

Discussion of  
“Learning and Efficiency in the Market  
for Physician Referrals”

Ian McCarthy & Seth Richard-Shubik

Discussant:

Andrew Ching

Johns Hopkins Carey Business School

# Research Questions

- How did primary care physicians (PCPs) learn about the quality of orthopedic surgeons?
- How did learning shape the allocation of patients to specialists via referrals?
- Setting: major joint replacement surgery among Medicare patients, using Part A/B claims from 2008–2018
  - final sample of just over 2 million procedures.

- The empirical exercises:
  - (i) a descriptive analysis of within-market heterogeneity in specialist quality;
  - (ii) a stacked DID event study showing that PCPs reduce referrals to a given specialist by about 16% following a bad outcome.
    - Use PCPs share the same specialist without bad outcome as control.
  - (iii) estimating a structural Bayesian learning model with beta-distributed beliefs about each specialist's quality,
    - allowing for **familiarity** effect and **congestion** effect.
    - updating those beliefs via observed patient outcomes – either success or failure.
- The authors argue that they provide identification argument which is more formal than the prior literature.

# Main Findings

- PCPs place little weight on patient outcomes (*alpha*), roughly 3 miles of driving distance on average, with many HRRevel estimates at or near zero;
- Counterfactual simulations that assume PCPs with perfect information about specialist quality, while preserving “existing familiarity”, lead to reallocation of about 23% of referrals and a small positive improvement in patient outcomes
  - 0.07 fewer failures per 100 patients, an approximately 0.8% relative reduction against the 9-per-100 baseline;
  - so those changes don’t lead to better referrals.
- When the counterfactual instead resets familiarity to zero, reallocation increases to about 32%, yet average health effects turn slightly negative.
- For the last counterfactual experiment finding, the authors interpret existing referral relationships themselves contain valuable quality information (another point I will discuss more below).

# Overall Take

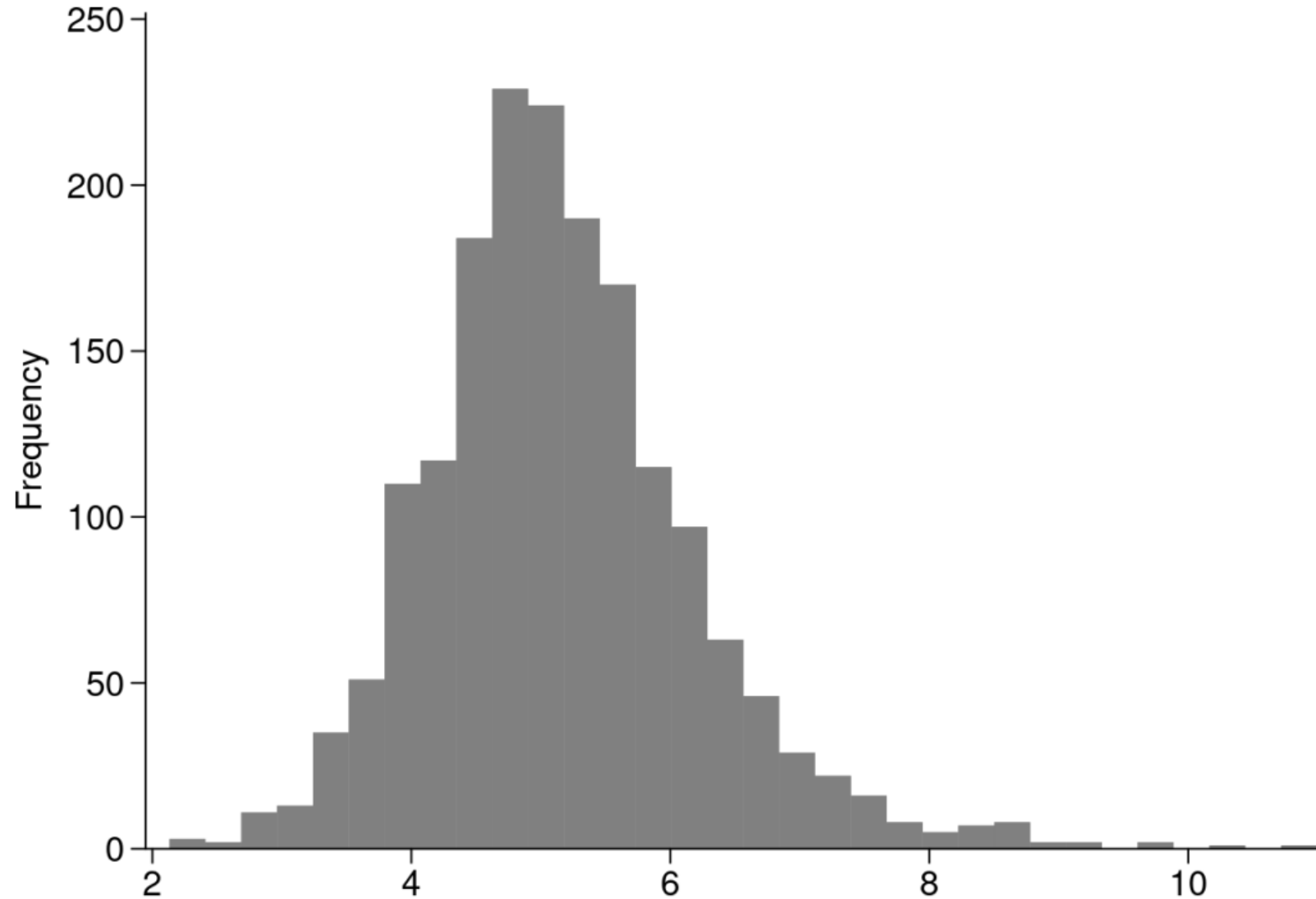
- This is a very serious and ambitious piece of research.
- The dataset is exceptionally rich and clearly required extensive cleaning.
- The authors provide thorough appendix analyses to support their empirical findings and anticipate many potential reader concerns.
- The paper is extremely carefully done!

# Comment 1

- I disagree with the authors' characterization that Ching, Erdem, and Keane (2013) offer only a “verbal” identification argument while their own presentation is formal.
- In a standard Bayesian framework discussed in Ching et al. (2013), as  $\# \text{ experience signals} \rightarrow \infty$ , the posterior variance approaches zero in the limit via standard Bayesian updating formulas.
- This step is mathematically straightforward and well understood in the literature.

# Heterogeneity in quality and cost

Interquartile ranges within markets of negative outcomes and episode spending



## 2. Surgeon Quality Heterogeneity vs. Patient Heterogeneity

- The authors interpret cross-surgeon dispersion in failure rates as direct evidence of quality heterogeneity.
- However, this dispersion could also reflect unobserved variation in patient severity (e.g., sicker patients selecting higher-quality surgeons).
- The authors recognize this issue and restrict the sample to elective, planned joint replacements, showing that 86.1% of their sample falls under a single standardized code: DRG 470 (assuming that this would make the patients pool more homogeneous in severity). Yet, there are two caveats:
  - 1. The remaining ~14% of non-DRG 470 cases may account for a disproportionate share of the observed dispersion in failure rates across markets.
- **\*\* I would be interested in seeing by just looking at DRG470 procedure, the distribution of success rates across physicians. \*\***

# DRG470 Caveat 2

- I asked Google “Is there a lot of heterogeneity in complications in this procedure?”
- “... there is a massive amount of clinical and statistical heterogeneity in complications under MS-DRG 470. Because this group is defined strictly by the *absence* of a Major Complication or Comorbidity (MCC), it serves as a clinical "catch-all" for highly diverse patient profiles and procedures.”
- **“The "Healthy" Outlier vs. The Comorbid Patient:** A healthy 65-year-old with isolated osteoarthritis has a radically different complication risk compared to an 82-year-old with mild chronic kidney disease, controlled diabetes, and moderate hypertension. Yet, because none of those conditions cross the bureaucratic threshold into an MCC, both surgeries are billed under DRG 470. [[1](#), [2](#), [3](#), [4](#), [5](#)]”
- **“Hidden Risk Profiles:** [Independent research on orthopaedic outcomes](#) indicates that many underlying conditions—like electrolyte disorders or neurodegenerative issues—drastically increase a patient's true risk and actual cost of care, yet fail to move the patient into the higher-paying DRG 469 (With MCC). [[1](#)]”
- **\*\* This suggest the paper should add noise to the signals. \*\***

# How to model noisy signal

- If sticking with Beta distribution belief, one can:
  - Fractional updating:  $a_{t+1} = a_t + w \cdot S_t$ , where  $w \in [0,1]$ .
  - Explicitly modeling classification error in success/failure outcome.
  - Add a parameter to scale up the pseudo-prior counts of the sample size in creating the initial prior,  $\eta = a_0 + b_0$ .
- Or, normal prior and normal noisy signals approach (Ching, Erdem and Keane, 2013 & 2017).

## Model specification

1. Each period  $t$ , the PCP  $i$  sends a patient (also  $t$ ) to a specialist  $j$  from a fixed choice set  $J$  (choice indicators:  $D_{ijt}$ )
2. Specialist treats the patient
  - binary outcomes:  $Y_{ijt} = 1$  (success) or 0 (failure)
  - specialist quality:  $q_j \equiv \Pr(Y_{ijt} = 1)$ , *unknown to PCP*
3. PCP's realized utility:

$$U_{ijt} \equiv \alpha Y_{ijt} + u(x_{ijt}) + f(e_{ijt}) + c(n_{jt}, z_j) + \xi_j + \epsilon_{ijt},$$

- $\alpha$  – weight on patient outcomes (e.g., altruism)
- $u(x_{ijt})$  – patient-specific factors (e.g., distance)
- $f(e_{ijt})$  – prior experience with specialist ( $\#$  patients, “familiarity”)
- $c(n_{jt}, z_j)$  – specialist current patient volume (congestion effect)
- $\xi_j$  – other demand factors (unobserved to econometrician)
- $\epsilon_{ijt}$  – idiosyncratic shock

### 3. Familiarity vs. Risk Aversion

- The high concentration of PCP referrals to specific specialists could reflect physician risk aversion toward surgeons with whom they have little or no experiences.
- A substantial body of structural learning literature incorporates risk aversion into the utility specification, which naturally generates a preference for familiar options (e.g., Erdem and Keane, 1996; Ching, 2010).
- The authors assume risk neutrality in the experience part, while incorporating a term to capture familiarity. But the estimated familiarity effect may just reflect risk aversion.
- The authors can consider a quadratic term for  $Y$  in their utility function. The coefficient for the quadratic term can be interpreted as risk aversion parameter.
  - The posterior variance will show up in the expected utility, qualitatively works like the familiarity term.
  - They might not need their familiarity term to explain the data.
  - Or, the impt of the familiarity term should drop.

## 4. Counterfactuals

- Counterfactual Expt 1. If the familiarity term simply represents the posterior variance & risk aversion, a counterfactual expt of setting everyone to have perfect info should also set the posterior variance to zero.
  - That would set the familiarity term to the limit.
- Counterfactual Expt 2. Setting familiarity to zero (no experience) should correspond to initial prior (not perfect information).

## 5. Event Study vs. Structural Model Findings

- The event study finds a 16% relative reduction in referrals following a bad outcome; the structural model estimates a mean partial effect of one additional failure of  $-0.32$  percentage points among “responsive” HRRs.
- In the event study, the authors restrict their attention to cases of overlapping networks, where multiple PCPs refer patients to the exact same specialist in a given quarter, and where at least one (but not all) of those patients experiences a surgical failure.
- In the structural model, some specialists have very low baseline referral probability (close to zero) creating a “floor” that dampens the negative outcome effect on choice behavior.

## 6. Bayesian or Other Forms of Learning (Future Research)

- The strong effect identified from the event study is surprising via the lens of Bayesian learning.
- If a PCP has already accumulated a large number of +ve experiences (success) with one specialist, just one failure should not change its posterior much at all.
- The fact that the relative reduction is 16% may suggest that the PCP pays a lot more attention to their recent experiences when updating their belief. This suggests that one should consider alternative learning models. One class of models that come to my mind is adaptive learning (see, e.g., Dorazelski et al. (2018), Li and Ching (2024)).
- There are many plausible reasons why PCP may adaptively learn (e.g., just pay attention to the most recent three months experiences if they refer patients frequently): surgeons' skills may improve due to learning-by-doing or their skills might decline as they age or due to some other factors that could -vely affect their performance persistently (e.g., burnout or emotional stress).

# Surprise Effect on Learning

- In Ching, Hossain, Tehrani and Zhao (2021, a new version coming), we run lab experiments to study how people update their beliefs.
- We elicit beliefs using an incentive-compatible payment mechanism, and study how subjects update their beliefs sequentially over a series of noisy signals.
- We find subjects dramatically lost confidence about their prior belief (posterior var. jumps up) when receiving a surprise signal.
- If they are Bayesian, this will never happen because posterior variance can only decrease in the number of signals received in Bayesian.

## 6. Learning-by-doing or Burn-out? (Future Research)

- Following the idea of PCP may adaptively learn, it would be very interesting and valuable if the authors can empirically investigate if surgeon's skills evolve over time.
- Given their long panel (2008-2018), this should be possible.