Health Economics and Organizational Economics, Part 1: Productivity and Microfoundations

David Chan

December 2019

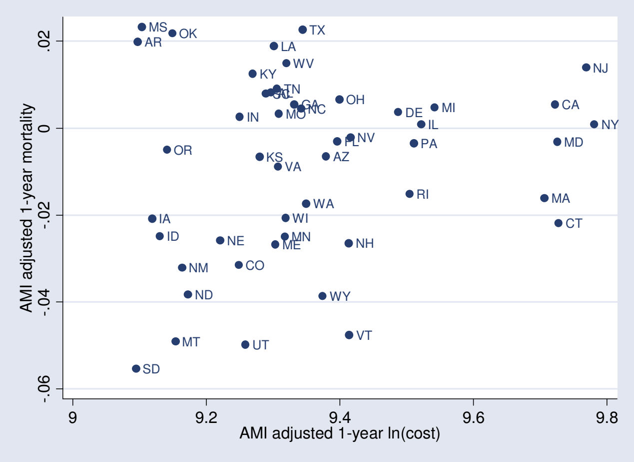
Outline

Motivation

- Stylized facts:
  - Large and growing spending in the US compared to other countries, not much better outcomes
  - Wide variation in spending and outcomes in the US, very little correlation
- Important questions in health economics:
  - Are we spending the right amount on health care?
  - Could we get more by spending the same amount?

Example

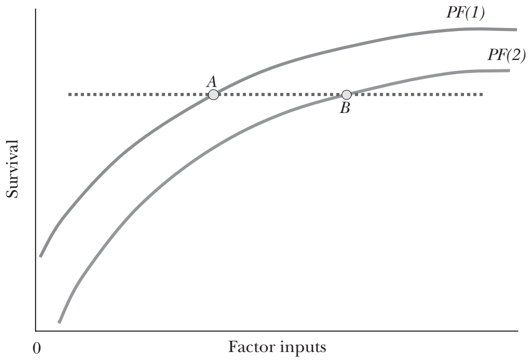
Heart attack mortality vs. spending across US states



Source: Skinner, Jonathan, and Douglas Staiger. “Technology Adoption From Hybrid Corn to Beta Blockers.” Working Paper. National Bureau of Economic Research, April 2005.

Allocative vs. Productive Efficiency

Figure 2b
Explaining "Flat of the Curve" Health Care Expenditures



Source: Garber, Alan M., and Jonathan Skinner. "Is American Health Care Uniquely Inefficient?" The Journal of Economic Perspectives 22, no. 4 (November 1, 2008): 27–50.

Outline

Chandra and Staiger (2007)

"Productivity Spillovers in Healthcare: Evidence from the Treatment of Heart Attacks"

- How should we rationalize the facts that (i) medical treatments seem to work (e.g., RCTs) and (ii) little correlation between spending and outcomes?
- Prominent explanation: "flat of the curve" medicine (i.e., overuse)
- Alternative explanation: productivity spillovers and Roy selection
  - Positive productivity spillovers: higher treatment rates *increase* benefits to treatment
  - Areas with more intensive treatment have worse outcomes for non-intensive treatment $\Rightarrow$ little difference in overall outcomes

Roy Model

- Rational selection into treatment
- Selection depends on production function

- Production function in turn depends on learning by doing, support services, physician sorting
  - All of these are interesting (more later)
  - At a high level, these factors may imply productivity spillovers

Basic Model

- Consider two mutually exclusive types of treatment: “intensive” ($D = 1$) or not ($D = 0$)
- Proportion of cases with intensive treatment is P
- Net utility from choosing intensive treatment as function of net benefits and net costs

$$\begin{aligned} U_D &= B_D + \lambda C_D \\ U &= U_1 - U_0 \\ &= f(X, P), \end{aligned}$$

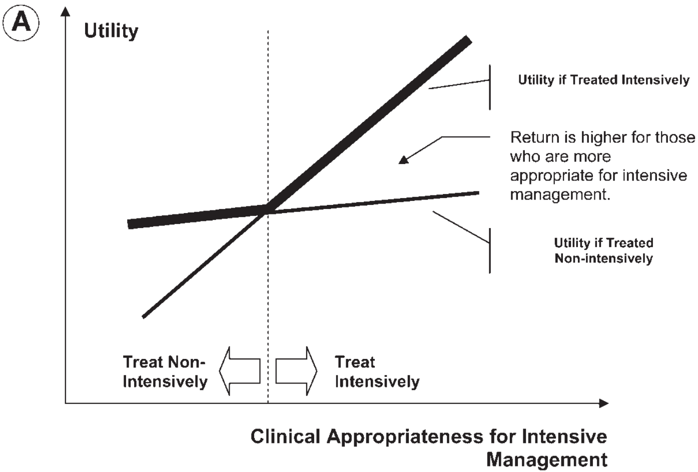
for patient characteristics X (assume no observable characteristics for now) and P.
  - Note: $f(X, P)$ embeds both productivity (B_0, C_0, B_1, C_1) and spending preferences λ

Basic Model

- Rationality: $D = \mathbf{1}(f(X, P) > 0)$
- Spillovers: $f(X, P)$ is increasing in P
  - Identical patients (same X) under same function f could be treated with different P
- Fixed point in equilibrium: $P = E[\mathbf{1}(f(X, P) > 0)] \equiv G(P)$
  - Equilibrium or equilibria depend on distribution of X and function $f(X, P)$, which jointly determine $G(P)$
  - Note: random assignment to areas + uniform productivity and spending preferences $\Rightarrow$ single $G(P)$

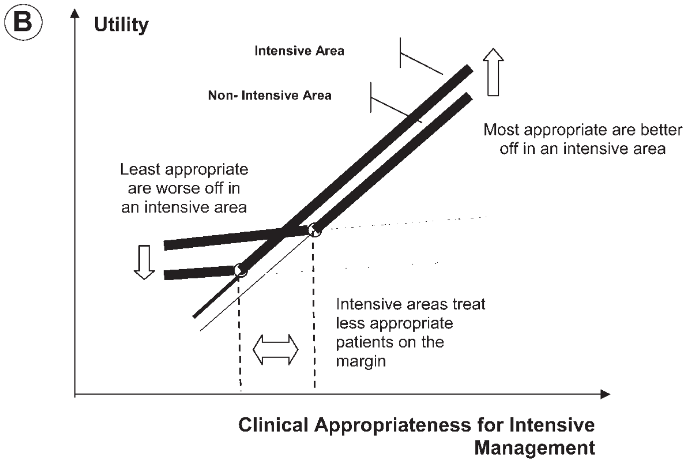
Graphical Selection Decision

Rationality: Patients who are more appropriate for intensive treatment will receive intensive treatment



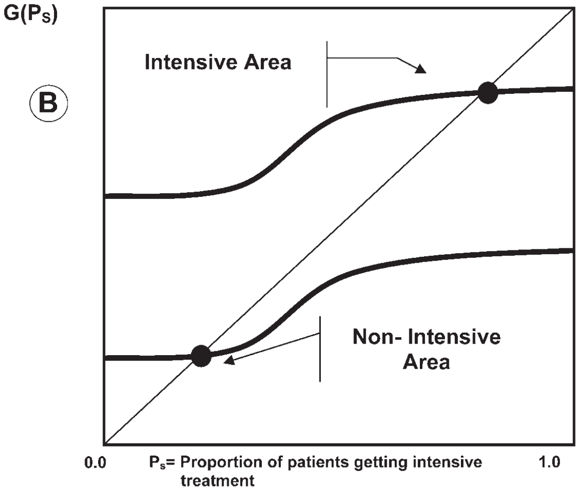
Graphical Selection Decision

Knowledge spillovers: Increasing *P* raises *U* for patients treated intensively, lowers *U* for patients not treated intensively



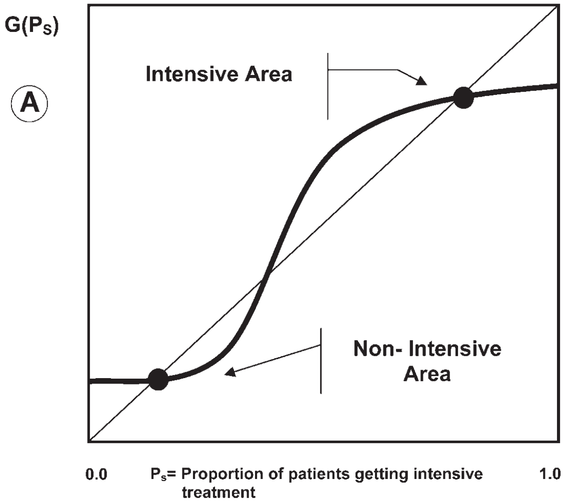
Possible Equilibria

Single Equilibrium, Different $G(P)$ Functions



Possible Equilibria

Multiple Equilibria, Single $G(P)$ Function



Use of Model

- Rationalize stylized facts:
  - Substantial differences in intensity (spending) across areas, not fully explained by patient characteristics
  - Spending not related to average outcomes
  - Large returns to receiving intensive intervention

Model Testable Predictions

1. Marginal patients receiving intensive treatment in intensive areas will be less appropriate than the average patient (rationality)
2. Utility associated with nonintensive treatment is worse in areas that are intensive (spillovers)
   - In intensive areas, utility will be higher for patients more appropriate for intensive treatment, lower for patients less appropriate
3. Net utility of intensive treatment should be higher among treated patients (treatment on treated) in intensive areas; gains highest for patients most likely to be treated intensively ($TOT > ATE$)

Distinguishing from Alternative Models

1. Differences in productivity of intensive treatments across areas, not necessarily related to spillovers
   - Probability of treatment for given patient should be unrelated to characteristics of *other* patients
   - Productivity between intensive and non-intensive treatment not necessarily negatively related
2. Flat of the curve medicine
   - High-intensity areas should have lower benefits and higher costs (lower net utility)
   - Also, predictions in #1 should hold, as no spillovers

Empirical Implementation

- Study case of AMI treatment: intensive treatment being catheterization or surgery; non-intensive being medical management (e.g., thrombolysis)
  - AMIs attractive for number of reasons (e.g., high mortality, common, little patient mobility, bimodal treatment choices)
- Study choice of AMI treatment according to area (Dartmouth HRR), estimate plausibly causal treatment effects within area on cost and survival
- Data from Cooperative Cardiovascular Project (CCP, collected by Medicare) on patient characteristics treated for AMI

Empirical Implementation

- Specification:
$$Outcome_{ij} = \beta_k Intensive_i + \mathbf{X}_i \Pi_k + \varepsilon_{ijk},$$
where k is some group of interest that is a function of i and j (e.g., high-risk patients, high-intensity HRR)
- β_k is object of interest, represents returns to intensive treatment in group k
- Estimated by using differential distance as IV for intensive treatment (c.f., McClellan et al, 1994)
- Predict appropriateness for intensive treatment based on patient characteristics, focus on differences between above- and below-median appropriateness

Results

First Stage, Reduced Form, Balance

TABLE 2

Relationship between Differential Distance (DD) and Probability of Catheterization and Survival, and Differential Distance and Observable Characteristics (%)

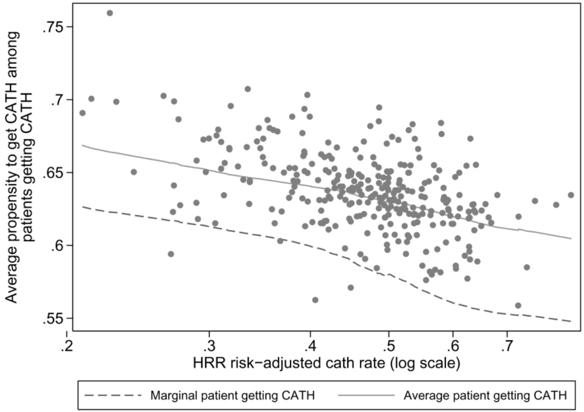
Sample	30-Day Cath Rate: DD Below Median (1)	30-Day Cath Rate: DD Above Median (2)	One-Year Survival: DD Below Median (3)	One-Year Survival: DD Above Median (4)	One-Year Predicted Survival: DD Below Median (5)	One-Year Predicted Survival: DD Above Median (6)	30-Day Predicted Cath Rate for Patients Getting Cath: DD Below Median (7)	30-Day Predicted Cath Rate for Patients Getting Cath: DD Above Median (8)
All patients (N = 129,997)	48.9	42.8	67.6	66.7	67.5	67.2	63.3	63.2
By cath propensity:								
Above the median (N = 64,733)	74.0	67.1	84.6	83.8	83.4	83.5	72.6	72.6
Below the median (N = 65,244)	22.9	19.5	50.1	50.4	51.1	51.6	32.3	32.5
By age:								
65–80 (N = 90,016)	61.1	54.9	74.3	73.5	73.9	73.9	67.4	67.3
Over 80 (N = 39,961)	20.3	16.5	52.1	52.1	52.6	52.7	34.6	34.1

Note.—Cath propensity is an empirical measure of patient appropriateness for intensive treatments. We define this measure by using fitted values from a logit model of the receipt of cardiac catheterization on all the CCP risk adjusters. Differential distance is measured as the distance between the patient's zip code of residence and the nearest catheterization hospital minus the distance to the nearest hospital.

- Implies average *positive* returns to treatment
  - Note: finding not new (McClellan et al 1994)

Results

Rational Selection



Spillover Results

TABLE 4
HRR-Level Measures of Intensive Treatment, Medical Management, Support of Medical Treatment, and Demographic Characteristics

HRR Indicator	Mean	Standard Deviation	10th Percentile	90th Percentile	Correlation with HRR Cath Rate
Measures of intensive treatment:					
Risk-adjusted 30-day cath rate	46.3%	9.1%	34.5%	58.3%	1.00
Risk-adjusted 30-day PTCA rate	17.7%	5.1%	11.3%	23.6%	.81
Risk-adjusted 30-day CABG rate	13.4%	2.9%	10.2%	16.9%	.51
Risk-adjusted 12-hour PTCA rate	2.7%	2.6%	.6%	5.8%	.52
Measures of quality of medical management:					
Risk-adjusted beta-blocker rate	45.6%	9.5%	34.2%	58.3%	−.31
Support for intensive treatment:					
Cardiovascular surgeons per 100,000	1.06	.27	.70	1.40	.33
Cath labs per 10,000	2.40	.76	1.50	3.30	.39
Demographic characteristics:					
Log of resident population	13.96	.89	12.72	15.18	−.05
Log of per capita income	9.55	.20	9.31	9.85	.02
Percent college graduates	19.3%	5.5%	13.1%	26.6%	−.05

Note.—HRR surgical and medical intensity rates are computed as the risk-adjusted fixed effects from a patient-level regression of the receipt of cath or beta-blockers on HRR fixed effects and CCP risk adjusters.

Spillover Results

TABLE 6
INSTRUMENTAL VARIABLE ESTIMATES OF INTENSIVE MANAGEMENT AND SPENDING ON SURVIVAL, BY SURGICAL INTENSITY OF HOSPITAL REFERRAL REGION

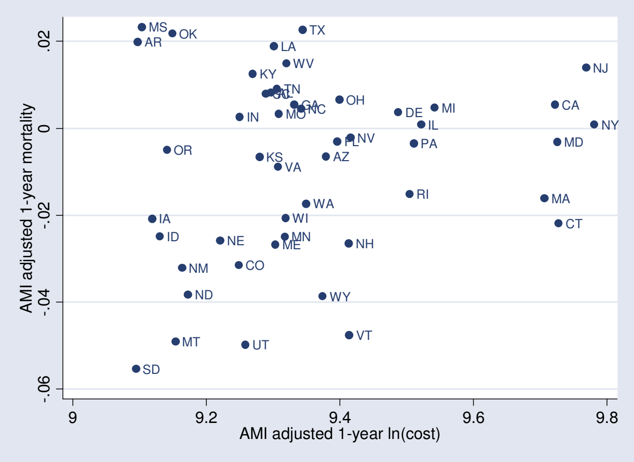
	INSTRUMENTAL VARIABLE ESTIMATES OF		
	Impact of Cath		
SAMPLE	On One-Year Survival (1)	On One-Year Cost ($1,000s) (2)	Impact of $1,000 on One-Year Survival (3)
A. All patients:			
HRR risk-adjusted cath rate:			
Above the median ($N =$ 63,771)	.256 (.061)	6.691 (3.510)	.038 (.021)
Below the median ($N =$ 66,124)	.09 (.059)	9.835 (3.155)	.009 (.007)
Difference	.166 (.085)	−3.144 (4.720)	.029 (.022)
B. Patients above the median cath propensity:			
HRR risk-adjusted cath rate:			
Above the median ($N =$ 32,388)	.271 (.064)	.347 (4.370)	.78 (9.820)
Below the median ($N =$ 32,411)	.168 (.046)	4.962 (2.890)	.034 (.021)
C. Patients below the median cath propensity:			
HRR risk-adjusted cath rate:			
Above the median ($N =$ 31,383)	.206 (.129)	16.21 (5.130)	.013 (.009)
Below the median ($N =$ 33,713)	−.139 (.165)	22.064 (6.870)	−.006 (.007)

NOTE.—HRR intensity rates are computed as the risk-adjusted fixed effects from a patient-level regression of the receipt of cath on HRR fixed effects and CCP risk adjusters. Differential distance (measured as the distance between the patient's zip code of residence and the nearest catheterization hospital minus the distance to the nearest hospital) is the instrument. Each model includes all the CCP risk adjusters, and the standard errors are clustered at the level of each HRR.

Interpretation

- Can reject flat of the curve hypothesis
- Strong evidence of productivity variation and some Roy selection [in this paper]
- Evidence of spillovers
  - Negative correlation between medical process measures (beta blockers) / outcomes and surgical intensity [in this paper]
- Possibly simpler story of productivity variation
  - Not necessarily with spillovers, nor with full Roy selection

How should this look like for “flat of the curve” hypothesis? For full Roy selection with single $G(P)$?



Source: Skinner, Jonathan, and Douglas Staiger. “Technology Adoption From Hybrid Corn to Beta Blockers.” Working Paper. National Bureau of Economic Research, April 2005.

Welfare Implications: Marginal vs. Average

- Spillovers $\Rightarrow$ externalities $\Rightarrow$ individual optimizing choices $\neq$ welfare optimizing choices
- Share of intensive treatment determined by marginal patient at which $U = 0$; welfare-optimal share depends on $E[U]$, averaged across all patients
  - Implies that welfare would be optimized by more extreme P (closer to 0 or 1) than in equilibrium; i.e., we should have *more variation*
- However, this depends:
  - $f(X, P)$ abstracts from moral hazard, other frictions
  - Reducing variation may still improve welfare under multiple equilibria
- Need to know microfoundations of production function (or $f(X, P)$)

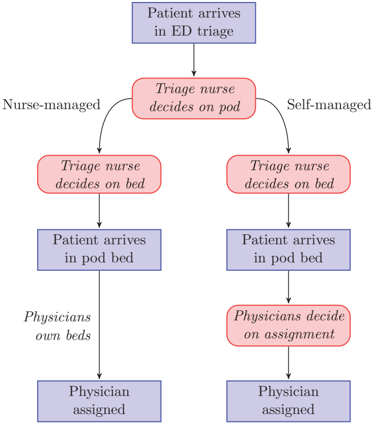
Outline

Moral Hazard

Chan (2016): "Teamwork and Moral Hazard: Evidence from the Emergency Department"

- Classically, **working in teams** $\Rightarrow$ moral hazard (Holmstrom 1982)
- Can **teamwork** reduce moral hazard (increase effort)?
- Empirical setting: Large ED with two systems of work assignment
  - "Nurse-managed system:" physicians directly assigned work by a triage nurse
  - "Self-managed system:" triage nurse assigns work to a team of physicians; physicians then decide who sees each patient

Empirical Setting



- Two locations ("pods"): Alpha always self-managed; Bravo changed from nurse-mangaged to self-managed system in March 2010

Overall Effect

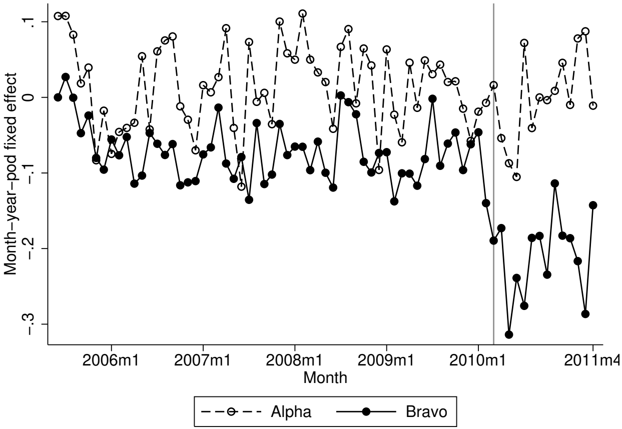
- Event study: for patient visit i,

$$Y_i = \sum_m \alpha_{m(i),p(i)} + \beta \mathbf{X}_i + \eta \mathbf{T}_i + \nu_{j(i)} + \varepsilon_i$$

  - Y_i: length of stay, orders, admissions, bouncebacks
  - pod $p(i)$, month-year $m(i)$
  - $\mathbf{X}_i$: patient characteristics
  - $\mathbf{T}_i$: time dummies
  - $j(i)$: physician-nurse team

- How do you think outcomes will differ with self-management?
- What would be the mechanisms?

Overall Effect



- 10-13% decrease in length of stay with self-managed system
- No effect on other outcomes: orders (~14 per visit), mortality (2% of sample), admissions (25% of sample), bouncebacks (8% of sample)

Mechanisms

- Potential mechanisms
  - Advantageous selection: better matching
  - “Free-riding”: waiting for peer to choose patients
  - Foot-dragging: pretending to be busy so won’t be assigned patient

- Free-riding and foot-dragging both forms of moral hazard on effort

Foot-dragging

How to empirically identify foot-dragging as a mechanism?

1. Suggestive: Advantageous selection would imply different outcomes (e.g., orders, costs, readmissions, mortality). Don't see this.
2. We can observe time to first order as measure of "free-riding". Don't see this.
3. Foot-dragging distinguished as response to *expected future work*
   - Temptation to foot-drag is greater when physicians expect more new work coming
   - No other mechanism should be affected by expectations
   - The flow of patients to the ED is exogenous to physicians

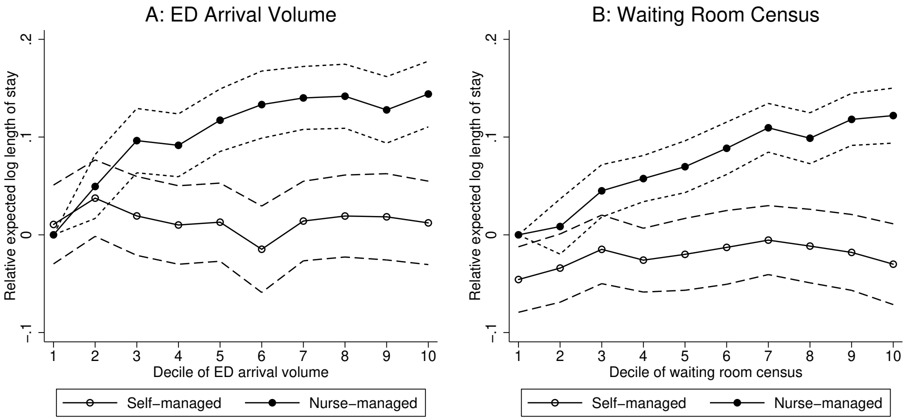
Foot-dragging

- Estimate effect of self-management interacted with expected future work (stock or flow of patients in waiting room) on length of stay:

$$
\begin{aligned}
Y_i \quad = \quad & \sum_{q=2}^{10} \alpha_0^d \left(1 - SELF_{p(i),t(i)}\right) D_q \left(EDWORK_{t(i)}\right) + \\
& \sum_{q=1}^{10} \alpha_1^q SELF_{p(i),} D_q \left(EDWORK_{t(i)}\right) + \\
& \beta \mathbf{X}_i + \eta \mathbf{T}_i + \zeta_{p(i)} + \nu_{j(i)} + \varepsilon_i
\end{aligned}
$$

  - $SELF_{p(i),t(i)}$: indicator for whether pod $p(i)$ is self-managed at $t(i)$
  - $EDWORK_{t(i)}$: measure of expected future work at $t(i)$

Foot-dragging



Joint Experience

Chen (2019): “Team-Specific Human Capital and Team Performance: Evidence from Doctors”

- What drives team performance?
  - Better match between team members, or
  - Team-specific human capital (requires experience working together)
- Broader economic concept: skills and knowledge on how to work together is a form of specific human capital

Empirical Setting

- Doctors treating heart attack patients in Medicare
  - Proceduralist: doctor who performs an intervention (PCI or CABG)
  - Physician: doctor who cares for the patient before and after the procedure
- Qualitative interviews: shared work experience $\Rightarrow$ communication, coordination, trust
- Identification strategies
  - Emergency cases quasi-randomly assigned to doctors on shift
  - [Two-way fixed effects for proceduralist and physician, interest in joint work experience between the two]

Specification and Identification

- For mortality y_i of patient i,

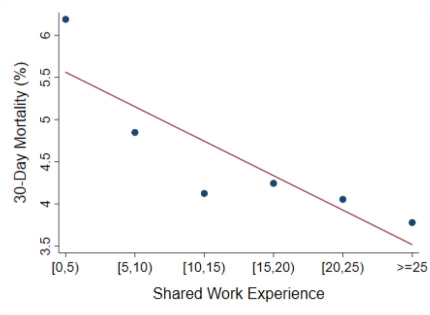
$$y_i = \alpha E_i + \theta_{k(i)} + \mathbf{T}_i \eta + \varepsilon_i,$$

  with shared work experience E_i, proceduralist fixed effect $\theta_{k(i)}$, and time dummies $\mathbf{T}_i$.
  - Also can control for patient characteristics $\mathbf{X}_i$, average characteristics of physician(s) $\mathbf{H}_{J(i)}$, proceduralist/physician individual work experience $\mathbf{F}_i$

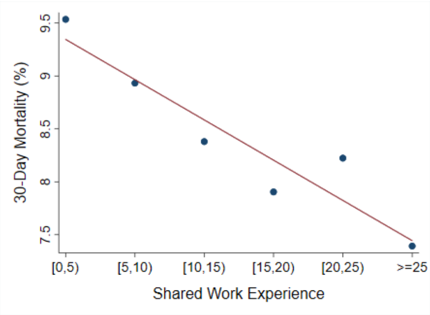
- Independence: patients quasi-randomly assigned to doctor teams
- Exclusion: effect is through shared work experience and not other characteristics (e.g., $\mathbf{H}_{J(i)}$ or $\mathbf{F}_i$)

Results

(a) PCI



(b) CABG



- 10-13% decrease in mortality with 1 s.d. increase in shared work experience (same magnitude as 1 s.d. increase in spending from Doyle et al 2015)
- Flat relationship between shared work experience and predicted mortality

Mechanisms

Production function, for physician j and proceduralist k

$$y_{jk}(e) = A_{jk}(e) + M_{jk}$$

1. Returns to shared work experience: $A_{jk}(e)$
   - Past collaboration experience e improves performance
2. Match quality M_{jk}
   - Better-matched doctors more likely to work together and to have better outcomes (invariant of e)
   - Limited institutional (e.g., shifts) or empirical evidence (match effects model from Card et al 2013) for this

Mechanisms

How does shared work experience improve outcomes?

$$A_{jk}(e) = a_{jk}(e) \cdot f(I_{jk}(e))$$

1. Improved productivity $a_{jk}(e)$
   - Proceduralists and physicians learn to work more efficiently (i.e., better outcomes with same inputs) $\Rightarrow$ welfare improvements
2. Increased inputs $I_{jk}(e)$
   - They increase inputs instead $\Rightarrow$ ambiguous welfare implications

- Inputs generally *decrease* with shared work experience

Adoption

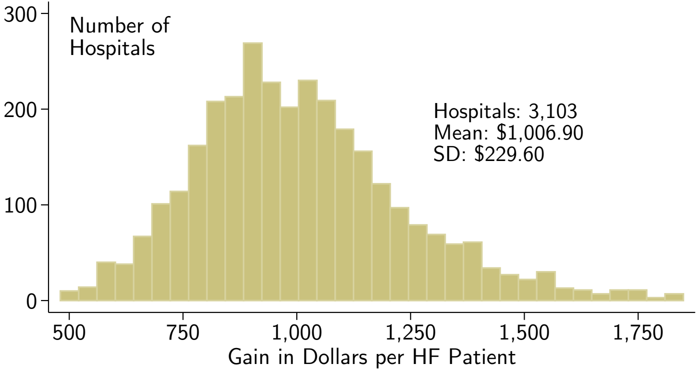
Sacarny (2018): "Adoption and Learning Across Hospitals: The Case of a Revenue-Generating Practice"

- A simple "revenue-generating practice"
  - In 2008, Medicare paid more for physicians to specify the type of heart failure in claims
  - If left unspecified, would be equivalent to leaving money on the table
- Yet hospitals only captured about half of this revenue

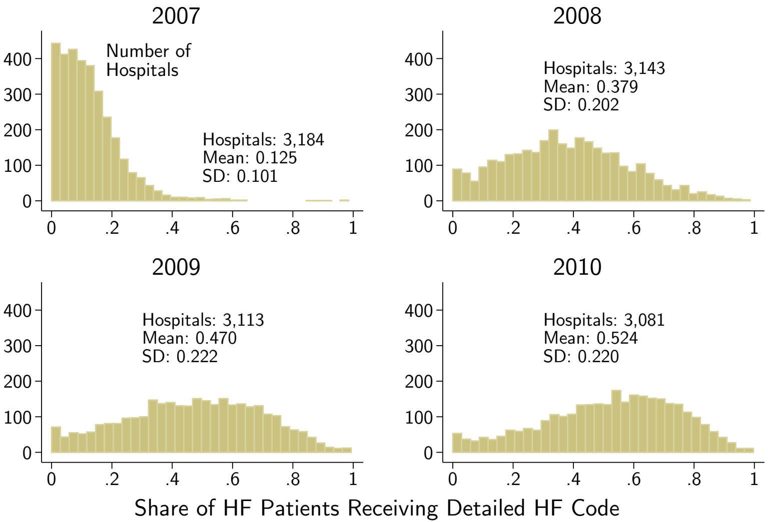
Adoption of Coding Practice Over Time

Figure plots the share of revenue available for detailed coding of HF that was captured by hospitals over time. Dotted line shows revenue that would have been captured in 2007 if hospitals had been paid per 2008 rules. See Appendix Section A.1.2 for more details.

Revenue at Stake per Heart Failure Patient



Hospital Adoption of Coding Practices Over Time



A hospital's adoption equals the share of its HF patients who received a detailed HF code in that year. Hospitals with fewer than 50 HF patients in the year excluded.

Hospital Determinants

- Using physician movers, about 80% of difference in adoption is due to hospitals
  - Reflects hospitals' ability to extract documentation from physicians
- Hospitals that extract more revenue ...
  - have better clinical outcomes
  - are more likely to be vertically integrated
  - have better management scores (Bloom)
- Similar variation in adoption across multiple settings

Outline

Identifying Productivity

Chan, Gentzkow, and Yu (2019): "Selection with Variation in Diagnostic Skills: Evidence from Radiologists"

- Can we separately identify production functions using cross-sectional data?
- Restrictions in how production functions can look like
- Simple setting of diagnostic productivity

Classification Problem

	Actual Positive	Actual Negative
Classified Positive	True Positive (TP)	False Positive (FP) Type I Error
Classified Negative	False Negative (FN) Type II Error	True Negative (TN)

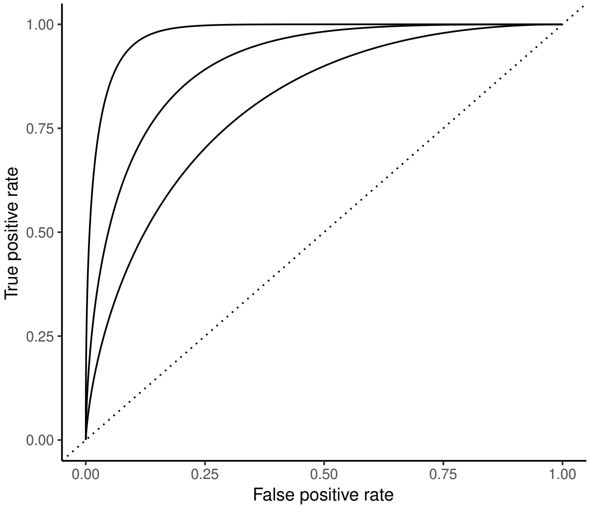
True Positive Rate

$$TPR = \frac{TP}{TP+FN}$$

False Positive Rate

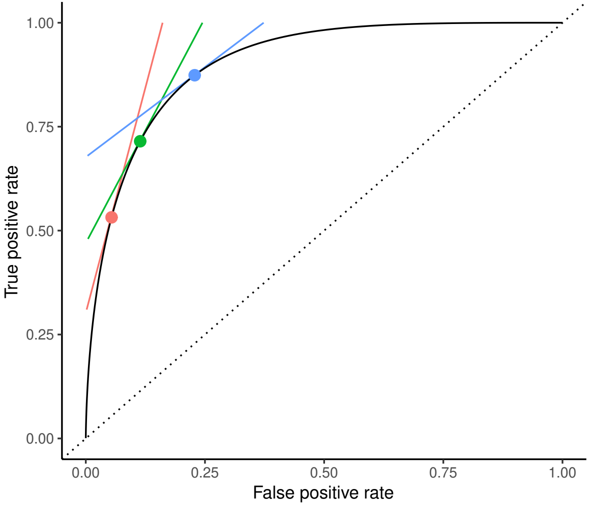
$$FPR = \frac{FP}{FP+TN}$$

Receiver Operating Characteristic (ROC) Curves



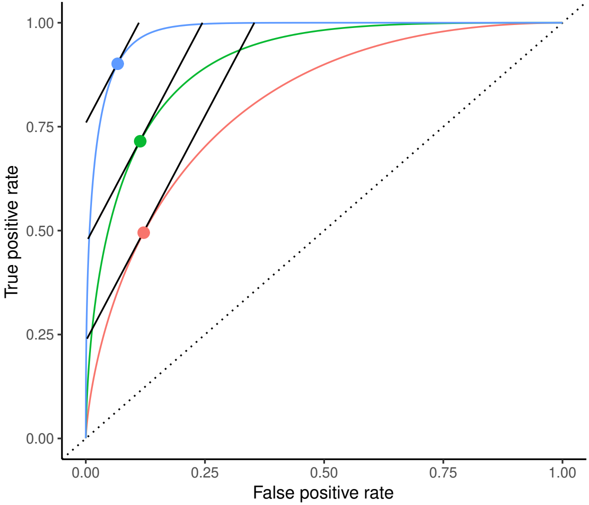
Standard Model

Fixed skill = standard monotonicity; variation in preferences $\Leftrightarrow$ variation in thresholds



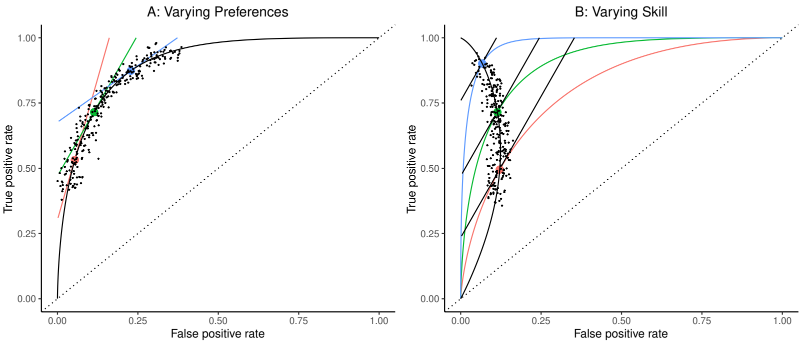
Alternative Model

Variation in skill; fixed preferences $\Rightarrow$ thresholds depend on skill and preferences



Distinguishing between Models

- Assume homogeneous populations between providers, observed conditions and classifications



- Preference vs. skill variation imply different patterns of data
  - Concepts: productivity = skill, allocative efficiency = preferences

Setting: Chest X-Rays for Pneumonia

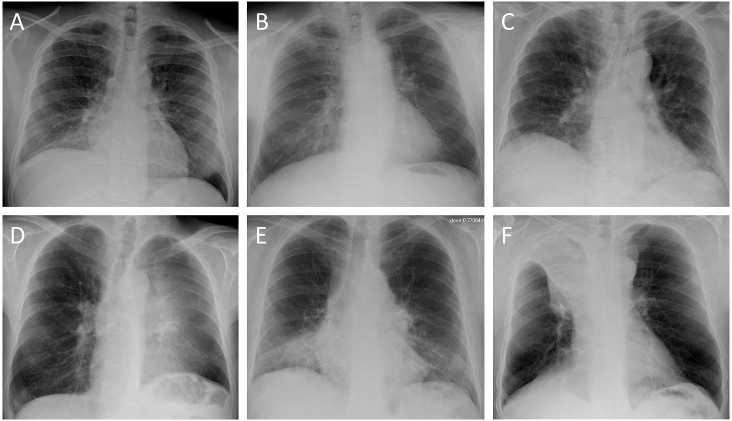


Figure 2. Typical examples of radiographs expected to mobilize detection skills (A—C) and interpretation skills (D—F). Experts' consensus diagnoses were: miliary tuberculosis — CXR#6 (A), lung nodule (cancer) in left upper lobe — CXR#19 (B), usual interstitial pneumonia — CXR#27 (C), left upper lobe atelectasis — CXR#3 (D), right lower lobe infectious pneumonia — CXR#14 (E) and right upper lobe atelectasis with Golden sign — CXR#36 (F).

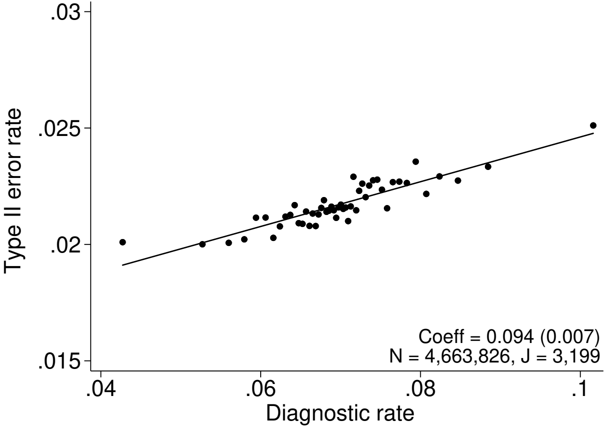
Source: Fabre, C., et al. "Radiology Residents' Skill Level in Chest x-Ray Reading." Diagnostic and Interventional Imaging 99, no. 6 (June 1, 2018): 361–70.

Mapping Between Spaces

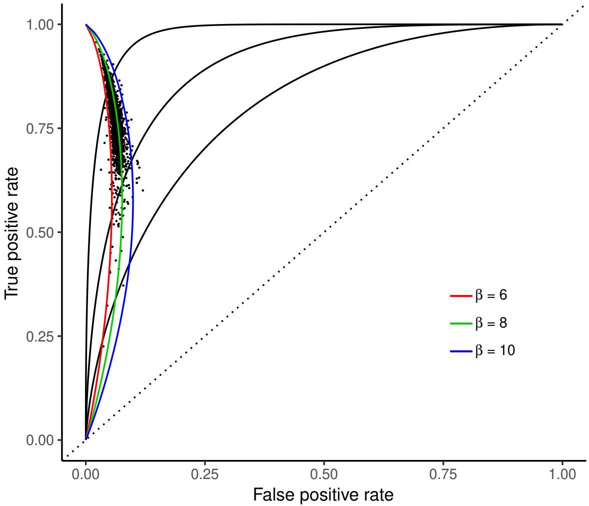
- One-sided selection:
  - Cannot distinguish true positives from false positives; only can observe diagnosis rate P_j, type II error rate FN_j for each radiologist j
  - Observable data in **"reduced-form space"** (P_j, FN_j)
- Random-assignment: prevalence of pneumonia S same for all j
  - One-to-one correspondence with **ROC space** (FPR_j, TPR_j) if S is known
- Upward sloping ROC curve $\Leftrightarrow$ for two agents j and j' with equal skill, $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} \in [-1, 0]$

Reduced-Form Results

Relationship between FN_j and P_j



Structural Results in ROC Space



Implications

- Variation in radiologist skill is large and drives a 55% of variation in decision
- Policy counterfactuals
  - Imposing uniform thresholds slightly *reduces* welfare
  - Small improvements in “skill” may substantially improve welfare (e.g., combining signals improves welfare by 33% to first best)
- Example of shape restrictions on production functions as identification

Outline

Summary

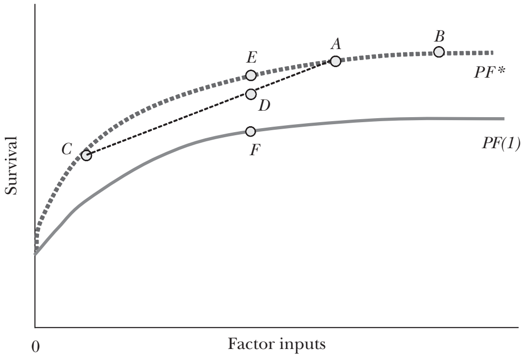
- Production functions at different levels
  - Chandra and Staiger (2007) at the area level
  - Determined by production at the hospital, provider, and team levels
- Preferences or allocative inefficiency at micro-levels could contribute to productive inefficiency at macro-levels

Simple Example

Within-area variation and Jensen's inequality

Figure 2a

A Health Care Production Function



Source: Garber, Alan M., and Jonathan Skinner. "Is American Health Care Uniquely Inefficient?" The Journal of Economic Perspectives 22, no. 4 (November 1, 2008): 27–50.

Future Directions

1. Shed light on informational component and related frictions
   - Chandra and Staiger (2007) is a seminal benchmark embedding productivity variation, but is very high level
   - Need to better understand frictions
2. Exploit multiple sources of variation
   - Need variation both across and within production functions
   - Can rely on rich institutional setting in health care
   - Can also imagine designing interventions
3. Assess interactions between various agents
   - Health care is very rich