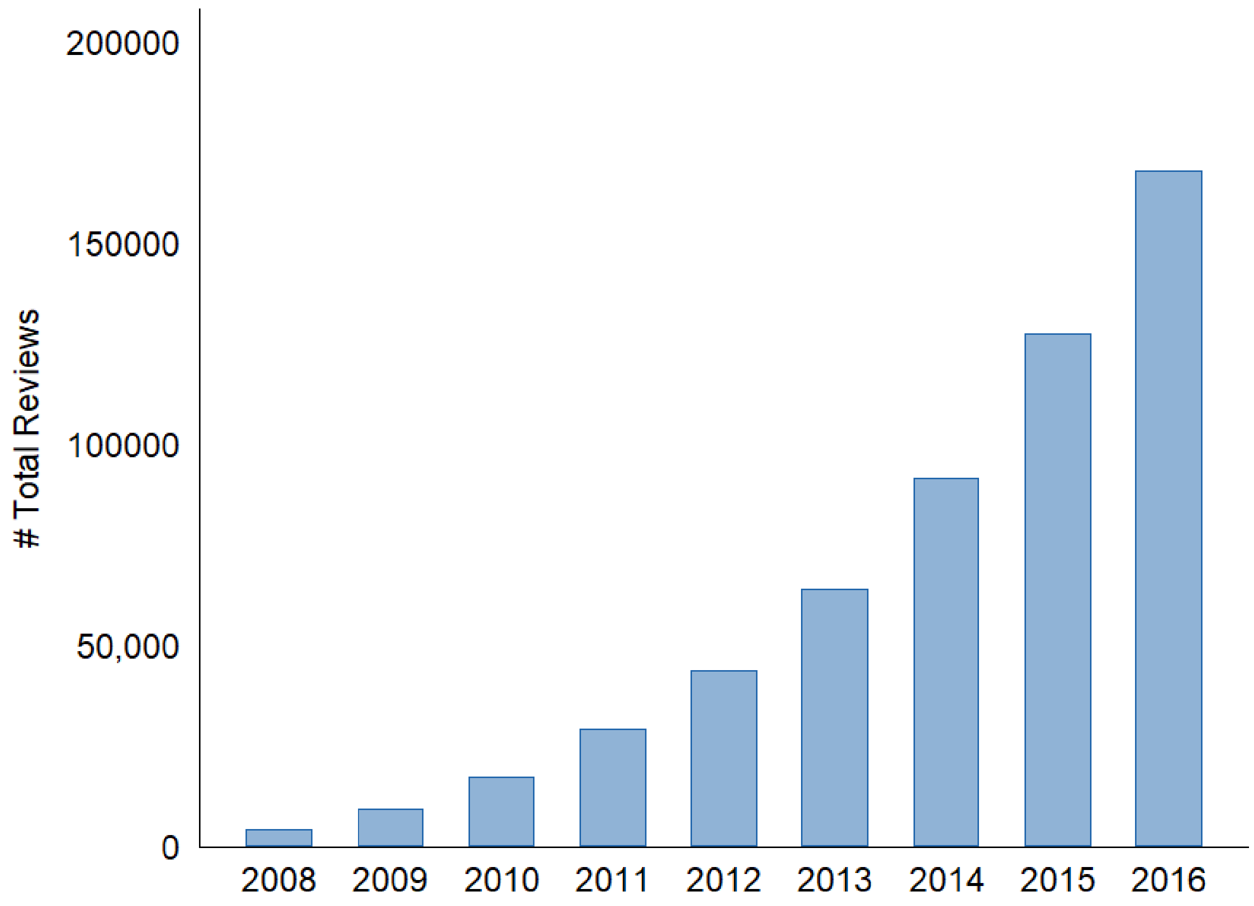
User-Generated Physician Ratings: Evidence from Yelp

Yiwei Chen

3.20.2019

Why study online user-generated ratings for physicians?

- 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,



Why study online user-generated ratings for physicians?

- Online ratings can potentially improve healthcare efficiency
  - Consumers have little information on which to base physician choices
  - In other industries, online ratings have steered consumers to better businesses 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 be distorted to prescribe harmful treatments e.g. opioids to please patients (multi-tasking)

Q&A: Do you think online physician ratings are good for patients (Pick one answer)

a. Yes
b. No
c. It’s hard to say

Research questions and preview of results

- Use Yelp ratings and Medicare claim data to study these challenges and examine whether online physician ratings are good for patients

1. What are the contents of Yelp physician reviews and do they correlate with clinical quality?
   - Text analysis shows Yelp 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?
   - Using reviewers' harshness in rating other businesses as instruments for physicians' ratings to retrieve the "causal" effects of ratings on annual revenue
   - 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?
   - Using a diff-in-diff comparing physicians who are rated earlier vs later
   - Suggestively, physicians slightly increase lab & imaging tests, but no statistically significant evidence shows they increase opioid prescriptions

Related Literature

- 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 that can be adopted for other rating studies

- 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
- 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

Medicare vs private

Sample Yelp page under “Doctor” tag

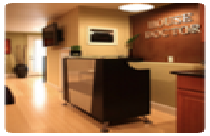
1. Dolhun Clinic

40 reviews

Family Practice, Internal Medicine, Home Health Care

Serving San Francisco and the Surrounding Area
(415) 923-3090

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

Potrero Hill

San Francisco, CA 94107
(415) 834-5364

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

Lakeside

2645 Ocean Ave
San Francisco, CA 94132
(415) 452-1200

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

Lower Pacific Heights

2300 Sutter St
San Francisco, CA 94115
(415) 928-7550

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, 542,977 historical reviews for 95,030 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 Yelp listings with “MD”, “DO” ,“OD” are uniquely matched
  - Most in small groups and in primary care and face-to-face specialties such as family medicine, internal medicine, dermatology, etc.

Matching sum stats

sum stats

1. What quality information do ratings convey?

Contents of Yelp reviews

- 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."

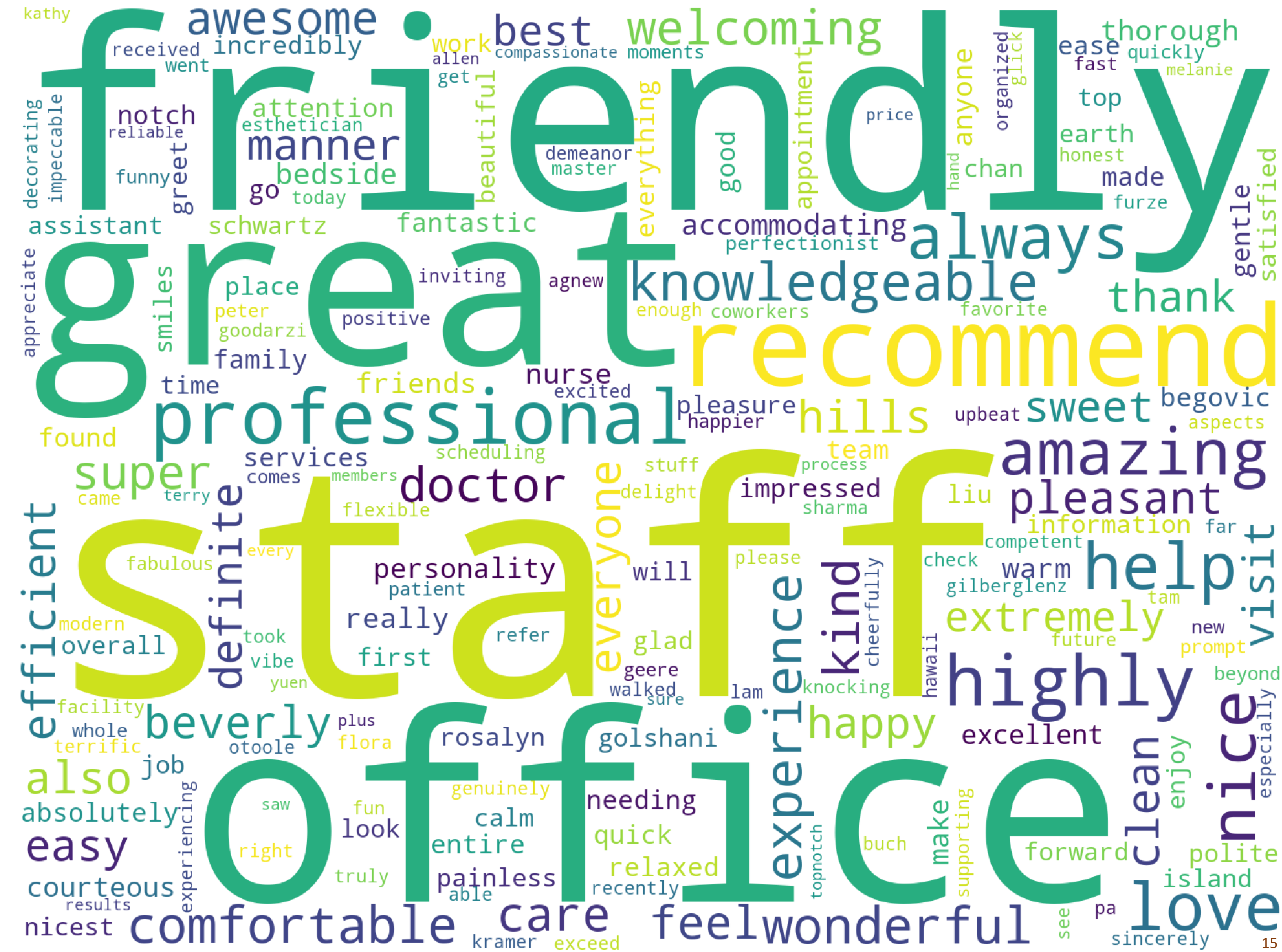
- Need methods to aggregate 200k+ reviews into categories

A machine learning approach

- 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.

name
usually
let
even
hard
long
laughed
something
smart
actually
forever
time
doctor
get
around
good
total
atmosphere
also
seriously
many
seen
happened
nothing
cool
sometimes
may
stop
super
though
anywhere
use
omg
always
best
now
every
ladies
mind
back
take
feel
anything
else
sure
little
guess
sweetest
rather
masterson
go
really
someone
remembers
done
doc
change
patient
thought
much
honestly
exactly
somewhere
might
found
point
got
big
bump
seem
gone
wrong
part
happy
amazing
miles
glad
look
guy
enough
personality
worth
whole
heard
know
job
probably
nice
everything
bad
away
way
sad
phoenix
especially
family
still
lots
pretty
find
great
another
almost
lived
drive
linda
keeps
make
right
love
can
see
yes
hope
problem
treat
new
earth
humble
needing
leave
work
straight
try
never
will
start
tell
just
believe
crazy
carried
reason
goes
sweet
different
types
lol
funny
first
like
man
price
real
matter
deal
oh
luck
trust
important
felt
plus
talk
one
better
ever
hear
give
figure
want
ok
sick
bit
care
says
must
everyone
moved
maybe
definite
extra
drs
days
fan
business
anymore
wow
fact
kind
far
imagine
things
means
place
years
think
comes
people
help
wish
anyone
hatessince
listened
awesome
understand
whatever
friends
honest


questions
time
concerns
answer
listened
feel
patient
explain
make
manner
care
understand
thorough
rushed
doctor
always
take
knowledgeable
like
great
detail
issues
kind
comfortable
bedside
never
felt
took
sure
personality
options
recommend
also
really
approach
everything
visit
attention
appreciate
health
information
humor
sense
needing
ask
good
help
best
thoroughly
decisions
clearly
confidence
medicated
plan
important
extremely
highly
spends
forward
clear
treat
gave
ever
respond
demeanor
response
ensures
direct
smart
peace
incredibly
conditions
providing
appointment
hand
opinion
future
possible
honest
patience
latest
may
education
value
found
solving
genuinely
understood
interested
experience
greenbaum
definite
things
seem
enough
regarding
reassuring
communications
says
can
get
happy
intelligent
paoletti
term
problem
obrien
empathetic
many
gentle
warm
history
will
talk
discussing
every
truly
follow
exam
find
wonderful
fully
unlike
addressed
lid
long
encouragement
spent
thought
treatment
ease
know
sandler
met
diagnosis
collins
compassion
shammas
calm
share
consideration
ayabe
speak
professional
offered
quickly
way
singh
trust
steps
methods
email
first
situation
available
necessary
overall
impressed
expect
effort
passionate
proactive
completed
tough
rare
much
engaged
even
one
experiencing
put
amazing
honesty
easy
tv
superb
straightforward
extra
give
shows
lots
mind
additionally
meeting
without
made
seen
advice
appropriate
compassionate
excellent
examinations
levine
respect
adberg
focusing
explaning
actually
integrative
g
integrative


kathy
awesome
best
welcoming
thorough
received
incredibly
work
allen
compassionate
moments
ease
quickly
fast
melanie
went
get
organized
glick
friendly
notch
attention
esthetician
reliable
price
top
decorating
impeccable
manner
demeanor
master
earth
honest
funny
bedside
beautiful
everything
good
appointment
anyone
chan
hand
made
greet
go
today
furze
assistant
schwartz
fantastic
accommodating
perfectionist
always
gentle
satisfied
great
appreciate
smiles
place
inviting
agnew
knowledgeable
thank
peter
positive
enough
coworkers
favorite
goodarzi
recommend
family
nurse
excited
time
friends
professional
pleasure
happier
hills
upbeat
sweet
begovic
aspects
found
services
scheduling
team
amazing
super
comes
members
doctor
stuff
process
came
terry
delight
impressed
liu
pleasant
flexible
sharma
information
far
efficient
definite
everyone
competent
every
please
check
fabulous
personality
will
warm
visit
help
patient
gilberglenz
tam
modern
really
extremely
new
staff
refer
future
prompt
overall
took
vibe
first
glad
kind
cheerfully
geere
yuen
walked
sure
lam
knocking
hawaii
highly
beyond
facility
experience
nice
especially
whole
beverly
plus
rosalyn
golshani
happy
excellent
terrific
otoole
flora
job
also
genuinely
clean
enjoy
absolutely
saw
fun
needing
office
calm
easy
look
quick
make
supporting
right
entire
topnotch
buch
forward
polite
experiencing
relaxed
island
truly
courteous
painless
recently
able
results
nicest
comfortable
care
feel
wonderful
see
pa
love
kramer
exceed
sincerely

Survey design

Correlations between ratings and clinical quality

- Reviews do not center on clinical quality
- If patients choose higher-rated physicians, unclear whether they match with physicians of better/worse clinical quality
- Do higher ratings correlate with better clinical quality measures?
  - If so, patients will visit physicians with higher clinical quality on average if visiting higher rated ones

Correlations between ratings and physician medical credentials

- 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

- Goal: Is R_j^{2017} predictive of physician's medical credentials

Higher Yelp ratings associated with better medical credentials

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

RHS	LHS: Measurement of Physician Clinical Ability/Quality		
	1(Board Certification)	Medical School Ranking	Log(# Self-reported accreditations)
β: Ratings	.0299*** (.00502)	1.736*** (.355)	.0280*** (.00836)
Implications of 1->5 stars	+.12 in probability	+7 in ranking	+11% in # reported
N	8,755	36,346	4,405
Mean of LHS	.73	59	.98
Sample	Healthgrades-Yelp (primary care physician only)	PhysicianCompare-Yelp	PhysicianCompare-Yelp

Standard errors two-way clustered at HSA and Specialty levels

Correlations between ratings and primary care health outcomes

- Do ratings correlate w/ better health outcomes?
  - 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

- Goal: Is R_j^{2017} predictive of patients' health outcomes?

Higher ratings correlated with better patient outcome and practices

RHS	**LHS: Measurement of Patient Health Outcomes and Medical Practices**		
	1(Eye Exam for Diabetics)	1(Mammogram for Breast Cancer Screening)	1(Preventable Inpatient Admissions)
β: Ratings	.00195** (.000787)	.00622*** (.00112)	-.000866*** (.000245)
1->5 stars	+.008 (+1.5%)	+.025 (+3.7%)	-.0036 (-8.7%)
N	810,464	751,746	3,013,423
Mean of LHS	.52	.67	.04
Control	Patient demographics, past year risk scores, year FE, location FE		

- Also use Charlson and CMS risk scores including current year diagnosis as LHS
- A 1-5 star rating change associates with -3% decrease in risk scores statistically significantly

1. What quality information do ratings convey?

2. Do ratings affect patients’ physician choices?

Research object

- 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 have higher 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

χ_j: Physician fixed effects

θ_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

Estimation results

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

RHS\LHS	Log Total Revenue	Log Unique Patients
Method	OLS	OLS
λ: D_{jt}	-.0125** (.00559)	-.00786* (.00423)
β: $R_{jt} * D_{jt}$	.0123*** (.00220)	.00740*** (.00156)
Obs	3,475,421	

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

Identification—Endogeneity Concerns

- Is β 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

Identification—IV strategy

- 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

- 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

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

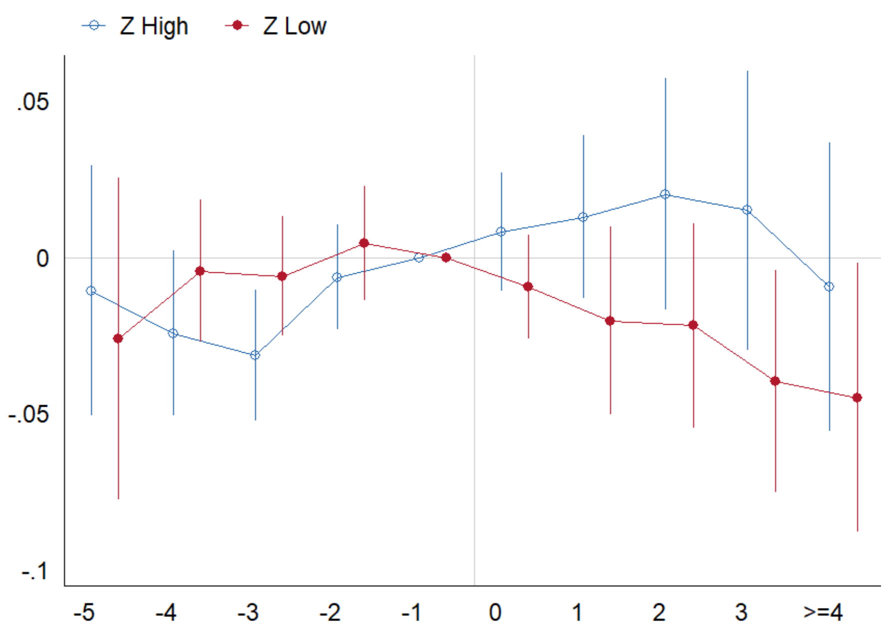
RHS\LHS	Log Total Revenue		Log Unique Patients	
Method	OLS	IV	OLS	IV
λ: D_{jt}	-.0125** (.00559)	-.0116* (.00595)	-.00786* (.00423)	-.00715* (.00430)
β: $R_{jt} * D_{jt}$	.0123*** (.00220)	.0186*** (.00705)	.00740*** (.00156)	.0121*** (.00477)
Obs	3,475,421			

First stage | Alt inst | Alt timing | Magnitude possible? | Heterogeneity

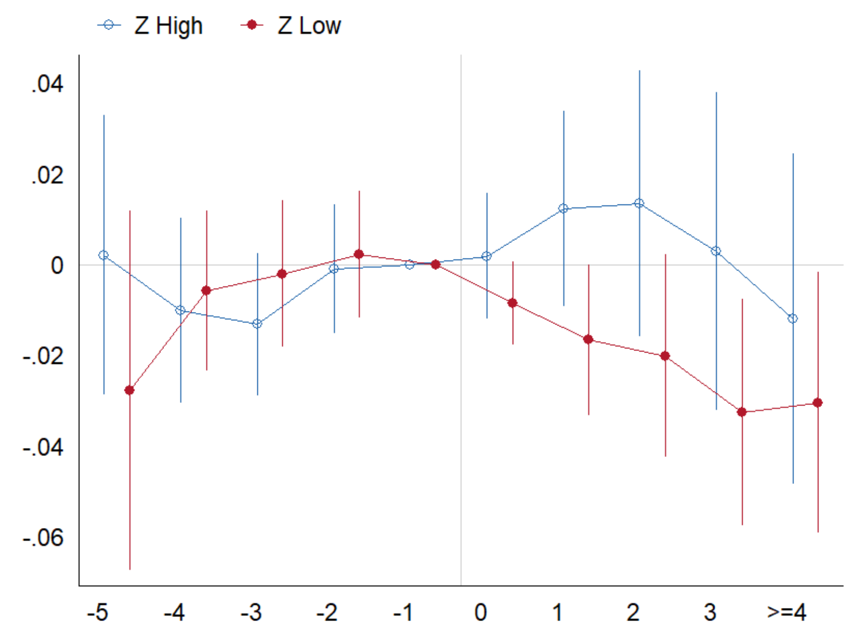
Event study of physician patient flow by different first year “harshness”

- Test whether physicians with different reviewer harshness have differential pre and post trends around being rated
- β_k^h: physicians whose first-year reviewers’ “harshness inst” is high
- β_k^l: physicians whose first-year reviewers’ “harshness inst” is low

LHS: Log revenue



LHS: Log # unique patients



Specification

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 $m \in 2009 \dots 2015$:

- 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

Empirical Strategy

- 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

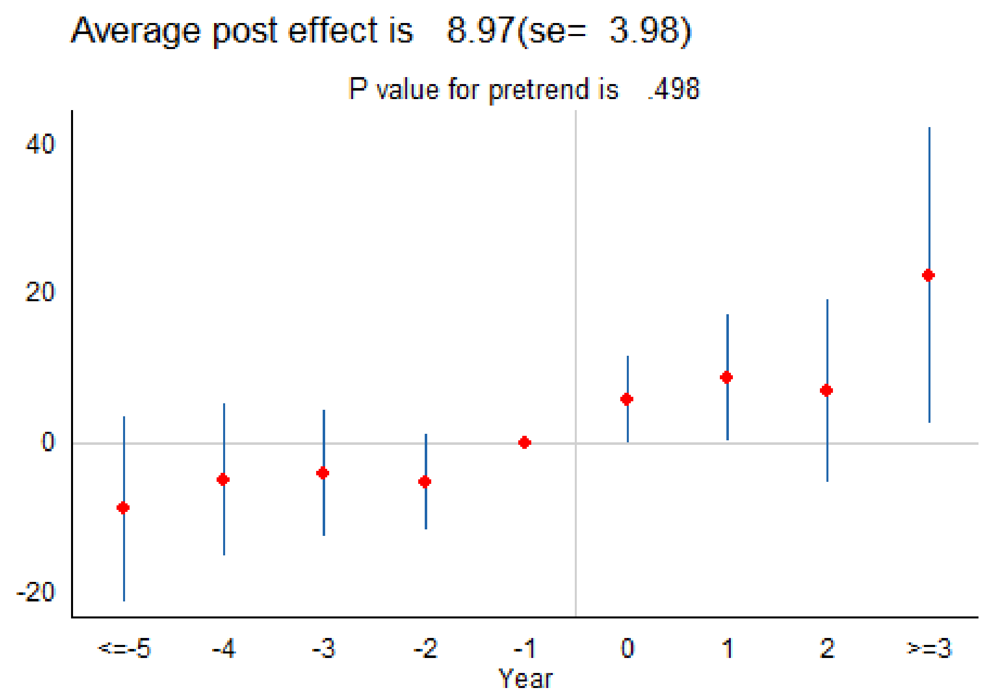
T_j^m: Whether physician j is in the "treatment" group (first rated in year m) in cohort m

Balance Check

$Outpatient, labs and imaging ↑; no changes for opioids

- $Outpatient spending per primary care visit on avg ↑$9 (se=4) for the treatment patients after first rating

$Outpatient spending per primary care visit

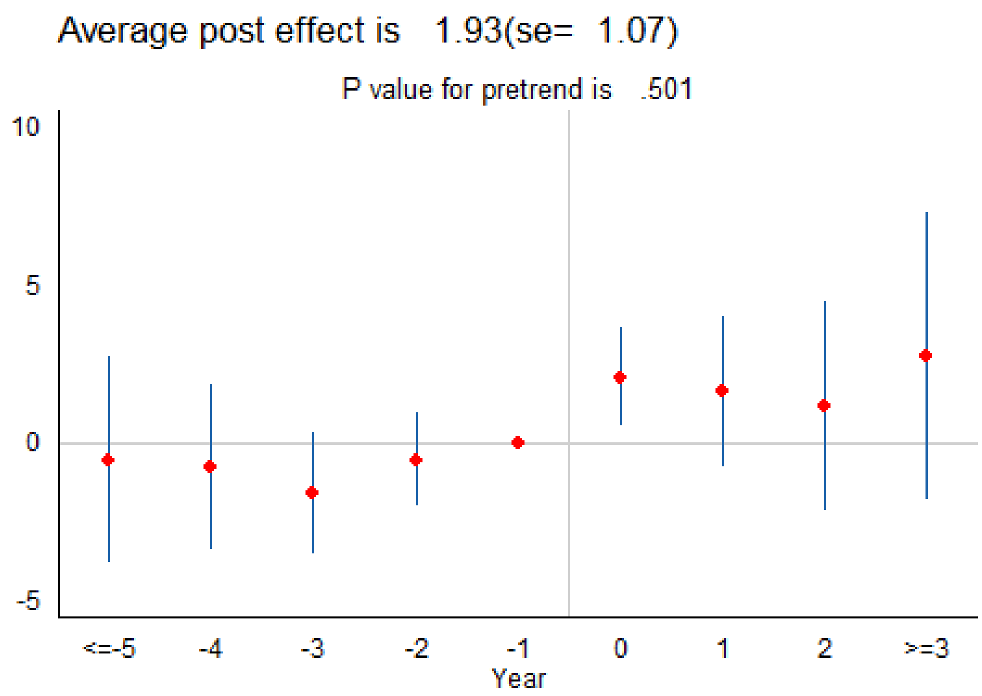


$Outpatient, labs and imaging ↑; no changes for opioids

- $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

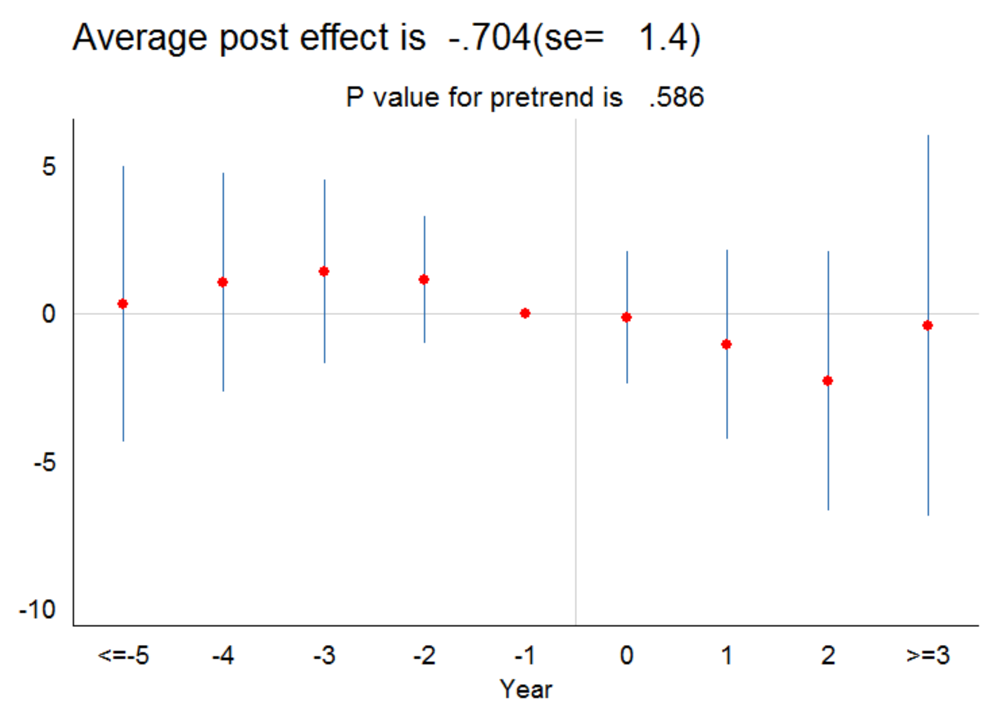
$Lab & imaging spending per primary care visit



$Outpatient, labs and imaging ↑; no changes for opioids

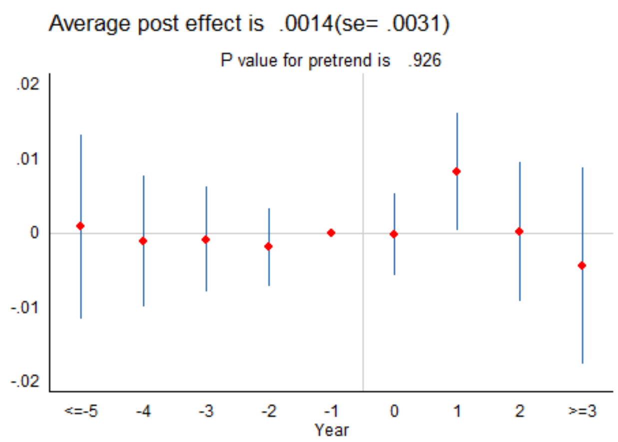
- $Outpatient spending per primary care visit on avg ↑$8 (se=4) for the treatment patients after first rating
- $Lab & imaging spending per primary care visit on avg ↑$2 (se=1) for the treatment patients after first rating
- $Opioid spending does not significantly change for the treatment patients after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

$Opioid spending per primary care visit

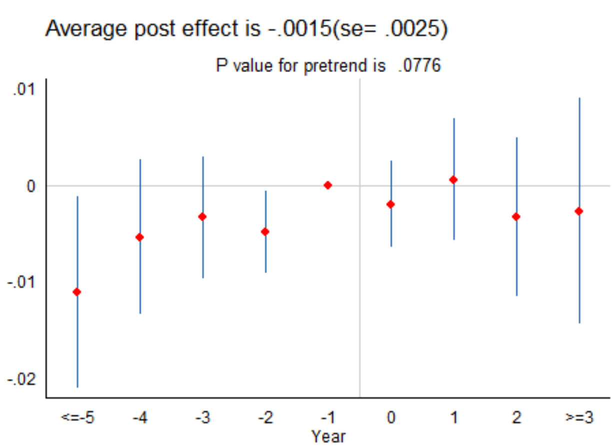


#ER visits and health risk scores hardly change

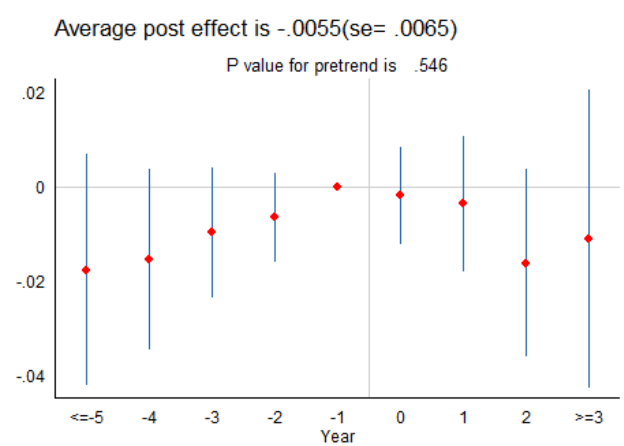
ER visits



CMS risk score



Charlson risk score



Low ratings only

Freyaldenhoven et al test

Conclusions and Policy Implications

- 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—little evidence shows they order more opioids

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

Q&A: Do you think online physician ratings are good for patients (Pick one answer)

a. Yes.
b. No.
c. It’s hard to say.
d. Probably yes but more research is needed.
e. Probably no but more research is needed.