User-Generated Physician Ratings: Evidence from Yelp

Yiwei Chen
10.22.2018
Why study online user-generated ratings in healthcare?

- Online user-generated ratings have been increasingly popular in healthcare
  - Surveys find 54% internet users use online physician ratings
  - # Online reviews for physicians grow rapidly. E.g., on Yelp,
Online ratings can potentially improve healthcare efficiency
- Consumers have little information on which to base physician choice decisions
- Online ratings may bring patients into better physicians and promote quality

However, they face challenges in healthcare to deliver the promise
- Users are better at evaluating patient centeredness than clinical effectiveness. Unclear what they inform to readers who value clinical quality
- Whether ratings actually affect patients’ physician choices
- Physicians may prescribe harmful treatments (opioids) to please patients

Why study online user-generated ratings in healthcare?
Research questions

- Use Yelp ratings and Medicare claim data to study these challenges

1. What are the contents of Yelp physician reviews and do they correlate with clinical quality?
   - Reviews primarily describe physicians’ interpersonal skills and amenities
   - Ratings are positively correlated with many measures of clinical quality

2. Do ratings affect patients’ physician choices?
   - A one-star higher physician average rating improves physician revenue and patient volume by 1-2% statistically significantly

3. Do physicians change practice behaviors after being rated?
   - Suggestively, physicians slightly increase lab & imaging tests, but no statistically significant evidence shows they increase opioid prescriptions
Broadly relates to how rating mechanisms affect consumers and suppliers, e.g., in public hygiene, education, consumer goods, restaurants, etc.

- Jin & Leslie (2003); Dewan et al. (2004); Rockoff et al (2012); Luca (WP)...
- Use a novel IV design utilizing reviewers’ “harshness” in rating other businesses to instrument for their physician ratings

Outcome-based health provider report cards have not elicited a large consumer response and not always had good impact on physician behavior

- Dranove et al. (2003); Bundorf et al. (2009); Kolstad&Chernew (2009); Kolstad (2013)...
- This paper extends to online physician ratings as a new format of health report cards

Young literature finds online ratings positively correlated with clinical quality

- Bardach et al. (2013); Ranard et al. (2016); Howard et al. (2016); Lu and Rui (2017)...
- This paper uses the universe of Yelp ratings for individual physicians VS small and specialized provider sample in existing literature
Data sources

• U.S. nationwide Yelp online “Doctor” ratings until June 2017 at an individual review level
  ▪ Yelp was the most used online physician rating website surveyed in 2014
  ▪ +0.36 correlations in ratings with second largest, Healthgrades.com

• Medicare database
  ▪ Annual physician level 100% payment data (2012-2015)
  ▪ Claim level data for 20% Medicare enrollees (2008-2015)
  ▪ Internet surveys found the elderly among the highest usage age groups of online physician ratings

• External data for medical credentials
1. Dolhun Clinic
   - 40 reviews
   - Family Practice, Internal Medicine, Home Health Care

   Dr. Dolhun was just the best Doctor that has ever taken care of me. He was really attentive, and easy to talk to. Some other doctors I've dealt with didn't care at all about me or... read more

2. The House Doctor
   - 85 reviews
   - Family Practice, Urgent Care, Internal Medicine

   Dr. Brian Hamway is an excellent, competent, and pleasant doctor, who has inspired our trust and confidence. Al (my partner) had been very weak, pale, and lacking energy. Being able... read more

3. David Shu, MD
   - 87 reviews
   - Internal Medicine

   Dr. David Shu is the best DOCTOR in San Francisco!! He's friendly, knowledgable and trustworthy. I really trust him with my life. He's an Internal medicine doctor on Ocean Ave so... read more

4. Edwin J Hassid, MD
   - 32 reviews
   - Internal Medicine

   could be the best doctor in San Francisco. I'm really grateful that I found Dr. Hassid to help me though this horrible experience. I guess this is my way of saying thanks. (No... read more

5. Arnold Lee, MD
   - 34 reviews
   - Internal Medicine, Family Practice

   Financial District
   San Francisco, CA 94111
Matching Yelp ratings with Medicare

- From Yelp, 638,489 historical reviews for 107,279 listings under “Doctor” are scraped until June 2017

- From Medicare, the payment data consists on avg 972k clinicians between 2012-2015

- Match Yelp with Medicare through matching Yelp physicians’ last name, first name, and Health Service Areas (HSA) with physician NPI directory

- 36,787 physicians are matched between Yelp and NPI directory
  - Only individual physician listings are matched
  - 70% of individual listings from Yelp profiles are uniquely matched
  - Most in small groups and in primary care and face-to-face specialties such as family medicine, internal medicine, dermatology, etc.
Yelp reviewers are usually “opinionated”

Avg # reviews per listing: 5.3
Avg rating at a listing level for physicians: 3.62
Avg rating at a review level for physicians: 3.79
Avg rating at a review level for all Yelp businesses: 3.71
1. What quality information do ratings convey?
A Yelp review may contain service information, clinical information, etc.

“*She encouraged him to exercise and lose weight which resulted in his much improved cholesterol ratio and energy level.*”

“What irked me slightly was that I did not get a reminder call about my appointment.”

Need methods to aggregate 200k+ reviews into categories
In a machine learning (LDA) model, a review is considered as a set of “topics”

Each topic is a cluster of words that tend to co-occur in a review
- E.g. “exercise, weight, energy,...”
- E.g. “reminders, appointment, ...”

The algorithm reads in reviews, generates possible topics, and classifies reviews into topics

I find the most frequent topics as describing attitude, interpersonal skills, amenities, etc.
Staff was not able to reach the patient. They scheduled a follow-up appointment for later in the week. They said they would call with the results of the test performed on Tuesday. They mentioned that the medication was running out and asked if the patient would like a refill. The patient requested a refill and was told to call the pharmacy. The patient was frustrated and wanted to know why they could not find the scheduled appointment in their records. They tried to schedule the appointment for the next available time, but the system was not showing anything. The patient asked if they could reschedule for a later date and the staff said they would try to accommodate them. The patient asked if they could call them if there were any updates and the staff said they would let them know. The patient was disappointed and decided to call the doctor's office directly.
A survey approach

• 1500 reviews are sent to human readers to be classified into “service quality related”, “clinical related”, “both”, or “neither”
  ▪ Service quality information, e.g. on attitude, friendliness, patience, amenity, billing, waiting, etc.
  ▪ Clinical quality information, e.g. on diagnosis, treatment, prescription, recovery, health outcome, etc.

• 81% of reviews are considered to contain “service” information and 44% contain “clinical” information
Yelp Ratings (+) correlate with various physician “quality”

• Are Yelp ratings positively correlated with physician clinical quality?

• Correlate ratings with physician medical credentials among all the rated physicians at a physician $j$ level:

$$y_j = \beta R_j^{2017} + HSA_j + Specialty_j + \epsilon_j$$

$y_j$: Physician $j$’s credential including board certifications, med school rankings, and #self-reported accreditations

$R_j^{2017}$: the latest cumulative average rating (2017) for physician $j$
Higher Yelp ratings associated with better medical credentials

\[ y_j = \beta R_j^{2017} + HSA_j + \text{Specialty}_j + \epsilon_j \]

<table>
<thead>
<tr>
<th>RHS</th>
<th>LHS: Measurement of Physician Clinical Ability/Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(Board Certification)</td>
<td>Medical School Ranking</td>
</tr>
<tr>
<td>(\beta): Ratings</td>
<td>.0299*** (.00502)</td>
</tr>
<tr>
<td>Implications of 1-&gt;5 stars</td>
<td>+.12 in probability</td>
</tr>
<tr>
<td>N</td>
<td>8,755</td>
</tr>
<tr>
<td>Mean of LHS</td>
<td>.73</td>
</tr>
<tr>
<td>Sample</td>
<td>Healthgrades-Yelp (primary care physician only)</td>
</tr>
</tbody>
</table>

Standard errors two-way clustered at HSA and Specialty levels
Correlations between ratings and primary care health outcomes

- Higher ratings corr w/ better outcome-based clinical quality measures?
- Link patient $i$ in year $t$ to most frequently visited primary care physician $j$
- Estimate a patient($i$)-year($t$) level regression among all patients of rated primary care physicians

$$y_{it} = \beta R_{j(it)}^{2017} + X_{it}'\gamma + \epsilon_{it}$$

$y_{it}$: patient $i$’s health outcome in year $t$
$R_{j(it)}^{2017}$: ratings of $j$ in 2017, who is patient $i$’s primary care physician in year $t$
$X_{it}$: patient characteristics including past year risk scores, patient demographics, location FE, year FE
Higher ratings correlated with better patient outcome and practices

<table>
<thead>
<tr>
<th>RHS</th>
<th>LHS: Measurement of Patient Health Outcomes and Medical Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1(Eye Exam for Diabetics)(^1)</td>
</tr>
<tr>
<td>β: Ratings</td>
<td>.00195(^{**}) (0.000787)</td>
</tr>
<tr>
<td>1-&gt;5 stars</td>
<td>+0.08 (+1.5%)</td>
</tr>
<tr>
<td>N</td>
<td>810,464</td>
</tr>
<tr>
<td>Mean of LHS</td>
<td>.52</td>
</tr>
</tbody>
</table>

Standard errors clustered at HSA levels

1 recommended procedure based screening criteria according to clinical
2 preventable inpatient conditions according to numerators from PQI index developed by AHRQ
1. What quality information do ratings convey?

2. Do ratings affect patients’ physician choices?
Does a higher Yelp rating bring more patient flow to a physician than a lower rating?

Ideal experiment: If randomly assigning Yelp average ratings to a physician, do physicians receiving higher ratings grow faster in patient flow than those receiving lower ones?
Estimation framework

- Estimate a physician $j$ year $t$ regression among all physicians from Medicare Part B payment data 2012-2015:

$$y_{jt} = \chi_j + \theta_{t,h} + \theta_{t,s} + \lambda D_{jt} + \beta R_{jt} D_{jt} + \epsilon_{jt}$$

$y_{jt}$: Physician $j$’s revenue and patient volume in year $t$

$\chi_j$: Physician fixed effects

$\theta_t$: Year fixed effects. HSA and specialty specific

$D_{jt}$: indicator of 1 since physician $j$’s first rating year (physician $j$ has a rating)

$R_{jt}$: Cumulative average rating of physician $j$ by year $t$, de-meaned to mean 0
\[ y_{jt} = \chi_j + \theta_{t,h} + \theta_{t,s} + \lambda D_{jt} + \beta R_{jt} D_{jt} + \epsilon_{jt} \]

- \( \lambda \), the common aggregate factor between before and after rating, is identified using physicians who change from being non-rated to rated.

- \( \beta \), the differential effects of ratings, is identified from two sources:
  - If ratings are fixed, whether physicians with higher ratings grow faster in patient flow than those with lower ratings, compared to before rating.
  - For physicians whose ratings evolve over time, how different levels \( R_{jt} \) are associated with her patient flow differently.
Data sample: Yelp data merged with 100% Medicare FFS payment data 2012-2015, including all physicians rated or never rated. Two year fixed effects are HSA(j) and specialty(j) specific. Standard errors two way clustered at HSA and specialty levels.

\[ y_{jt} = \chi_j + \theta_{t,h} + \theta_{t,s} + \lambda D_{jt} + \beta \cdot R_{jt} \cdot D_{jt} + \epsilon_{jt} \]

<table>
<thead>
<tr>
<th>RHS\LHS</th>
<th>Log Total Revenue</th>
<th>Log Unique Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>OLS</td>
<td>OLS</td>
</tr>
<tr>
<td>[ \lambda: D_{jt} ]</td>
<td>-.0125**</td>
<td>-.00786*</td>
</tr>
<tr>
<td></td>
<td>(.00559)</td>
<td>(.00423)</td>
</tr>
<tr>
<td>[ \beta: R_{jt} \cdot D_{jt} ]</td>
<td>.0123***</td>
<td>.00740***</td>
</tr>
<tr>
<td></td>
<td>(.00220)</td>
<td>(.00156)</td>
</tr>
<tr>
<td>Obs</td>
<td></td>
<td>3,475,421</td>
</tr>
</tbody>
</table>
Is $\beta$ the treatment effect of differential ratings on patient flow?

Physicians’ *time-varying* inner ability or budget before ratings may co-determine likelihood to receive high/low ratings and new patient flow

- E.g., physicians are already improving staff training
- E.g., physicians spend budget on marketing but do not have enough staffing

Measurement errors of ratings due to yearly aggregation

- Ideal treatment: average rating each patient sees before choosing physicians
- The cumulative end-of-year ratings is not the ideal treatment as it contains new ratings not-yet-observed by patients visiting physicians early in the year
Reviewers have intrinsic “harshness” in rating all businesses
  ▪ Generates ratings independently of physicians’ inner quality

IV: a physicians’ cumulative average reviewer’ “harshness”
  ▪ Reviewer “harshness” measured by her average rating in non-j businesses

\[ z_{jt} = \frac{1}{\#\text{review}_j} \sum_k r_{-j}^k \]

Exclusion Restriction: having a panel of observed “harsh” versus “lenient” reviewers does not correlate with physicians’ time-varying factors that co-determine the likelihood of receiving high/low ratings and patient flow
Estimation results

\[ y_{jt} = \chi_j + \theta_{t,h} + \theta_{t,s} + \lambda D_{jt} + \beta R_{jt} * D_{jt} + \epsilon_{jt} \]

<table>
<thead>
<tr>
<th>RHS \ LHS</th>
<th>Log Total Revenue</th>
<th>Log Unique Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>OLS</td>
<td>IV</td>
</tr>
<tr>
<td>\lambda: D_{jt}</td>
<td>-.0125** (0.00559)</td>
<td>-.0116* (0.00595)</td>
</tr>
<tr>
<td>\beta: R_{jt} * D_{jt}</td>
<td>.0123*** (0.00220)</td>
<td>.0186*** (0.00705)</td>
</tr>
<tr>
<td>Obs</td>
<td></td>
<td>3,475,421</td>
</tr>
</tbody>
</table>

Data sample: Yelp data merged with 100% Medicare FFS payment data 2012-2015, including all physicians rated or never rated. Two year fixed effects are HSA(j) and specialty(j) specific. Standard errors two way clustered at HSA and specialty levels.
<table>
<thead>
<tr>
<th></th>
<th>Physician</th>
<th>Restaurant (Luca WP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect per star increase</td>
<td>1-2% increase in patient volume</td>
<td>5-9% increase in revenue</td>
</tr>
<tr>
<td>New consumer share</td>
<td>35% (pcp only) -43% (all Yelp)</td>
<td>36%-49%</td>
</tr>
<tr>
<td>Implied response rate per new</td>
<td>3.8%</td>
<td>16.5%</td>
</tr>
<tr>
<td>consumer per star increase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% new consumers use Yelp</td>
<td>10%</td>
<td>75%</td>
</tr>
<tr>
<td>Implied response rate per new</td>
<td>38%</td>
<td>22%</td>
</tr>
<tr>
<td>consumer using Yelp per star</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Event study of “Harshness” on patient flow

- Test whether physicians with different reviewer harshness have differential pre and post trends around being rated

\[ y_{jt} = \chi_j + \theta_{t,h} + \theta_{t,s} + \sum_k \lambda_k 1(t - d_j^f = k) + \sum_{g \in \{h,l\}} \sum_k \beta_{k}^g 1(t - d_j^f = k) z_j^g + \epsilon_{jt} \]

\( \lambda_k \): time trends of \( k \)th year since being first rated
\( d_j^f \): first rating year of physician \( j \)
\( z_j^h \) & \( z_j^l \): an indicator that first year reviewer harshness of physician \( j \) higher than 67th percentile or lower than 33th percentile
\( \beta_k^h \) and \( \beta_k^l \): differential time trends of \( k \)th year since being first rated for a physician with high or low first-year harshness, compared to those with medium harshness
Event study of physician patient flow by different first year “harshness”

- Each dot shows $\beta_k^h$ and $\beta_k^l$ in the $k$th year since being rated

LHS: Log revenue

LHS: Log # unique patients
Discussion on exclusion restrictions

- **Threats to the exclusion restrictions:**
  - “Harsher” reviewers are choosy and gravitate towards physicians with high ability
  - “Harsher” reviewer more likely post when the physician quality is low=> the observed “harshness” correlated with unobserved factors

- Time-unvarying factors are controlled by physician FE=>only need to worry about time-varying factors

- Previous event study shows no differential patient flow before being rated for physicians with high vs low first-year harshness

- Observable physician quality characteristics are not correlated with instruments and indicator of being rated

- Construct alternative IV as the avg rating in non-medical businesses only and residualize the impact of baseline business ratings
1. What quality information do ratings convey?

2. Do ratings affect patients’ physician choices?

3. Patterns of physician behavior change?
Do physicians change practice behaviors after being rated?

- Being first rated makes future reviews more likely and increases a physician’s salience on internet

- Physicians may now have more incentives to please patients in order to potentially improve their ratings

- Will physicians try to please patients by ordering possibly wasteful (lab & imaging) and harmful (opioids) substances and impact health?
Empirical Strategy – Diff-in-diff

- Define patients of cohort $m \in 2009 \ldots 2015$:

- For primary care physicians first rated in year $m$ (treatment) and those first rated in 2016/17 (control), compare their patients’ health services received before and after year $m$
  - Link a patient to her primary care physician
  - Restrict to preexisting “before-m” patients who first visit their physicians before $m$

- Identification assumption:
  - Timing of first review is exogenous to physician and preexisting patient behaviors
  - Preexisting patients do not desire more health services after physicians are rated
  - Differences in patients’ utilization/outcomes after ratings reflect physician efforts
Including treatment and control preexisting patients of all cohorts, a cohort\((m)\)-patient\((i)\)-year\((t)\) level diff-diff from Medicare claim 2008-2015:

\[
h_{it}^m = \chi_{ij(i,t)} + \theta_{t,h}^m + \theta_{t,s}^m + \sum_k \alpha_k 1(t - m = k) T_{j(i,t)}^m + \epsilon_{it}^m
\]

\(h_{it}^m\): Patient \(i\)’s health utilization/outcome in year \(t\), who is of cohort \(m\)

\(\chi_{ij(i,t)}\): A patent \(i\)-physician \(j\) relationship FE. Constant if \(i\) stays within her physician \(j\). If \(i\) switches to \(j'\) in some year, a new FE for the new relationship

\(T_{j(i,t)}^m\): Whether physician \(j\) is in the “treatment” group (first rated in year \(m\)) in cohort \(m\)
• $Outpatient spending per primary care visit on avg ↑$8 (se=4) for the treatment patients after first rating
Outpatient, labs and imaging ↑; no changes for opioids

- Outpatient spending per primary care visit on avg ↑$8 (se=4) for the treatment patients after first rating

- Lab & imaging spending per primary care visit on avg ↑$2 (se=1) for the treatment patients after first rating

Average post effect is 1.93 (se= 1.07)
P value for pretrend is .501

Year

Year

≤-3
-2
-1
0
1
2
≥3
$Outpatient, labs and imaging ↑; no changes for opioids

- $Outpatient spending per primary care visit on avg ↑$8 (se=4) for the treatment patients after first rating
- $Lab & imaging spending per primary care visit on avg ↑$2 (se=1) for the treatment patients after first rating
- $Opioid spending does not significantly change for the treatment patients after first rating
# ER visits and health risk scores hardly change

**CMS risk score**

**Charlson risk score**

- Average post effect is -0.015 (se = 0.025) with a P value for pretrend of 0.0776.

- Average post effect is -0.055 (se = 0.0095) with a P value for pretrend of 0.546.

Low ratings only

Freyaldenhoven et al test
1. Ratings convey physicians’ interpersonal skills and (+) correlate with clinical quality

2. Patients’ choices of physicians respond significantly to ratings
   ▪ ↑ One star in physician rating => ↑ about 1-2% in a physician’s patient volume and revenue annually

3. After rated, physicians slightly ↑ lab & imaging tests, but do not order more opioids or impact health

Re-cap of positive economics findings
Conclusions

- Yelp ratings significantly impact patients and physicians

- Patients benefit from ratings despite the potential concerns
  - Not the wrong measure—better rated physicians good in many dimensions
  - High acceptance—ratings bring consumers to higher-rated physicians
  - Physicians do not hurt patients—they do not order more opioids

- Potential costs for other players:
  - Small extra costs to taxpayers
  - Investment costs and risks to physicians

- How to design a better, more optimal physician information system?
  - How much do patients understand the quality implications of ratings?
  - Combining both patient satisfaction scores and objective quality metrics?
  - Combining ratings with an appointment system?