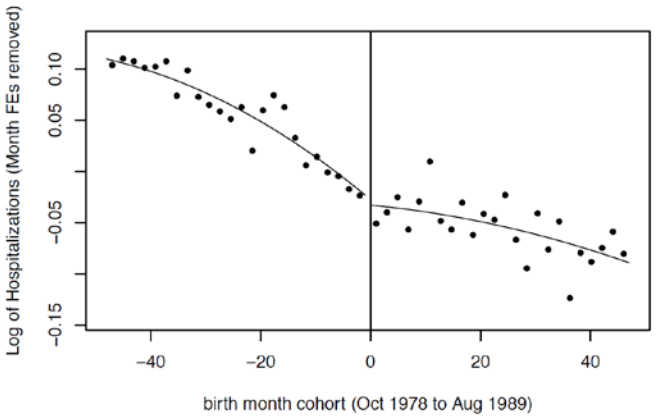


(b) Blacks

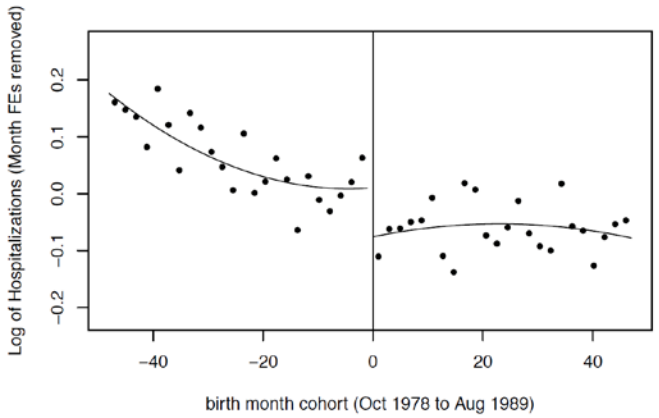


(c) Non-Blacks

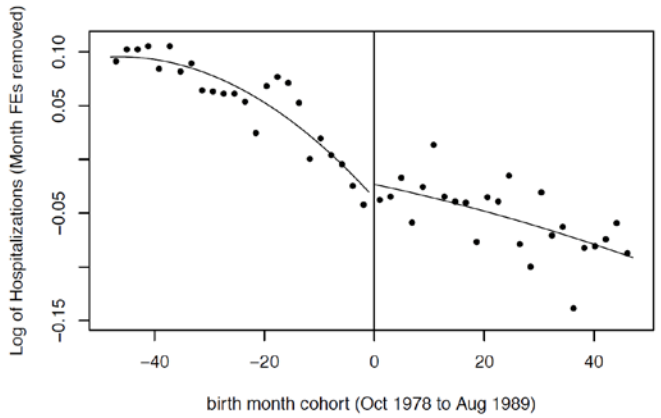
Hospitalizations in 2009 (mid 20s) by Month of Birth



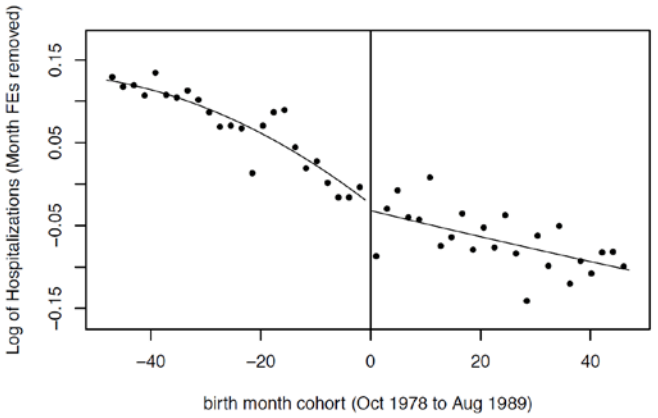
(a) All Hospitalizations, All Races



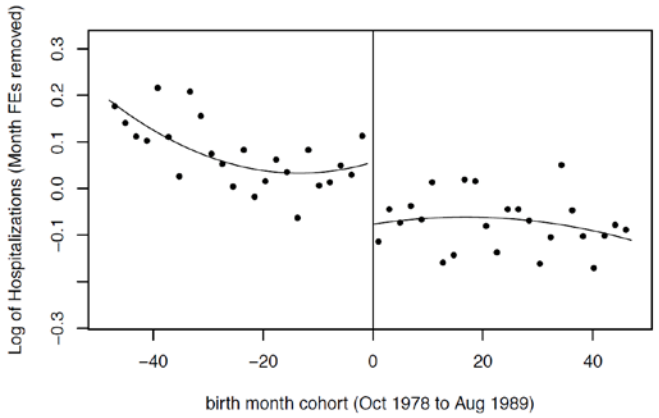
(b) All Hospitalizations, Blacks



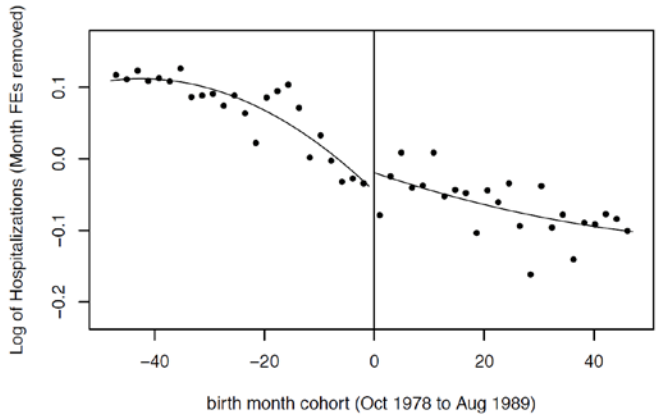
(c) All Hospitalizations, Non-Blacks



(d) Chronic Illness Hospitalizations, All Races

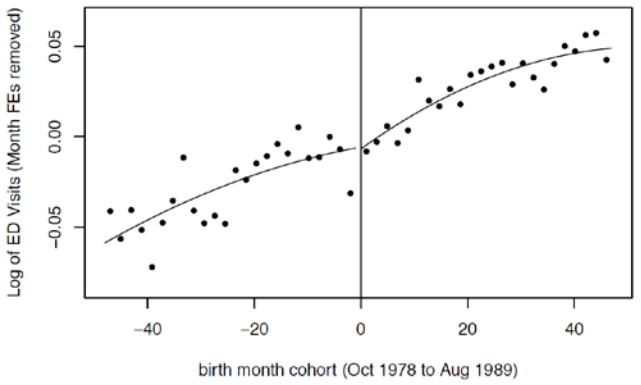


(e) Chronic Illness Hospitalizations, Blacks

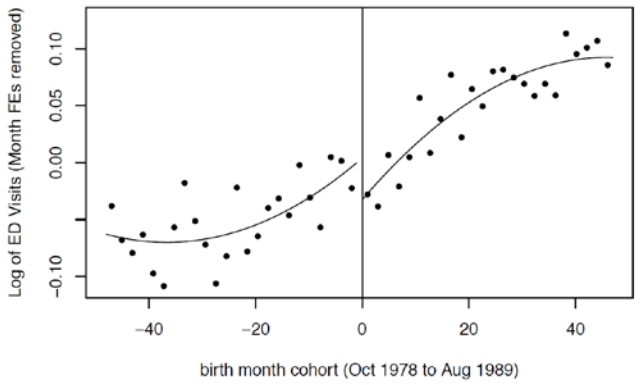


(f) Chronic Illness Hospitalizations, Non-Blacks

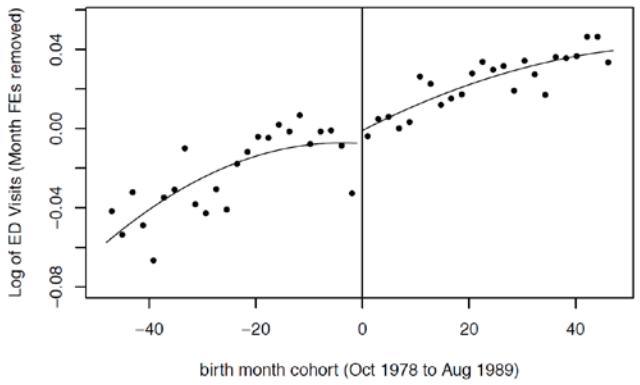
Emergency Department Visits in 2009 (mid 20s) by Month of Birth



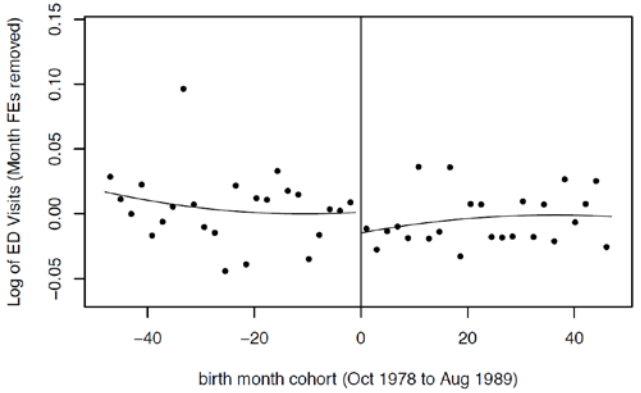
(a) All ED Visits, All Races



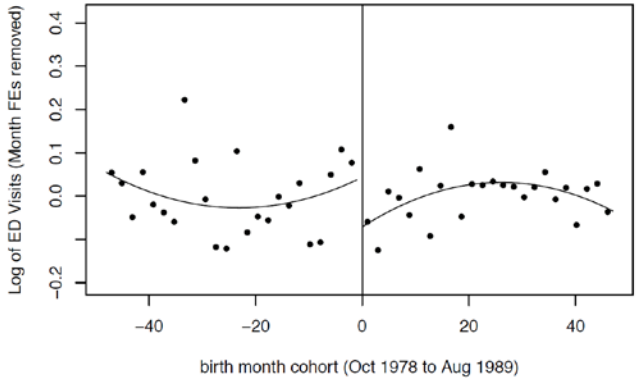
(b) All ED Visits, Blacks



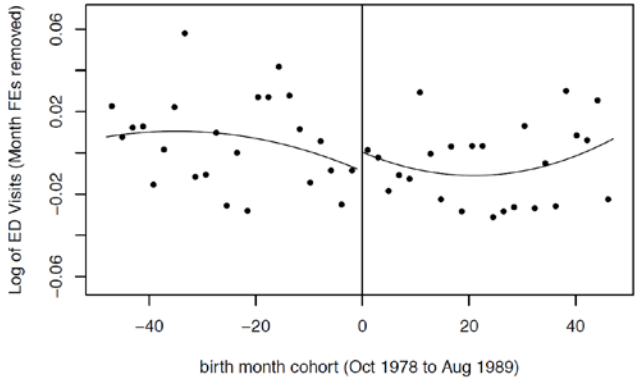
(c) All ED Visits, Non-Blacks



(d) Chronic Illness ED Visits, All Races

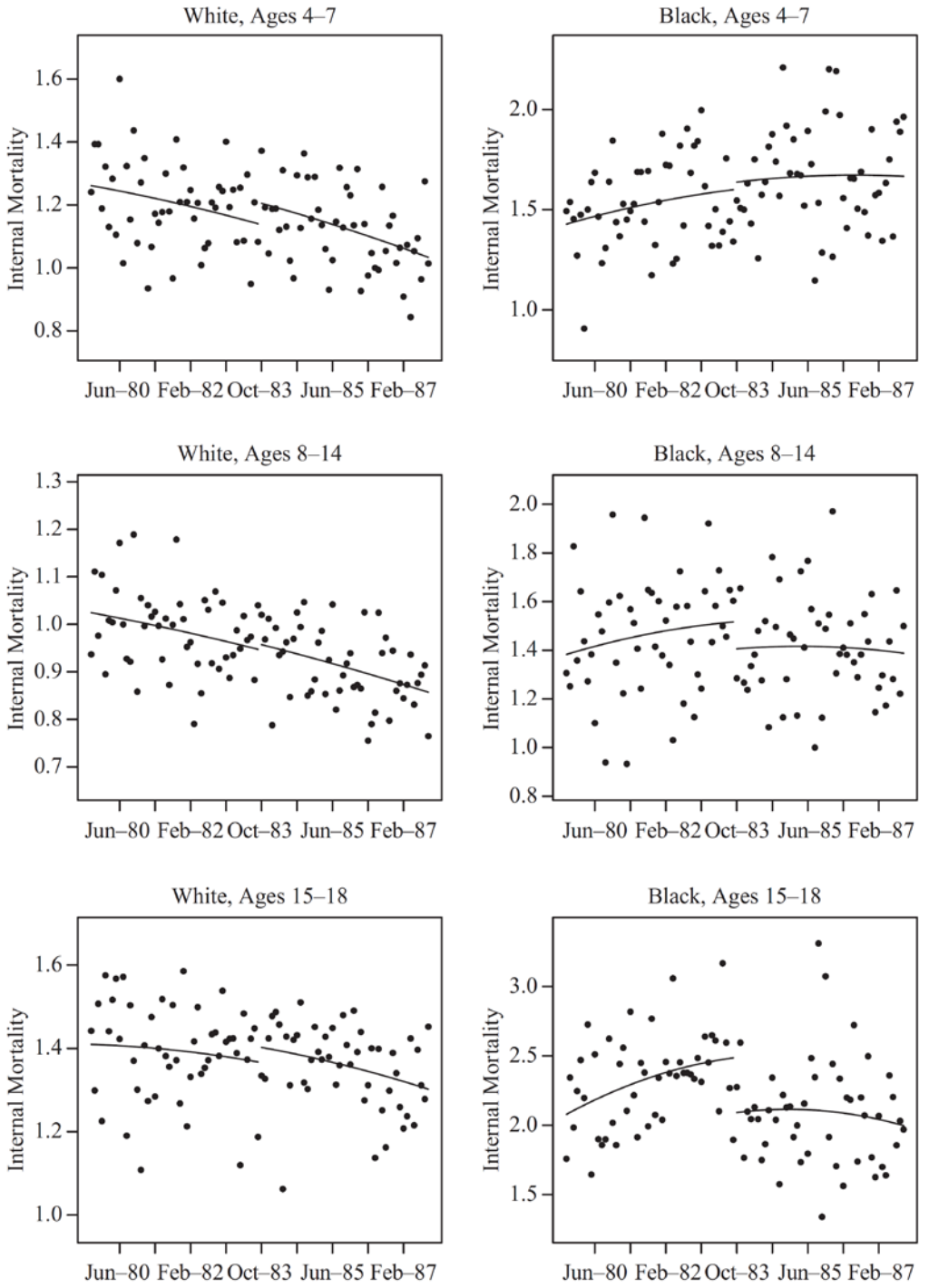


(e) Chronic Illness ED Visits, Blacks

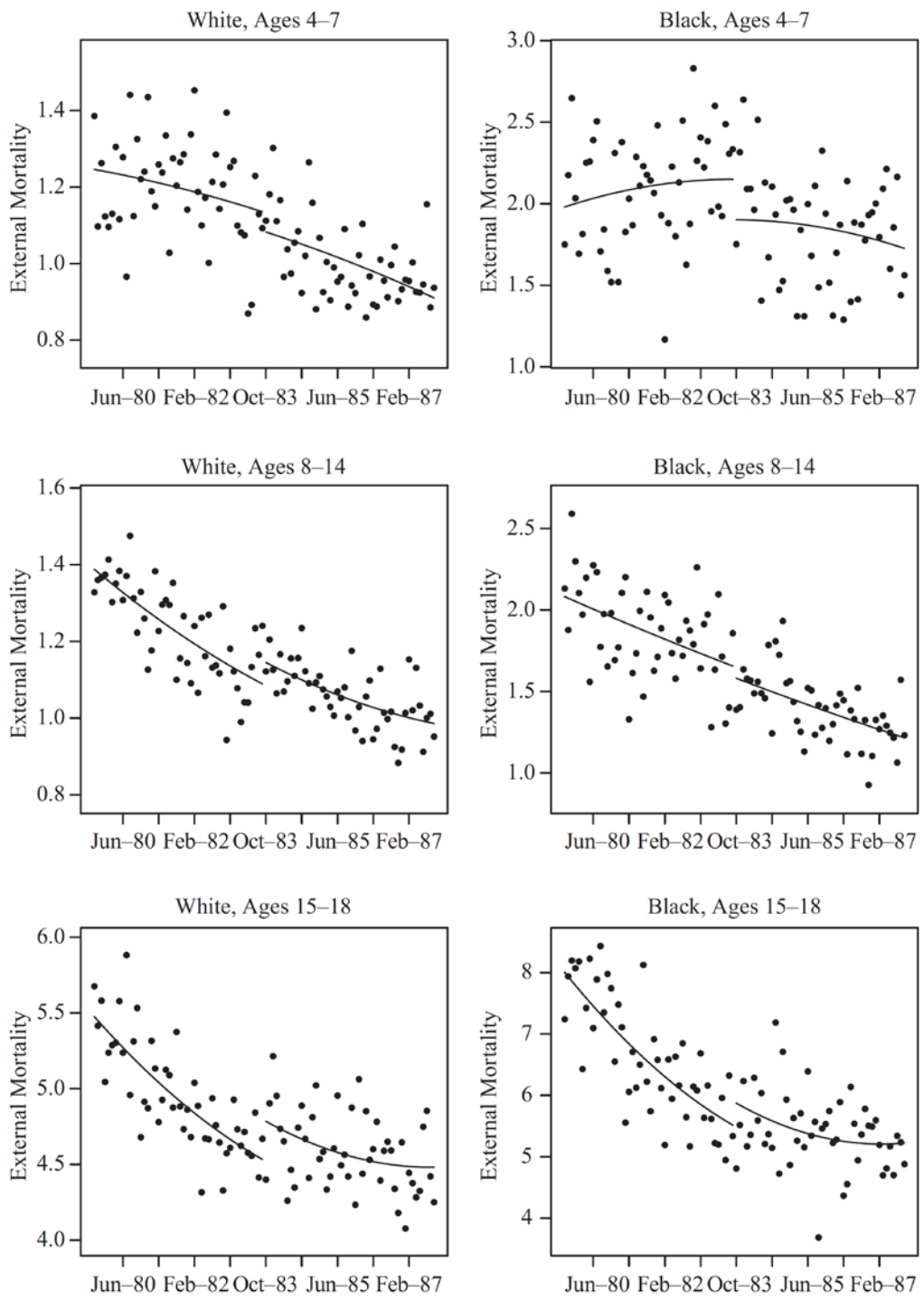


(f) Chronic Illness ED Visits, Non-Blacks

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, especially in long-run

- 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?

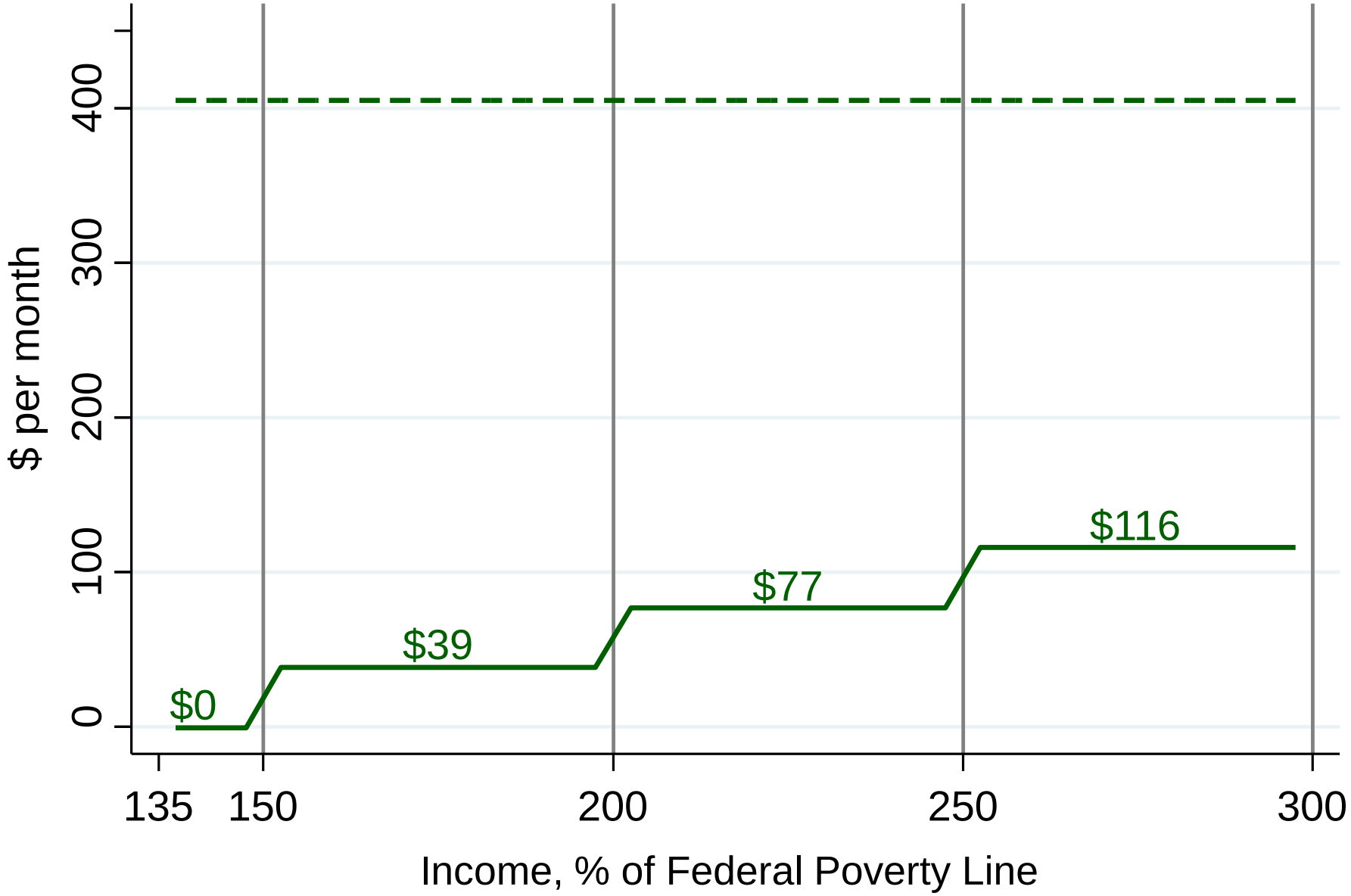
Demand for Health Insurance

- Finkelstein, Hendren, and Shepard (2017) show why government intervention is essential to sustain markets for insurance

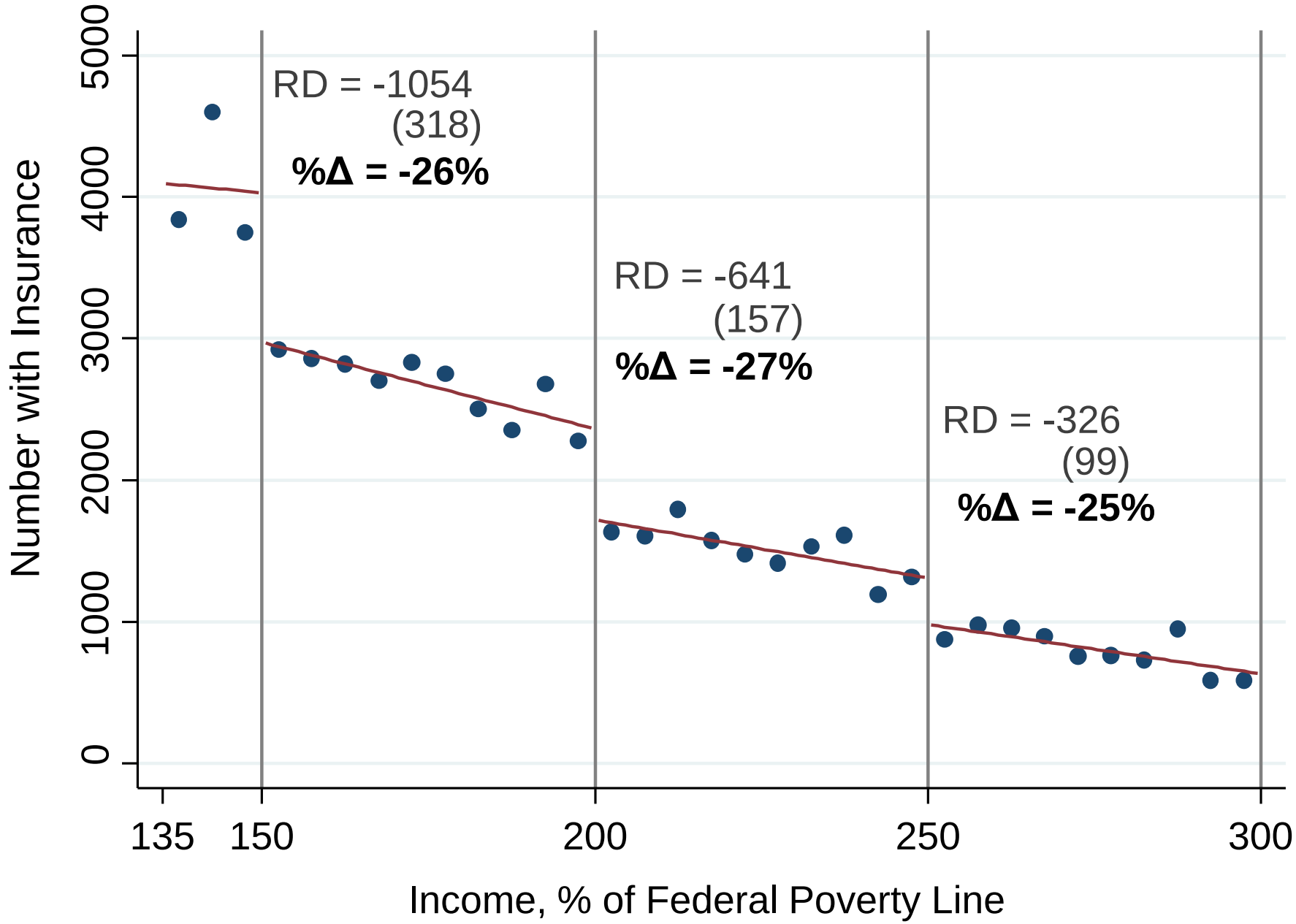
- Study Massachusetts public universal health insurance program
  - Introduced in 2006; predecessor to the national Affordable Care Act

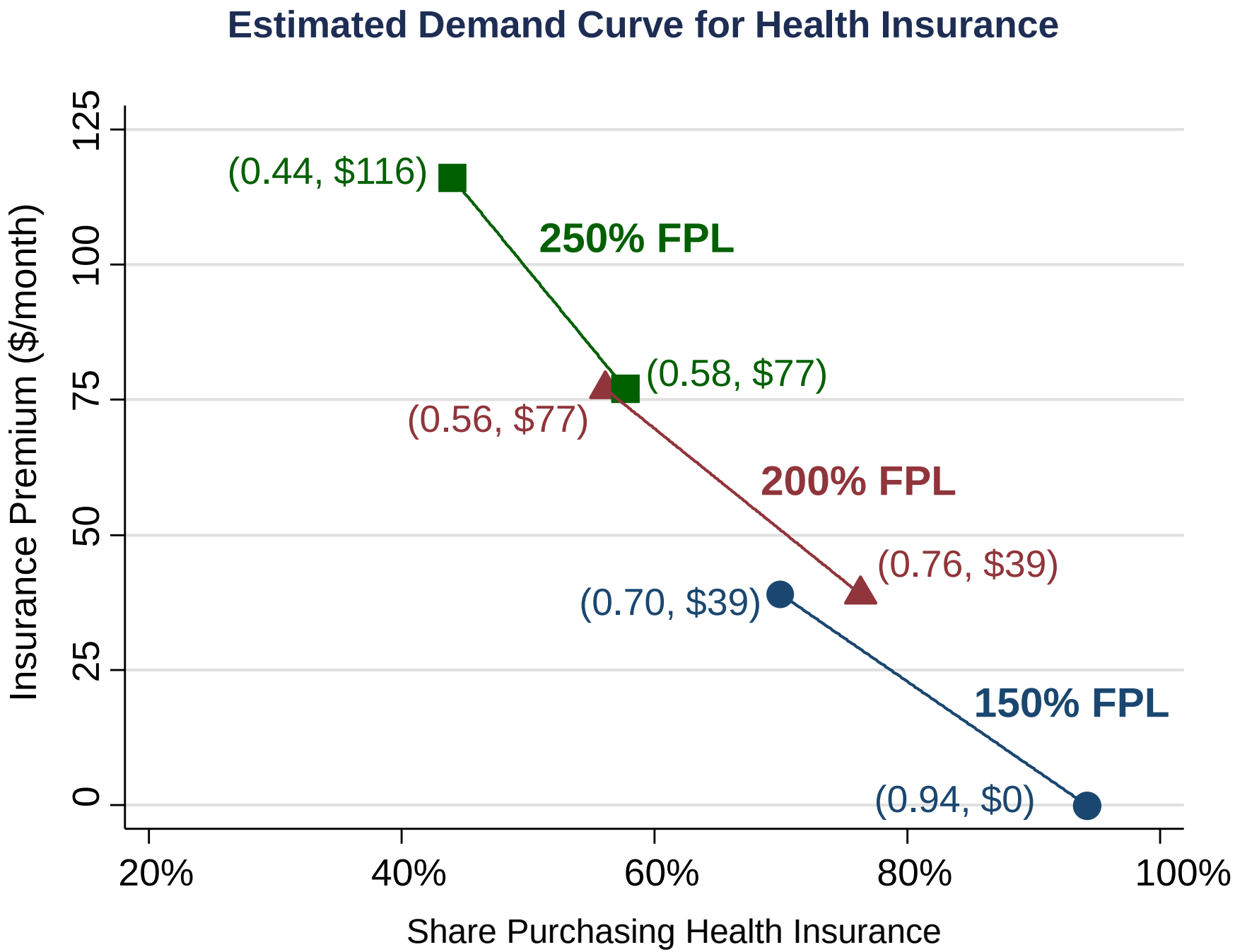
- Research design: exploit discontinuities in subsidies for insurance based on income level

Subsidy and Premium Discontinuities in Massachusetts in 2011



Number of Individuals Enrolled in a Health Insurance Plan by Income Level

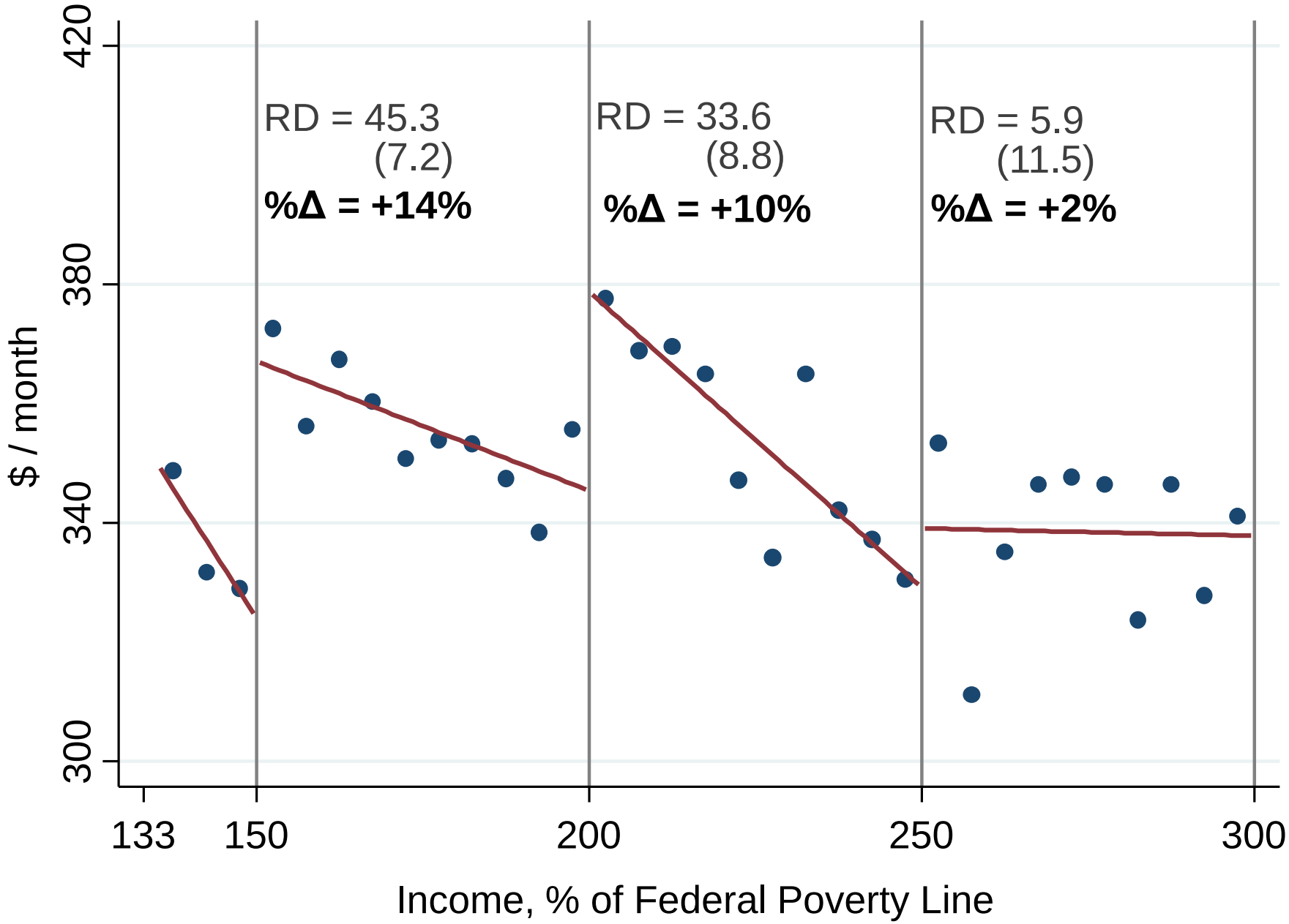



Estimated Demand Curve for Health Insurance
Insurance Premium ($/month)
0
25
50
75
100
125
20%
40%
60%
80%
100%
Share Purchasing Health Insurance
(0.44, $116)
250% FPL
(0.58, $77)
(0.56, $77)
200% FPL
(0.76, $39)
(0.70, $39)
150% FPL
(0.94, $0)

Demand for Health Insurance

- Demand for health insurance among the poor falls very rapidly as price rises
  - Reducing subsidies would drastically reduce the number of individuals insured
- Moreover, *sicker* people remain insured, increasing average costs for insurers – **adverse selection**

Adverse Selection: Average Costs Paid by Insurance Companies



Lessons on Markets for Health Insurance

- Government intervention is critical to sustain markets for health insurance for two reasons:

  1. Low-income individuals are very sensitive to price → will not buy insurance if not subsidized or provided by government

  2. Healthiest low-income individuals are least likely to buy → insurance companies get stuck with higher costs, making market collapse

Lessons on Markets for Health Insurance

- Adverse selection can lead to a “death spiral” in private insurance markets [Akerlof 1971]
  - Actually happened when Harvard changed health plans for its employees in late 1990s [Cutler and Reber 1998]
- Hendren: Trumpcare would effectively end enrollment in insurance markets for families that make less than $75,000 a year

The New Study That Shows Trumpcare's Damage

David Leonhardt MAY 3, 2017

From left, Republican representative Fred Upton, Michael Burgess, Greg Walden and Billy Long after meeting with President Trump on Wednesday. Stephen Crowley/The New York Times

When Massachusetts expanded health insurance a decade ago, state officials unknowingly created an experiment. It's turned out to be an experiment that offers real-world evidence of what would happen if the House Republicans' health bill were to become law.

The findings from Massachusetts come from an academic paper being released Thursday, and the timing is good. Until now, the main analysis of the Republican health bill has come from the Congressional Budget Office, and some Republicans have criticized that analysis as speculative. The Massachusetts data is more concrete.

Epidemiology and Public Health: Forecasting Pandemics

Epidemiology and Public Health in the Era of Big Data

- Public health/epidemiology focuses on improving health by:
  - Changing individual behaviors: e.g., smoking and exercise
  - Population health: e.g., improving water quality, reducing spread of diseases
- Focus here on how big data is starting to enter this field by focusing on one classic question: forecasting and preventing health pandemics

Forecasting Pandemics

- Contagious diseases like flu spread exponentially → large returns to taking action quickly when disease emerges
- Most common method to monitor contagious diseases in developed countries: aggregated data from local clinics
- Problem: slow reporting and small samples → data not very fine-grained

Forecasting Pandemics: Google Flu Trends

- Ginsberg et al. (2009) propose a new data source to monitor spread of the flu: Google search data

- Idea: people often search for terms like “antibiotics” or “how to treat cough” when getting sick

- Use aggregated search data to get predictions of spread of flu that are (a) more timely and (b) available at fine geographies

Forecasting Pandemics: Google Flu Trends

- Method: predictive modeling
  - Get historical data on truth from CDC and estimate a statistical model using Google search data to predict that data
  - Then evaluate the model using future data that was not used for estimation to evaluate model's predictive accuracy

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

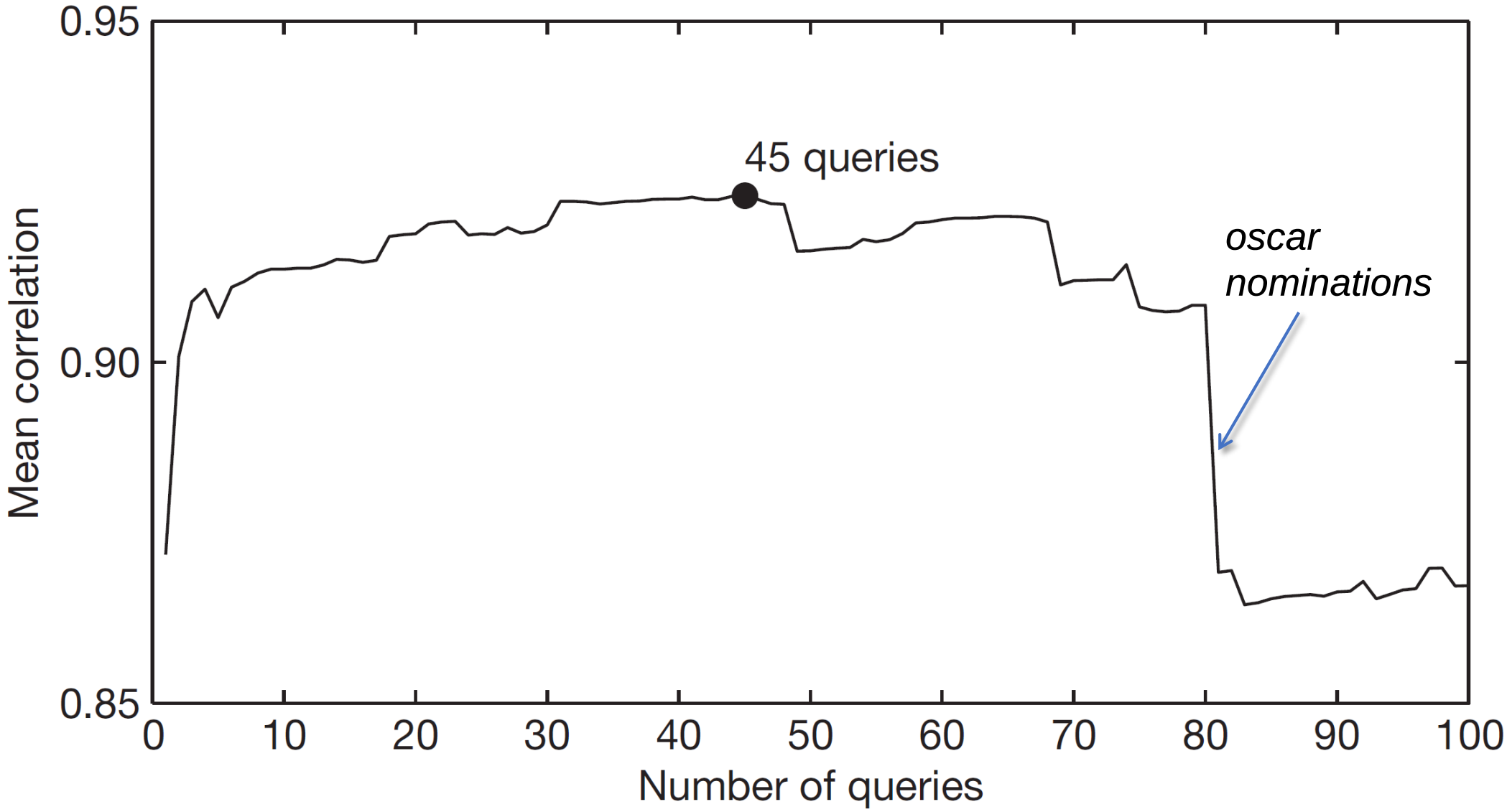
- Construct predictive model in a series of steps:

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

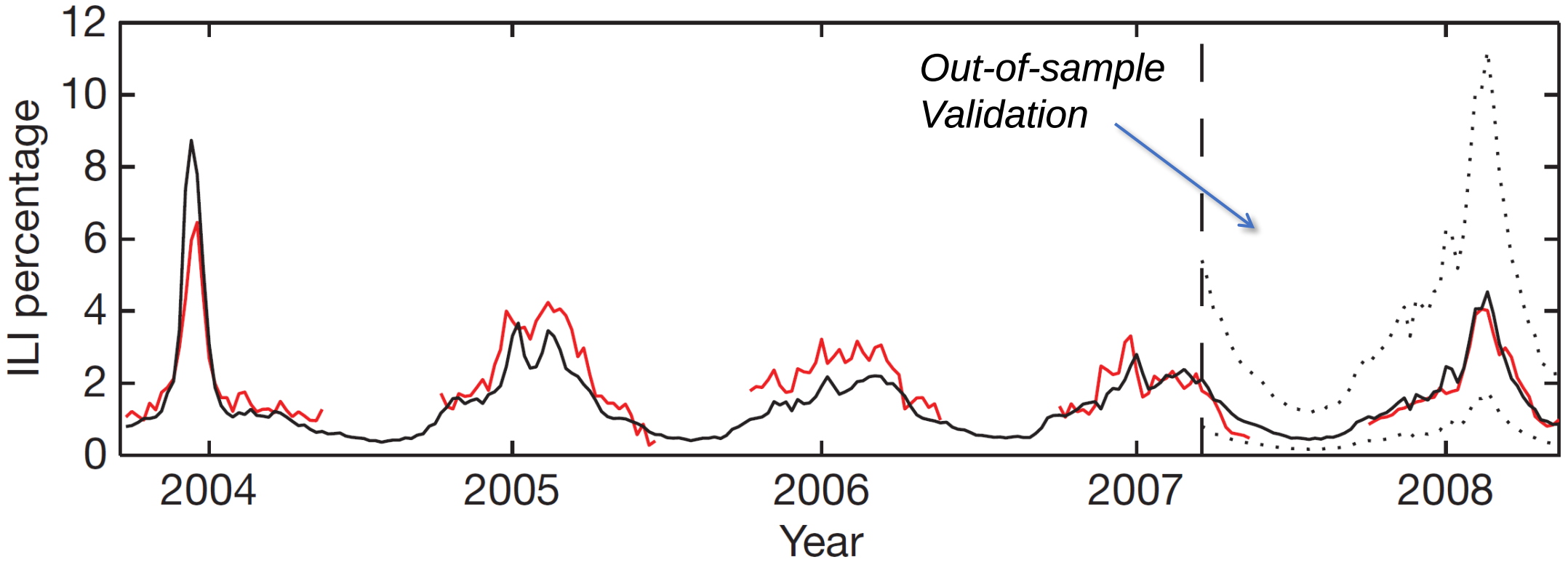


Google Flu Trends: Methodology

- Construct predictive model in a series of steps:

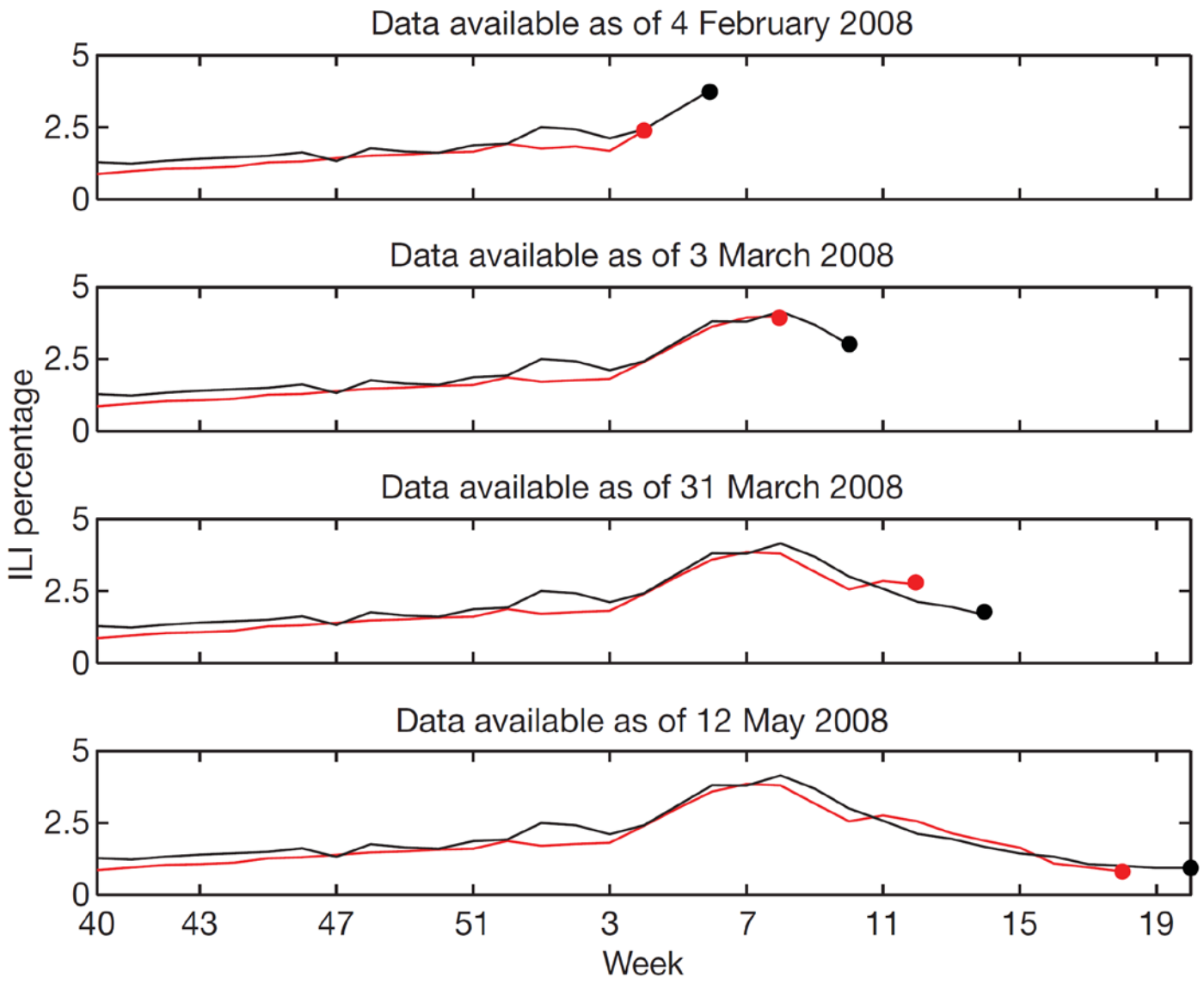
3. Finally, evaluate model fit and out of sample predictive accuracy using subsequent data that was not available when model was estimated

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



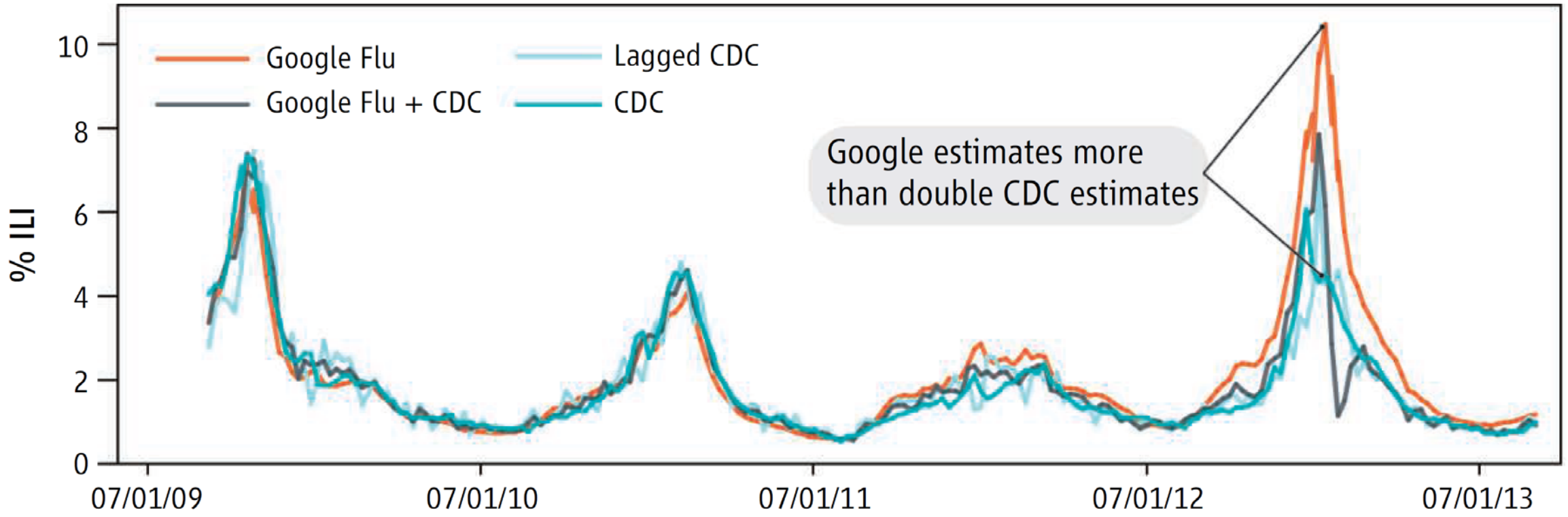
Note: CDC official statistics in red; Google trends forecast in black

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