User-Generated Physician Ratings—Evidence from Yelp

Yiwei Chen[*]

11.8.2018

Please click HERE for the latest version

Abstract

User-generated physician ratings from online sources are increasingly popular, but since consumers typically lack the ability to evaluate clinical quality, it is unclear whether these ratings actually help patients. Using the universe of Yelp physician ratings matched with Medicare claims, I examine what information on physician quality Yelp ratings reveal, whether they affect patients' choices of physician, and how they influence physician behavior. Through text and correlational analysis, I show that although Yelp reviews primarily describe physicians' interpersonal skills, Yelp ratings are also positively correlated with various measures of clinical quality. Instrumenting physicians' average ratings with reviewers' "harshness" in rating other businesses, I find that a one-star increase in physicians' average ratings increases their revenue and patient volume by 1-2%. Using a difference-in-differences strategy, I test whether, in response to being rated, physicians order more substances that are desirable by patients but potentially harmful clinically. I generally do not find that physicians substantially over-prescribe. Overall, Yelp ratings seem to benefit patients—they convey physicians' interpersonal skills and are positively correlated with their clinical abilities, and they steer patients to higher-rated physicians.

[*] I am indebted to my advisers Liran Einav, Jonathan Levin, Mark Duggan, Kate Bundorf, and Grant Miller for their guidance and encouragement throughout the project. I am also grateful to numerous feedbacks from my informal adviser Matthew Gentzkow. Comments and suggestions from Jay Bhattacharya, Nicholoas Bloom, Tim Bresnahan, David Chan, Amitabh Chandra, Michael Chernew, Raj Chetty, Han Hong, Caroline Hoxby, Daniel Kessler, Michael Luca, Emily Oster, Maria Polyakova, Jesse Shapiro, Dan Zeltzer, and seminar participants at Stanford labor/public economics and industrial organization workshops and ASHEcon 2018 conference are also appreciated. I also thank institutional knowledge and suggestions from Anna Lembke and Harold Luft. Generous funding from Leonard W. Ely and Shirley R. Ely Graduate Student Fellowship is acknowledged and appreciated. All errors are my own.

1 Introduction

Consumers have increasingly been turning to user-generated ratings from online sources when choosing products and services. Healthcare is no exception to the trend. A recent survey found that more than half of Internet users consult online physician ratings when looking for physicians.[1] In various industries such as restaurants, hotels, and consumer goods, these user-generated ratings appear to help promote efficiency—they disseminate quality information, bring consumers to better businesses, and motivate businesses to improve quality (Dranove & Jin, 2010; Luca, 2015). While they may similarly improve efficiency in healthcare, many challenges may prevent them from delivering on such promise.

Perhaps most notably, healthcare quality is inherently difficult for patients to evaluate. Among other things, healthcare should be clinically safe, effective, efficient, and patient centered.[2] Consumers can more easily assess some quality dimensions such as patient-centeredness, but many find it challenging to appraise clinical dimensions as consumers generally lack the knowledge and experience to evaluate them. If most users review their physicians based only on patient-centeredness and the resulting physician ratings are uncorrelated or even negatively correlated with clinical quality, the ratings may steer patients who actually prioritize clinical quality toward the wrong providers. Another challenge is the degree to which online ratings actually impact patients' choice of physicians. Perhaps patients do not find the user ratings meaningful with respect to clinical quality or they consider the numbers of reviews for physicians too sparse, so they place little importance on the ratings when making physician choice decisions. Furthermore, physicians may, in response to being rated, work harder to please patients, which would usually benefit them. However, due to patients' lack of abilities in evaluating all quality dimensions, one may worry that physicians would "teach to the test": they may work to improve patient-centeredness that is easier to observe by patients, but neglect clinical quality that is harder to observe. An example of such behaviors is that physicians may start to over-prescribe opioids to make patients happier at the cost of clinical safety.

I examine these questions—whether user-generated physician ratings from online sources signal physician quality information, affect patients' physician choices, and change physician behaviors—using data from the universe of Yelp physician ratings matched with Medicare claims. This empirical setting provides several advantages. First, Yelp is the leading physi-

[1] Surveyed from Software Advice in 2017. https://www.softwareadvice.com/resources/how-patients-use-online-reviews/

[2] Healthcare quality dimensions defined by Institute of Medicine. http://nationalacademies.org/hmd/Global/News%20Announcements/Crossing-the-Quality-Chasm-The-IOM-Health-Care-Quality-Initiative.aspx

cian rating website and the numbers of reviews have grown exponentially in recent years. These facts make Yelp an important and representative example of online physician ratings as a whole. Second, since I collect the Yelp data at the individual review level over time, I am able to create a longitudinal physician review sample and know when and how the ratings change for a physician. Third, as Medicare allows one to link ratings with the claims data at the individual physician level, I am able to analyze granular individual physician behavior responses for more than 30,000 rated physicians.

To understand the content of Yelp reviews and ratings, I first use a machine-learning algorithm to detect the common topics of the written reviews.[3] The algorithm automatically categorizes the reviews' text into a set of "topics," each of which is a probability distribution of keywords that tend to occur together. From my reading of the keywords from the most common "topics," I interpret that reviewers primarily describe physicians' interpersonal skills and office amenities in writing reviews. To understand how human readers may interpret reviews, I conduct a survey recruiting respondents from Amazon Mechanical Turk and asking them to interpret a random sample of 1,500 Yelp reviews. The survey finds 81% of reviews containing "service quality related" information, such as friendliness, and 44% containing "clinical quality related" information, such as treatment procedures. These results suggest that readers regard the information from Yelp sources as focusing more on physicians' patient-centeredness than on their clinical effectiveness.

To evaluate the clinical quality of higher- versus lower-rated physicians, I further correlate the 1-5 star ratings with various measures of clinical quality. Empirically, higher-rated physicians have better educational and professional accreditations. Primary care physicians with higher ratings show higher adherence to clinical guidelines in ordering screening exams. Their patients also display better health outcomes after controlling for observed patient characteristics—they have lower preventable inpatient admissions and better ex-post risk scores that are indicators of predicted health conditions. The possible patient selection may confound the interpretation of clinical ability, although it is perhaps less of a concern for the measures of physicians' educational and professional backgrounds and adherence to clinical guidelines. To the extent that these measures reflect clinical abilities rather than patient selection, patients visiting higher-rated physicians will be matched with physicians with better clinical quality. In addition, I also find that the positive correlations between ratings and clinical quality measures remain strong even when controlling for observable physician quality from other non-Yelp sources, such as medical school rankings, suggesting the correlations between ratings and clinical quality measures are beyond what other sources

[3] The algorithm is latent Dirichlet allocation. See Gentzkow et al. (2017) for a discussion in economics applications.

may already tell the patients.

Next, I explore whether Yelp ratings influence patients' physician choices. If consumers do base their decisions on these ratings, according to the findings above, they will match with better overall physicians. I examine this question by testing whether physicians with higher Yelp ratings grow faster in patient flow than those with lower ratings. An OLS estimation shows that after physicians are rated, physicians who are 1-star higher in average ratings are associated with 1.3% and 0.7% faster growth in annual revenue and patient volume than before being rated. However, one may be concerned that physicians may receive different average ratings for reasons that directly affect patient flow, which would confound the causal interpretation of Yelp effects in the previous results. For example, physicians may have already been improving their office amenities, causing improvement in patient flow and increase in the likelihood of receiving higher ratings. Or, a physician may have decided to spend her budget on marketing but does not have enough staffing capacity, causing improvement in patient flow but resulting in increase in the likelihood of receiving lower ratings.

To get around the potential endogeneity concerns, I use an instrumental variable approach. I define a reviewer's "harshness" as her average ratings when rating other businesses on Yelp. Due to the nature of the platform, those businesses are mostly restaurants. The measurement captures a reviewer's baseline "harshness" in reviewing any businesses, which would influence her ratings for her physicians while also possibly being orthogonal to the endogenous factors that directly affect patient flow, such as physicians' time-varying quality. I calculate a physician's annual cumulative reviewer "harshness" as the average "harshness" for all reviewers who have rated that physician by each year. Using that variable as an instrument for a physician's yearly cumulative average Yelp rating, I find that a 1-star increase in a physician's average rating statistically significantly increases her annual patient revenue and volume by 1.9% and 1.2%. I also run a series of tests to show that the exclusion restriction is plausible. For example, I find that physicians' cumulative "harshness" is uncorrelated with observable physician quality characteristics. An event study also shows that physicians who receive high or low levels of first-year "harshness" do not differ in their patient flow prior to being rated; however, after being rated, the physicians with more "harsh" reviewers have a downward trend in patient flow. In addition, I also discover that the effects of ratings on patient flow are stronger for listings with more reviews and for physicians with younger and more educated patients, all of which are consistent with the predictions from patients' choice responses to Yelp ratings.

Last, I present some suggestive evidence regarding whether being rated changes physician behaviors. Critics have worried that in response to being rated, physicians may try to please patients by over-prescribing services that patients tend to overvalue and desire, such as opioids and lab and imaging services (Pham et al., 2009; Lembke, 2012; Hoffmann & Del Mar, 2015). At the margin, these services may result in wasteful or harmful consequences if physicians prescribe them only to please patients and potentially improve their ratings. I use a difference-in-differences strategy to test whether physicians order more of these services after being rated. I focus on physicians who receive their first ratings earlier (treatment group) and those who receive first ratings later (control group) and compare their patients' health services received before and after the treatment group physicians receive their first ratings (treatment date). The analysis is restricted to patients who have already visited their physicians before the treatment date in order to remove the potential selection of new patients after Yelp ratings are posted. I find that after the treatment date, the treatment patients receive slightly higher amounts of total outpatient and lab and imaging services than those in the control group, but no significant differences in the amount of opioid prescriptions and general health outcomes are detected. Assuming that the timing of first ratings is uncorrelated with physician behavior and patient selection, and patients do not desire different amounts of health services after their physicians are rated, the evidence would suggest that physicians do not seem to order more opioids or negatively impact patients' health after being rated, although they order slightly more lab and imaging services that are possibly wasteful.

Overall, my findings indicate that Yelp ratings actually benefit patients despite the potential concerns. Although Yelp reviews focus primarily on physicians' service quality and interpersonal skills, Yelp ratings are positively correlated with various measures of physician clinical quality, suggesting that higher-rated physicians have better quality in multiple dimensions. As Yelp effectively brings patients toward higher-rated physicians, it also brings patients toward better physicians. At the same time, I do not find significant patterns suggesting that physicians prescribe harmful treatments or harm patient health after they are rated on Yelp.

This paper contributes to several strands in the economics and medical literatures. First, many studies examine the traditional outcome-based report cards for health providers and find that they do not seem to have elicited a large consumer response and have not always had a positive impact on physician behavior (Dranove et al., 2003; Werner et al., 2012; Kolstad, 2013). In response to some of these concerns, many authors have instead proposed that the next-generation report cards should be easy to use and understand for consumers (Schneider & Epstein, 1998; Fung et al., 2008; Faber et al., 2009; Kolstad & Chernew, 2009; Hibbard et al., 2010; Sinaiko et al., 2012). This paper represents a step in that direction by analyzing the impact of user-friendly and subjective Yelp physician ratings, a new type of report card, on both patients and physicians. A young and growing literature has started to scrutinize the online rating setting for physicians, typically using regional data and focusing on large specialized providers, and finds that the ratings are positively correlated with both medical and non-medical quality (Bardach et al., 2013; Ranard et al., 2016; Howard et al., 2017; Lu & Rui, 2017). My paper supplements this literature by instead focusing on a large representative physician rating sample consisting of individual physicians from mostly face-to-face specialties and examining both the correlations of ratings and quality and the implications for patients' and physicians' behaviors.

Second, this paper contributes to the economics literature on how rating mechanisms affect consumer demand by proposing a new causal research design. The literature has covered a wide array of industries including health plans, restaurants, education, and consumer goods (Scanlon et al., 2002; Jin & Leslie, 2003; Dewan & Hsu, 2004; Dafny & Dranove, 2008; Kolstad & Chernew, 2009; Rockoff et al., 2012; Luca & Smith, 2013; Luca, 2016). Methodologically, the studies in the literature often rely on cross-sectional or panel variation of the demand of the business unit in response to ratings. Some papers utilize the institutional design of the rating systems and exploit a regression discontinuity strategy (e.g., ratings above or below some thresholds are rounded to different numbers) (Anderson & Magruder, 2012; Luca, 2016). My paper proposes a different causal approach using an instrument variable design that exploits Yelp reviewers' "harshness" in other reviews. This approach allows me to use data variations of ratings not only within the rounding boundaries and can be applied in other rating demand estimations.

Third, this paper adds to the literature on how rated suppliers respond to quality disclosure, particularly in settings of supplier multitasking. The economics literature has long recognized that, if agents are monitored only in specific areas, they may invest in the rewarded areas but overlook the unrewarded ones (Holmstrom & Milgrom, 1991; Dranove et al., 2003; Jacob & Levitt, 2003; Werner et al., 2009; Mullen et al., 2010; Feng Lu, 2012). My paper contributes new empirical evidence to the applications of healthcare by documenting how physicians change their efforts after receiving their first Yelp ratings, in contrast to the previous correlational evidence (Pham et al., 2009; Fenton et al., 2012). The findings may also have implications for current efforts in using patient satisfaction scores for physician and hospital compensations in various public and private programs.[4]

The rest of paper is organized as follows. Section 2 describes the contents of the empirical setting, the data sources and construction, and the general trends of Yelp physician ratings. Section 3 explores what Yelp ratings reveal about physician quality. Section 4 studies the

[4] For example, Medicare now pays hospitals depending on their scores on the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, a measure of customer satisfaction.

impact of Yelp ratings on physicians' patient flow. Section 5 examines the behavior responses of physicians after being rated. Section 6 presents a stylized normative model to discuss who and how much each player benefits and loses from the Yelp ratings. Section 7 concludes the paper with a discussion of future research areas and policy implications.

2 Setting and Data

2.1 Empirical Setting

I use Yelp ratings as my empirical setting of online physician ratings. Yelp is the leading online platform on which consumers voluntarily leave feedbacks on their experiences with the visited businesses. In the platform, a business profile is created by a consumer, an owner, or directly by Yelp from scraping business directories. Once a profile is created, users can leave an impression on the business by leaving a text review and post a 1-5 star rating. When a consumer searches for a business via Google or directly within Yelp, a snapshot of the business profile is displayed. It shows the business contact information as well as key indicators such as the average ratings of the businesses and the numbers of reviews in total. After clicking on the snapshot, the detailed profile page of the "doctor" will be opened and display each individual review of the "doctor," the individual star associated with the review, and its posting date. Yelp also has a sophisticated algorithm to prevent fraudulent reviews. For example, it only allows reviews and ratings to be displayed if a reviewer has posted multiple reviews for different businesses as a way to screen out potential fraudulent reviews.

I study Yelp for physician ratings for the following reasons. First, Yelp has a long history in physician ratings and is one of the top players in online physician ratings. The firm was created in 2004 originally to rate physicians as the founder had a difficult time looking for a good one.[5] In 2014, a survey found that among all online rating websites, Yelp was used the most often: 27% of respondents indicated they used Yelp as the go-to-source of physician rating website.[6] Among physicians who are rated on both Yelp and the second largest website Healthgrades, their star ratings also have a strong +36% correlation. These features make Yelp a representative and appealing platform to study online physician ratings as a whole. Second, Yelp also provides very detailed rating information. Not only does it provide the content and the time stamp of each review, it also provides detail information on a reviewer—how many businesses she has reviewed, the ratings of these reviews, etc. These

[5] http://archive.fortune.com/magazines/fortune/fortune_archive/2007/07/23/100134489/index.htm

[6] Cited from https://www.softwareadvice.com/resources/medical-online-reviews-report-2014/.

pieces of information allow researchers to understand the time trends and the contents of ratings and the general rating patterns of reviewers.

The Yelp data in this paper is collected using an algorithm outlined in Appendix A.1. For each physician rated on Yelp, I collect the date of each individual review, the text of each review, the star rating of each review, and the historical distribution of each reviewer's ratings of all her rated businesses. All of these information are collected up to June 2017.

I capture a physician's patient flow and patients' health outcomes from the Medicare fee-for-service claims data. In 2013, the program covered more than 37 million U.S. residents, 30 million of which are elderly aged 65+.[7] The program is also widely accepted among physicians. Among primary care physicians alone, surveys found 93% of them accept Medicare insurances.[8] The wide coverage of Medicare allows me to analyze the impact of Yelp on a large and important sector of patients, the elderly, and the behavior responses from the vast majority of physicians, if rated on Yelp.

I use two sources of Medicare claims data. The first one is the 100% Medicare payment data from 2012 to 2015, which contains all physicians' annual revenue and numbers of unique patients served in a year from Medicare Part B fee-for-service program.[9] Second, I also obtain the research-identifiable-files (RIF) of Medicare Fee-For-Service inpatient and outpatient claims and Part D drug event files for a random 20% sample of Medicare enrollees between 2008 and 2015. They contain granular claim level information including procedure codes, prescriptions filled, the amounts of bills, dates of services, etc.

I also obtain physicians' educational and professional accreditations from external websites. From Healthgrades,[10] a large physician information website, I collect board certification status of primary care physicians, which is a voluntary test to demonstrate a physician's mastery of the minimum knowledge of and skills for their subject. From Physician Compare,[11] the official physician information website of physicians who bill Medicare endorsed by CMS, I obtain the medical school rankings of every physician.[12] From Physician Compare, I also obtain the total number of every physician's self-reported quality metrics, such as accreditations in "Preventive Care and Screening: Influenza Immunization," and "Documentation of Current Medications in the Medical Record."

[7] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CMSProgramStatistics/2013/Downloads/MDCR_ENROLL_AB/CPS_MDCR_ENROLL_AB_13.pdf

[8] https://www.kff.org/medicare/issue-brief/primary-care-physicians-accepting-medicare-a-snapshot/

[9] Obtained from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html.

[10] https://www.healthgrades.com/

[11] https://www.medicare.gov/physiciancompare/

[12] I merge medical schools with the rankings from U.S.News & World Report (short for Usnews in this paper) and StartClass (now defunct). I reverse the ranking so that the higher rankings the better schools, with unranked schools are imputed as rank 0.

2.2 Data Construction

2.2.1 Sample Construction

To bring together the Yelp and Medicare data, the main data sample is constructed by matching a physician's rating profile on Yelp with the National Provider Identifier (NPI) directory using physician last name, first name, and practice Health Service Area (HSA). As an NPI identifier is required for all physicians who bill Medicare, the directory is a super-set of physicians who bill Medicare. The algorithm is detailed in Appendix A.2.

In Yelp, I collect 95,030 physician profiles in total who are either individual physicians or group practices. The matching algorithm using names is only effective in finding individual physicians, restricting the scope of the paper. After applying the matching algorithm, 36,787 physicians from the Yelp profiles are uniquely matched with the NPI directory, which is the main physician sample for analysis. 10% of Yelp profiles have duplicates matches with the NPI directory even within an HSA and are discarded. To gauge the accuracy of the algorithm, I assume that listings with suffixes "MD," "DO," or "OD" are individual physicians and an ideal algorithm should find a unique match from the NPI directory. Among the 95,030 total listings collected, 43,906 listings contain such suffixes, and the algorithm uniquely matches 30,729 of them with the NPI directory. That is, the average matching rate is 70%.

2.2.2 Variable Definitions

A physician's Yelp average rating by each year is re-constructed as the physician's cumulative average rating up to the end of that year, which mimics what Yelp displays as the average rating of the business.

The variables on physicians' patient flow are directly obtained using Medicare payment data. I measure a physician's revenue and patient volume as the annual revenue a physician receives and number of distinct patients a physician sees from Medicare Part B.

A patient's health outcomes are constructed using Medicare claims. Using patient demographics and current year diagnosis, I compute a patient's Charlson comorbidity index, under which high scores mean higher predicted mortality, and the CMS-HCC 2015 risk scores, under which higher scores mean higher predicted spending.[13] Higher values of those health measures reflect more sickness of a patient in general. I compute whether a patient during a year receives the HEDIS recommended clinical procedure among eligible patients—eye ex-

[13] The Charlson comorbidity index model is adapted from https://www.lexjansen.com/wuss/2013/119_Paper.pdf. The CMS risk score model uses the V2213L2P version from https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors-Items/Risk2015.html and uses the community score. Both models are only calculated using the demographics and diagnoses of the current year for enrollees who are either new to Medicare or have fully enrolled for 12 months in the current year.

ams among diabetic patients and mammograms for breast cancer screening among eligible females.[14] Those measures reflect whether a patient receives a clinically recommend exam. A preventable inpatient admission indicator is also constructed to capture whether patients receive certain inpatient admissions that good primary care services may avoid.[15]

A patient's health utilization is also measured by the claims data. I first assign patient i in year t to her most frequently visited primary care physician j in that year to let the physician in "charge" of her spending. Then, I compute patient i's medical spending per primary care visit in year t as i's total spending from all service providers in that year divided by i' number of visits to physician j in year t.[16] These measures are constructed to reflect a patient's total medical spending including referral amount per primary care visit.

2.3 Data Patterns

2.3.1 Physicians in the Main Sample

To gauge the representativeness of the matched physicians among all Yelp individual physician listings, I compare some key summary statistics among the matched and unmatched Yelp listings that contain suffixes "MD", "DO", or "OD". These listings are most certainly individual physicians rather than group practices. Table A.1 shows that the listings that are matched and unmatched with the NPI directory have overall similar characteristic. The matched listings have similar average ratings as the unmatched ones, although they have slightly higher average numbers of reviews and are rated slightly earlier.

To understand the scope of the physicians that bill Medicare from the main sample compared to all Medicare physicians, the 36,787 Yelp-rated physicians in the main sample contain 26,822 physicians that billed Medicare in 2015, which is 5% of 560,626 physicians who billed that program in 2015.[17] Those physicians in the main sample are more common in primary care specialties: 35% of physicians in the main sample are in general medicine, internal medicine, family medicine, or geriatric medicine, compared to 28% overall. Figure A.1 displays the top 10 specialties among the physicians in the main sample versus all physicians that bill Medicare in 2015. The top specialties in the matched sample are generally

[14] These procedures are recommended by The Healthcare Effectiveness Data and Information Set (HEDIS), developed and maintained by the National Committee for Quality Assurance (NCQA). https://provider.ghc.org/open/providerCommunications/hedisInsight/coding-guidelines.pdf.

[15] Defined using the numerators from http://www.qualityindicators.ahrq.gov/Modules/PQI_TechSpec_ICD09_v60.aspx among the eligible patients by AHRQ.

[16] All spendings are top-coded at the 99th percentile within each year and converted to U.S. 2015 dollars using CPI-U.

[17] These numbers are computed by matching the main physician sample with Physician Compare demographic data in 2015 and the Medicare payment data in 2015. The computation also excludes health workers that are nurses, physician assistants, social workers, physical therapists, or chiropractors.

"face-to-face" specialties, such as family medicine, internal medicine, and dermatology, and do not contain non-"face-to-face" specialties such as anesthesiology and radiology that are common among all Medicare physicians. The matched physicians also work in smaller practice groups. I estimate a regression at the physician level including all physicians that billed Medicare in 2015, using a physician's organization size in 2015 as the dependent variable and whether a physician is in the matched sample as the independent variable, controlling for physician HSA and specialty fixed effects. The average organization sizes of the unmatched physicians are 130% bigger than those from the main sample.

The comparisons above show that physicians who are rated on Yelp are different from average physicians that bill Medicare. One should be aware that the scope of this study is focused on the rated physicians, who are more common in face-to-face and primary care specialties and in smaller practice groups.

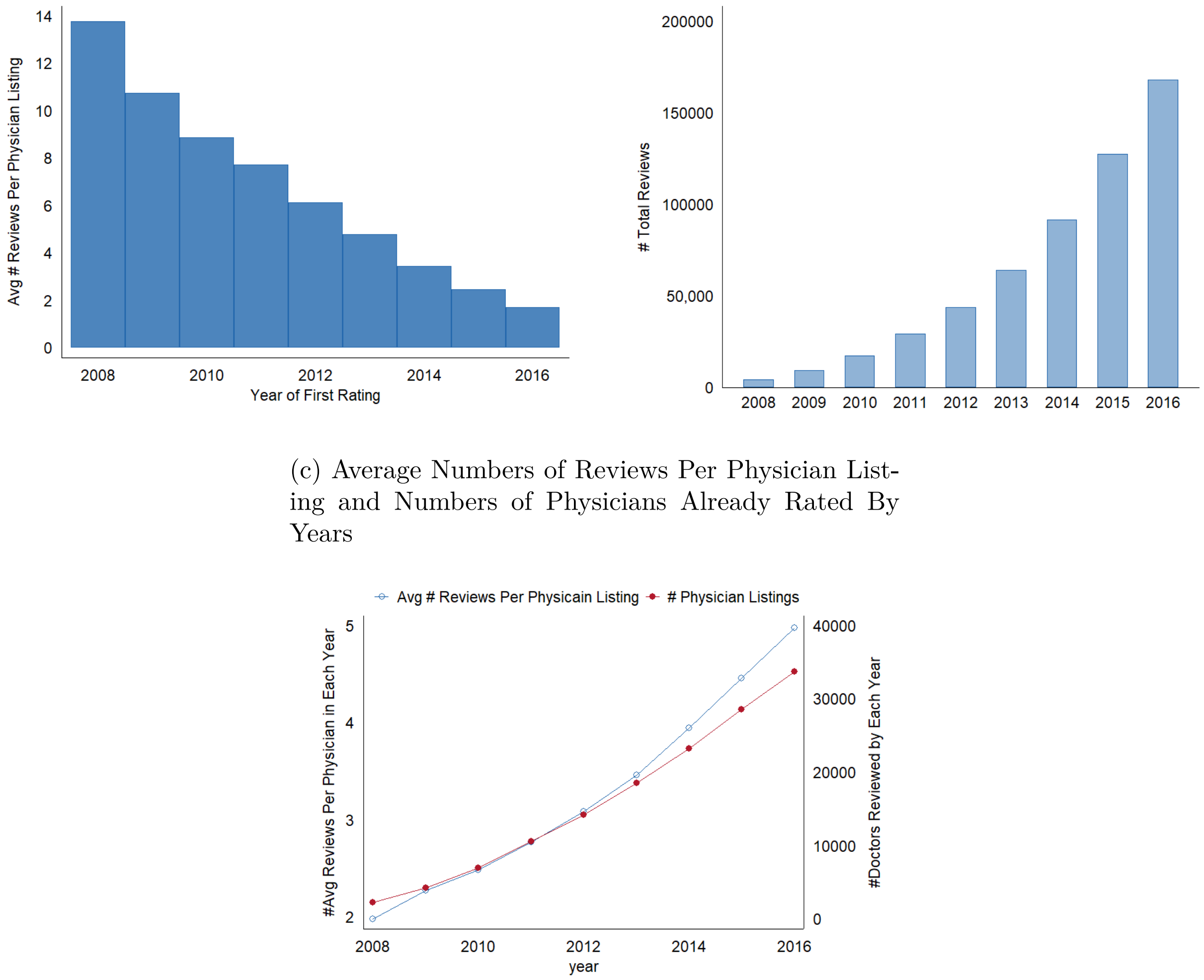
2.3.2 Yelp Ratings and Reviews

From the main physician sample, Panel (a) of Figure 1 plots the numbers of reviews per physician listings in June 2017 by the date (year) of the first rating. For example, for physicians who were first rated in 2008, the number of reviews on average was about 14. In contrast, for physicians who first received a rating in 2016, their average number of review was about 2 for the year 2017. On average, physician reviews were sparse, with an average number of review equaling 5.3, but physicians who received ratings earlier start to gain more reviews over time. This can also be seen in Panel (b) of Figure 1, which shows that the total numbers of cumulative Yelp physician reviews in the sample grew rapidly over years. In Panel (c), one can see that both the numbers of reviews per physicians and the numbers of physicians reviewed also increased quickly.

Panel (a) of Figure 2 plots the histogram of cumulative average Yelp ratings of each physician listing in the main rated physician sample by June 2017, rounded to the closest 0.5 star. The average rating at the physician listing level was 3.62. Panel (b) contains the histogram of Yelp ratings at the review level for both Yelp physician ratings in the main sample and all Yelp businesses ratings. Reviewers tend to give 5 or 1 stars for either physicians or other Yelp businesses.[18] The average rating at the individual review level for physicians was 3.79, compared to 3.71 for all businesses on Yelp. The fact that these figures somewhat resemble each other suggests that reviewers possibly rate similarly for physicians and other businesses on Yelp.

[18] Calculated from https://www.yelp.com/factsheet.

Figure 1: Average Numbers of Reviews Per Physician Listing By Years of First Rating



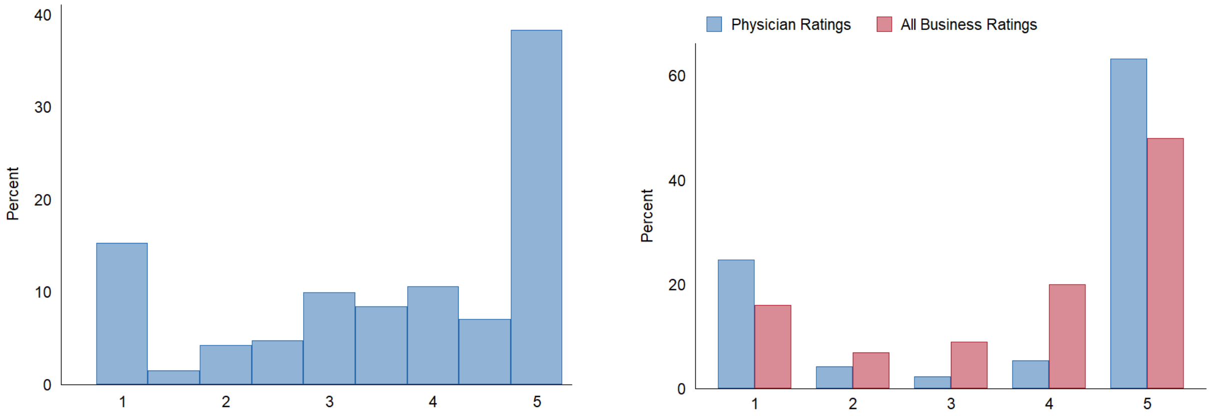
(a) Average Number of Reviews Per Physician Listing By Years

(b) Numbers of Total Reviews By Years

(c) Average Numbers of Reviews Per Physician Listing and Numbers of Physicians Already Rated By Years

Notes: Panel (a) shows the numbers of reviews per physician listings by June 2017 (on the y-axis) against the first year with recorded ratings (on the x-axis). For example, the first column shows the number of reviews per physician listings for physicians who were first rated in 2008. On average, among the main rated physician sample, each physician had 5.3 reviews up to June 2017. The figure in Panel (b) plots the numbers of cumulative Yelp physician reviews in the main rated physician sample in the U.S. by year In Panel (c), the blue line shows the average number of reviews per physician listings (on the y-axis on the left) among physicians who were already rated by year (on the x-axis). The red line plots the total number of physicians already reviewed (on the y-axis on the right) by year.

Figure 2: Distribution of Yelp Ratings



(a) Distribution of Physician Average Ratings

(b) Distribution of Individual Review Ratings

Notes: Panel (a) contains the histogram of cumulative average Yelp ratings of each physician listing in the main rated physician sample in June 2017. Each average rating is rounded to the closest 0.5 star. The average rating at the physician listing level was 3.62. Panel (b) contains the histogram of Yelp ratings for each review of the physician listings from the main rated physician sample in June 2017 in blue and the histogram of each Yelp review for all Yelp businesses in August 2018 in red, downloaded from https://www.yelp.com/factsheet. The average rating at the individual review level for physicians was 3.79. The average rating at the individual review level for all businesses on Yelp was 3.71.

3 What Quality Information Do Yelp Ratings Convey?

3.1 Contents of Yelp Reviews

First, I analyze the texts of Yelp reviews to understand what information are written by reviewers. As a method to understand the common themes of Yelp reviews and to reduce review information dimensionality, I use a machine-learning algorithm (latent Dirichlet allocation) to automatically categorize the texts of reviews into a small set of "topics," which one may intuitively interpret as a cluster of keywords that tend to show up together in a review. Technically, in the algorithm, a review is regarded as an unordered word collection generated by a small number of topics. Each topic is a distinct probability distribution that generates keywords in such a way that a small set of keywords will be drawn frequently. After pre-specifying the total number of possible topics, the algorithm reads in the keyword distributions from all reviews and uses a Bayesian algorithm to infer the distribution of topics among reviews and keyword distributions among topics. The details of the algorithm specification I use can be found in Appendix B.1.

Using all of my collected Yelp reviews for physician listings with suffixes "MD," "DO," or "OD" to represent individual physicians, the algorithm shows the following top 10 topics, their top relevant keywords within each topic, and my subjective interpretation in Table 1, assuming 20 possible topics in total. The topics are ranked by the probability of a review belonging to a topic. The top relevant keywords are the words with highest adjusted probability within each topic.[19] Then, I subjectively offer an interpretation for each topic. From the table, most of the topics reflect a physician's interpersonal skills and office amenities. For example, Topic 2 is associated with keywords such as "question," "feel," "time," "always," "answer," "concern," etc. It seems that users are describing a physician with strong empathy in this topic, indicating the physician's superior interpersonal skills. I find similar results when performing the same analysis for reviews of high ratings (higher than or equal to 4 stars) and those of low ratings (lower than or equal to 2 stars) in Appendix Tables B.1 and B.2.

Second, to understand how human readers may interpret Yelp reviews, I randomly select 1,500 reviews from all my collected Yelp reviews for physician listings with suffixes "MD," "DO," or "OD" and run a survey that recruits human readers to categorize them. Using Amazon Mechanical Turk, I hired two human readers per review and asked them to classify a review into "service quality related" (e.g., friendliness, attitude, and amenity), "clinical quality related" (e.g., diagnosis, treatment, prescription, and outcome), "both of the above,"

[19] The probability of a keyword within a topic is adjusted by the algorithm in Sievert & Shirley (2014) to penalize a keyword if it appears too often across all reviews. See Appendix B for details.

Table 1: Top 10 Topics of Reviews from Yelp

Topic	Probability	Avg Rating	Top Relevant Keywords	Subjective Interpretation
1	11.6%	3.9	go, like, doctor, really, know, just, can, say, get, see	generic
2	11.3%	4.4	question, time, concern, answer, patient, feel, listen, care, explain, take	office amenities, interpersonal skills
3	10.6%	4.7	staff, friendly, office, great, recommend, profession, highly, love, help, nice	office amenities, interpersonal skills
4	8.4%	2.5	call, appoint, office, phone, get, back, told, day, said, ask	office amenities, interpersonal skills
5	8.2%	4.4	care, year, physician, patient, doctor, family, primary, medic, recommend, many	generic
6	7.4%	3.0	test, went, told, said, blood, came, ask, saw, doctor, go	clinic related
7	6.1%	3.0	wait, minute, time, hour, room, appoint, long, late, min	office amenities, interpersonal skills
8	5.6%	4.4	surgery, procedure, surgeon, consult, result, done, plastic, perform, went, lip	clinic related
9	5.1%	2.1	front, rude, desk, office, staff, worst, horrible, doctor, unprofessional, custom	office amenities, interpersonal skills
10	3.7%	3.3	review, star, yelp, read, write, negative, give, rate, experience	generic

Notes: The sample includes 231,215 reviews for all physicians with suffixes “MD,” “OD,” or “DO." The model assumes that there are in total 20 topics and runs through the LDA algorithms 200 times. Topic numbers are ranked by the probability that a Yelp review in the sample is classified according to each topic. “Avg rating” refers to the weighted average of each review rating over all reviews, weighted by the probability of the focal topic belonging to each review. Top relevant words are derived using the formula and modules provided by Sievert & Shirley (2014), setting $\lambda = 2/3$, which is the weight balancing a keyword’s probability in a topic, and its probability in a topic divided by the overall probability of the keyword in all usage. See Appendix B for details. Subjective interpretation consists of my personal interpretation of the keywords of each topic.

or “other.” Figure B.1 describes the survey questions. Among the 781 out of 1,500 reviews that receive the same answers from both human readers, the respondents find 81% of reviews describing service quality and 44% of reviews describing clinical quality.[20] The results are consistent with previous machine-learning results indicating more reviews are describing a physician’s service quality, although there is also a significant amount of information that the human readers interpret as clinically related. Perhaps the reviewers just do not write clusters of clinical jargon, and thus the machine-learning algorithm fails to recognize those information. More details about the survey results can be found in Appendix B.2.

3.2 Correlations between Ratings and Physician Clinical Quality

If on average, a physician’s interpersonal skills are negatively correlated with a physician’s abilities in delivering effective clinical outcome, many patients who value clinical outcomes may suffer from visiting a high-rated physician. For example, Dr. House from the popular TV show *“House M.D.”* may easily receive low star ratings because of the character’s cynical attitude toward his patients, but many people are amiss to avoid him due to his extraordinary clinical ability.[21] This section attempts to infer whether, on average, physician ratings are positively or negatively correlated with physician clinical quality.

The first set of clinical quality indicators I use is physicians’ educational and professional credentials—board certification status, ranks of medical schools, and number of self-reported accreditations. Although one may argue that those pieces of information are readily available outside of Yelp, if there are positive correlations between Yelp ratings and those quality measures, one may have more confidence that Yelp ratings actually convey some unobserved physician clinical quality as well. I estimate the following physician j level regression using all physicians from the main rated physician sample:

$$y_j = \alpha R_j + HSA_j + Specialty_j + \epsilon_j, \tag{1}$$

where y_j represents a physician’s educational and professional credentials, R_j denotes the latest cumulative average physician rating by June 2017, and physician HSA and specialties fixed effects are also included on the right hand side. The results are displayed in columns 1–4 of Table 2. 73% of physicians in the sample are board-certified. Column 1 illustrates that

[20] If the answer is “both above,” I count it as both service and clinical quality in the percentage calculation. The option of “both above” also explains why only a third of reviews are mutually agreed upon. Many survey respondents would only find a review to be either “clinical quality related” or “service quality related” while others find it “both of the above.”

[21] Dr. Gregory House from *“House M.D.,”* a popular American TV show by Fox. https://www.fox.com/house/

a change from 1 to 5 stars in physician rating is associated with +12 percentage points in a physician's board certification probability. Physicians' average medical school rankings are 18 and 59 from Usnews and StartClass, respectively. From columns 2 and 3, a change from 1 to 5 stars in physician rating is associated with +2 and +7 in better rankings for Usnews and StartClass (the order is reversed so that the higher rankings the better). Column 4 shows that a change from 1 to 5 stars in physician rating is associated with +11% in the number of self-reported quality indicators. To benchmark how large these effects are, I also include a physician's medical school ranking from StartClass on the right-hand side as an additional control variable. Columns 1 and 2 in Table B.3 show that the effects of 1-5 star changes on various dependent variables do not diminish compared to before controlling for the additional regressor. The coefficients of ratings are also much bigger than those of medical school rankings. These findings suggest that Yelp ratings provide positive information on physician credentials beyond what simple observables such as physician school rankings would predict.

Next, I investigate whether Yelp ratings convey patient health outcome information. I examine whether the patients of higher-rated primary care physicians have better primary care related health quality. I link patient i to her mostly frequently visited primary care physician $j(i,t)$ in year t to let the physician in "charge" of patient i's health. Among all the patients of physicians in the main rated physician sample, I estimate the following patient(i)-year(t) level regression using the Medicare Part B claims data from 2008 to 2015:

$$y_{it} = \alpha R_{j(it)} + X'_{it}\gamma + \epsilon_{it}, \tag{2}$$

where $j(i,t)$ is patient i's primary care physician in year t; $R_{j(it)}$ represents the latest cumulative average Yelp ratings for each listing by June 2017 for a patient i's primary care physician j in year t; X_{it} denotes patient characteristics including age effects up to cube terms, gender, race, risk scores of the previous year, and HSA-year fixed effects; and y_{it} includes a variety of patient health outcomes that primary care physicians may affect: whether the patients receive recommended eye exams and mammograms if eligible, whether they receive preventable inpatient admissions, and health risk scores such as CMS-HCC risk scores and Charlson comorbidity index.

The estimation results are displayed in columns 5–9 of Table 2. In column 5, I find that a change from 1 to 5 stars in physician rating is associated with +0.008 in probability in receiving a recommended eye exam, which is 1.5% of the mean value of the dependent variable 0.52. In column 6, a change from 1 to 5 stars in physician rating is associated with +0.025 in probability receiving recommended mammograms, which is 3.7% of the mean value of the dependent variable 0.67. In column 7, a change from 1 to 5 stars in physician rating in physician ratings is associated with -0.0036 in probability of a patient receiving

Table 2: Correlations between Yelp Ratings and Physician Clinical Ability

	(1)	(2)	(3)	(4)
Dep Variables:	Board	Usnews	Startclass	Log(#Accreditations)
Ratings	0.0299***	0.581***	1.736***	0.0280***
	(0.00502)	(0.116)	(0.355)	(0.00836)
Observations	8,755	36,346	36,346	4,405
R-squared	0.192	0.165	0.221	0.293
Mean(LHS)	0.73	18	59	0.98

	(5)	(6)	(7)	(8)	(9)
Dep Variables:	Eye Exam	Mammogram	PQI	Risk Score (Charlson)	Risk Score (CMS)
Ratings	0.00195**	0.00622***	-0.000866***	-0.0137***	-0.00826***
	(0.000787)	(0.00112)	(0.000245)	(0.00280)	(0.00173)
Observations	810,464	751,746	3,013,423	2,891,537	2,891,537
R-squared	0.057	0.066	0.056	0.411	0.386
Mean(LHS)	0.52	0.67	0.04	1.81	1.26

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above contains the estimation results of equations (1) in columns 1–4 and (2) in columns 5–9. The estimation sample in column 1 includes the Healthgrades data for primary care physicians matched with the main rated physician sample using a physician's last name, first name, and HSA. The dependent variable in column 1 is an indicator variable for whether physician j is board certified. In columns 2 and 3, the sample includes Physician Compare demographic data from 2015 matched with the main rated physician sample. A physician's ranking of medical school is obtained by manually matching her medical school information from Physician Compare with rankings from Usnews or StartClass. I reverse the order so that the best schools receive the highest number and unranked schools are input as 0. In column 4, the sample includes Physician Compare performance data from 2015 matched with the main rated physician sample using NPI numbers. The dependent variable is the log of the number of self-reported quality indicators of physician j. Columns 1–4 are estimated at the physician level, include physician specialty and HSA fixed effects, and are two-way clustered at the specialty and HSA levels. Columns 5–9 include the main rated physician sample linked with Medicare Part B non-institutional claims data between 2008 and 2015. The dependent variables for columns 5 and 6 are whether a patient receives eye exams and mammograms among eligible diabetic patients and female patients younger than 74. The dependent variable for column 7 is an indicator of whether patient i receives preventable inpatient admissions in year t, defined by the numerators of the PQI index from AHRQ. The dependent variables for column 8 and 9 are the computed Charlson and CMS-HCC-2015 risk scores using a patient's medical history up to year t. Columns 5–9 are estimated at the patient-year level, include physician specialty and HSA fixed effects, and are clustered at the physicians' HSA level.

a preventable inpatient admission, which is 8.7% of the mean value 0.04. Using Charlson and CMS-HCC risk scores as health outcome measures, columns 8 and 9 show that changes from 1 to 5 stars in physician rating are associated with -0.06 and -0.04 decrease in risk scores after controlling for the previous year risk scores, which are 2.8% and 3.2% of the mean value of the dependent variable, implying better health condition developments and potentially slower developments in comorbidities.

All the above findings suggest that there are small but positive and meaningful correlations between physicians' Yelp ratings and various conventional measures of clinical quality—physicians' educational and professional backgrounds, their adherence to clinical guidelines, and their patients' risk-adjusted outcomes. One may worry that these measures may reflect patient selection rather than true clinical abilities of physicians. This is perhaps unlikely the case for educational and professional backgrounds and less likely for physicians' adherence to clinical guidelines in ordering exams. The clinical quality measures using patient health outcome are of most concern. One thing to notice is that R_j from equation (2) is physicians' average ratings in 2017. For many of the observations in the estimation sample, patients would not be able to observe these ratings yet since many physicians were not rated until 2017. This fact potentially removes much of the patient selection due to observing Yelp ratings, and the remaining selection is partly controlled by patients' characteristics X_{it}. With these caveats in mind, to the extent that these clinical measures reflect physicians' clinical abilities instead of patient selection, patients from visiting higher-rated physicians will be matched with physicians with better clinical quality. To benchmark how large these positive correlations are, in columns 3-7 of Table B.3, I again include a physician's StartClass medical school ranking as an additional regressor and find that the associations of patient health outcomes and ratings do not diminish and are larger in magnitude than the control variable. These results imply that the correlations remain strong beyond what other sources would predict.

Overall, the evidence suggests that the texts of Yelp reviews are written mainly on reviewers' satisfactions with their physicians' interpersonal skills. However, physicians' Yelp ratings are also positively correlated with their medical credentials, adherence to clinical guidelines, and patients' risk-adjusted health conditions, all of which possibly indicate better clinical abilities. Although there are many quality dimensions related to healthcare, Yelp online ratings seem to signal better physician quality in multiple dimensions.

4 The Impact of Yelp Ratings on Physicians' Patient Flow

This section investigates how much different Yelp ratings causally affect physicians' annual patient flow differently, measured in either revenue or patient volume. In other words, the object is to obtain the "treatment effect" of higher versus lower ratings—if one randomly assign different levels of Yelp average ratings to physicians, how much physicians receiving higher ratings grow faster in revenue and patient volume than those receiving lower ones.

4.1 OLS Strategy

I first approach the treatment effect using the following OLS framework. Among all physicians—whether rated or not from Medicare—the following panel regression is specified at the physician j and year t level using the physician payment data from 2012 to 2015:

$$y_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \alpha D_{jt} + \beta R_{jt} D_{jt} + \epsilon_{j,t}. \tag{3}$$

In the specification, y_{jt} is a physician's annual revenue or patient volume. χ_j represents a physician fixed effect that controls for physician time-unvarying characteristics. $\theta_{s,t}$ and $\theta_{h,t}$ denote flexible year fixed effects that are specific to physician j's specialty $s(j)$ and HSA $h(j)$, capturing time trends of physician specialties and locations. They are identified from physicians who are ever rated as well as from those who are never rated. D_{jt} is an indicator that is 1 if year t is equal or after physician j's first rating year and 0 otherwise, capturing the aggregate factors between being rated and being non-rated. It is identified from physicians who change from not being rated to being rated. Conditional on a physician is now rated by year t, R_{jt} represents the cumulative average Yelp rating of physician j by the end of year t, which mimics what readers see online in year t. It is normalized to mean 0 in the regression. β is the key coefficient of interest and captures how different levels of R affect patient volumes differently. It is identified from two sources. First, suppose that ratings, if rated, are constant: that is, $R_{jt} = R_j$. Ignoring θ_t, β captures whether treatment physicians j who are rated higher grow faster in patient flow (their patient flow after rating is $\chi_j + \alpha + \beta R^{high}$) than the control physicians j' who are rated lower ($\chi_{j'} + \alpha + \beta R^{low}$), compared to before they receive ratings (χ_j and $\chi_{j'}$). Second, since a physician's rating actually evolves over time, β is also identified from how different ratings R_{jt} correlate with the physicians' different rates of patient flow.

The estimation results using OLS are displayed in columns 1 and 3 of Table 3. In column 1, the left-hand side is the log of a physician's annual revenue. The coefficients mean that,

Table 3: Regression Results for Equation (3)—Effects of Yelp Ratings on Patient Flow

	(1)	(2)	(3)	(4)
Dep Variables:	Log(Revenue)	Log(Revenue)	Log(#Unq Patients)	Log(#Unq Patients)
Method	OLS	IV	OLS	IV
$R_{jt}D_{jt}$	0.0125***	0.0186***	0.00762***	0.0121**
	(0.00208)	(0.00705)	(0.00157)	(0.00477)
D_{jt}	-0.0125**	-0.0116*	-0.00786*	-0.00715*
	(0.00559)	(0.00595)	(0.00423)	(0.00430)
Observations	3,474,085	3,474,085	3,474,085	3,474,085
R-squared	0.914	0.914	0.923	0.923

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation results of equation (3). The sample combines the main rated physician sample with the Medicare physician payment data from 2012 to 2015 and includes all physicians with positive payments. Columns 1 and 2 use the log of a physician's total revenue in a year as the dependent variable. Columns 3 and 4 use the log of a physician's number of unique patients as the dependent variable. Columns 1 and 3 use OLS estimation. Columns 2 and 4 use instrumental variable estimation. In the first stage regression putting $R_{jt}D_{jt}$ as the left-hand-side variable and $z_{jt}D_{jt}$, physicians' "harshness" interacted with the indicator whether the rating exists, the coefficient is 0.570 with an standard error 0.0163. All the regressions include physician fixed effects, physician specialty-specific time fixed effects, and physician HSA-specific time fixed effects. Standard errors are two-way clustered at the physicians' HSA and specialty levels.

once a physician has been rated, with an average rating (3.65 star), her revenue would decrease by 1.3% ($\hat{\alpha}$). For every extra star over the average rating, the physician would gain 1.3% more in revenue ($\hat{\beta}$). Similarly, in column 3, the left-hand side is the log of a physician's annual number of unique patients, or the patient volume. After being rated, with an average rating, patient volume would decrease by 0.8%, but every extra star over the average rating would gain 0.8% in patient volume. One way to think about the results is that patients may regard the average rating—3.65 stars—as a signal of below-average physician quality in the overall market. Without near perfect ratings, physicians' patient flow would decrease after being rated; and the lower the ratings, the lower the patient flow would become.

The OLS results depict a time relationship before and after physicians receive ratings with different levels. If the ratings are uncorrelated with physicians' other time-varying quality that may affect patient flow, and cumulative end-of-year average ratings R_{jt} accurately measure the ratings a patient observes before visiting a physician,β would pick up the average treatment effect of receiving rating R rather than $R-1$ on a physician's patient flow.

In reality, ratings are not randomly assigned and the end-of-year cumulative average ratings R_{jt} do not perfectly measure what patients observe, causing two endogeneity issues. First, physicians' time-varying characteristics, such as their changes of inner ability, quality, or budget, may co-determine their likelihood of receiving high or low ratings and new patient flow. For example, physicians may have already been improving their office amenities and thus have better chances receiving higher ratings. However, the improvement in office amenities will also directly improve patient flow itself. In another example, a physician may have decided to spend her budget on marketing efforts but does not have enough staffing capacity to accompany the increase in patient flow. This may result in more chances receiving bad ratings but will improve the physician's patient flow overall. *Ex ante*, the bias is unclear for β. Second, the ideal treatment variable is the Yelp ratings that patients actually observe before they visit their physicians. The yearly cumulative end-of-year average ratings R_{jt} are not the ideal treatment ratings because a patient may visit her physician before some of the new ratings arrive by the end of the year. In a sense, R_{jt} contain noises—the new ratings—for the ideal treatment ratings. From the point of view of a classical measurement error, β is biased toward zero.

4.2 Instrumental Variable Strategy

In the second strategy, I use an instrumental variable approach to tackle the above concerns. The intuition is that some reviewers may be nicer in reviewing any businesses while others are harsher. Since physicians have different lucks in getting nicer or harsher reviewers, their reviewers' "harshness" creates variations in their Yelp ratings while also being plausibly orthogonal to the endogenous factors—physicians' time-varying quality that co-determines the likelihood of receiving high or low ratings and patient flow, as well as measurement errors. In the same fashion of cumulative ratings R_{jt}, I construct a physician's cumulative reviewer "harshness" as the instrument. Let n_{jt} denotes the number of reviewers for physician j by year t. Let $k \in K(jt)$ be a reviewer who reviewed physician j by the end of year t. Let r^k_{-j} be the average rating from k's reviewed businesses excluding physician j. Then, the cumulative average reviewer "harshness" of physician j by year t is measured as follows:

$$z_{jt} = \frac{1}{n_{jt}} \sum_{k \in K(jt)} r^k_{-j}. \tag{4}$$

In equation (3), I instrument the endogenous variable $R_{jt}D_{jt}$ with $z_{jt}D_{jt}$.

The inclusion restriction of the instrument is that the average yearly cumulative "harshness" affects a physician's yearly cumulative Yelp rating. The exclusion restriction assumes that observed physicians' reviewer "harshness" z_{jt} should not be correlated with other time-varying factors that co-determine likelihood to receive high or low ratings and patient flow. However, since the observed "harshness" are not assigned to a physician randomly, the assumption may be invalidated. For example, if "grumpier" reviewers are very choosy and tend to gravitate toward physicians who have unobserved high inner abilities that co-determine ratings received and patient flow. Or "harsher" reviewers would more likely to post a bad review when the unobserved physician quality is low, so the observed reviewer "harshness" becomes correlated with the unobserved physician quality.

The main results of the instrumental variable estimation from equation 3 are shown in columns 2 and 4 of Table 3. In the first stage regression not shown in the columns, a 1-star increase in average reviewer "harshness" increases a physician's average rating by 0.57 stars. Columns 2 and 4 contain the second stage results for the log of a physician's total revenue in a year and the log of a physician's number of unique patients in a year as the dependent variables. In column 2, with an instrumental variable (IV) estimation, a physician after being rated for an average rating is associated with a -1.2% decrease in revenue, but a 1-star increase in ratings is associated with a 1.9% increase in a physician's annual revenue. In column 4, with an IV estimation, a physician after being rated for an average rating is associated with a -0.7% decrease in patient flow, but a 1-star increase in physician's rating is associated with a 1.2% increase in a physician's number of unique patients. In general, $\hat{\alpha}$, the aggregated differences between before and after being rated, are similar to the OLS estimations: they are negative, and for the average rated physicians (3.65 stars), their patient flow would decrease after being rated. The IV estimates of β, the differential effects of ratings, are generally positive and slightly larger than the OLS estimates: overall, they suggest that a 1-star increase in Yelp ratings is associated with a 1-2% increase in a physician's revenue and patient volume. Assuming that $\hat{\beta}$s using IV estimations are the true "causal" treatment effects of differential ratings to physicians, one may interpret them in at least two ways—that customers value physician interpersonal skills and react to such signals, or customers are using Yelp ratings as an easy proxy for good clinical and overall quality. I cannot differentiate between these two hypotheses and would leave the question for future research.

To gauge the possibility of magnitude of $\hat{\beta}$, I conduct the following back-of-envelope calculation. In the Medicare claims data, I find that 35% (if only considering primary care physicians) or 43% (if considering all physicians) of the patients of the main rated physician sample are new patients in a year: those patients have never encountered their physicians before. Assuming that only new patients use Yelp ratings, since the overall response rate of revenue and volume per star increase is about 1–2%, the implied response rate per Yelp star increase for new elderly customers becomes about $(1\% + 2\%)/(35\% + 43\%) = 3.8\%$. I assume that 10% of elderly new consumers would consult Yelp when looking for physicians either directly by searching on Yelp or indirectly by Googling a potential physician's name and looking up results on Yelp, which is in line with survey findings.[22] The resulted implied response rate per Yelp star increase among new elderly customers using Yelp becomes $3.8\%/10\% = 38\%$. In comparison, Luca (2016) found a 5–9% increase in restaurant revenue per Yelp star increase. The new customer share for restaurants is about 35–49%.[23] If one divides the average of 5–9% by the average of 35–49%, the implied response rate per new consumer per star increase is about 16.5%. If one further assumes that about 75% of new restaurant consumers use Yelp when looking for restaurants, the implied response rate per Yelp star increase among new restaurant customers using Yelp becomes about 22%, not too far off from the back-of-envelope physician rating estimate.

4.3 Suggestive Tests of Exclusion Restriction

As a suggestive check to the exclusion restriction assumption of the IV strategy and as a reduced form estimation, I estimate an event study showing how physicians who receive a general level of high or low "harshness" compare in terms of patient flow before and after receiving their first ratings. The event study can test whether there are differential trend between the two groups of physicians before they receive any ratings. If there are no differential trends, it would give us confidence that physicians' reviewer "harshness" is not correlated with other time-varying factors that affect patient flow before being rated. Following this intuition, using the same estimation sample of equation (3), I estimate the following event study regression at the physician j year t level among all physicians whether rated or not from Medicare payment data between 2012-2015, defining the event as being first rated with high or low first-year reviewer "harshness" instrument z_j^f (High value instruments are less "harsh."):

[22] Internet survey by Software Advice found that 42% of online Internet users had used online physician ratings by 2014. In 2013, a similar survey found that the elderly were among the high utilization age group of online physician ratings. And Internet usage among elderly during that time was not low. According to Pew Research Center, 47% of elderly owned a home broadband service back to 2013 and 67% of them owned one in 2016. In addition, adult children could have been actively helping their parents in medical decisions. A study in 2014 found that 47% of seniors had some surrogate involvements in making medical decisions within 48 hours of hospitalizations Torke et al. (2014). These findings suggest that maybe a 10% usage rate among new elderly customers of physicians is possible.

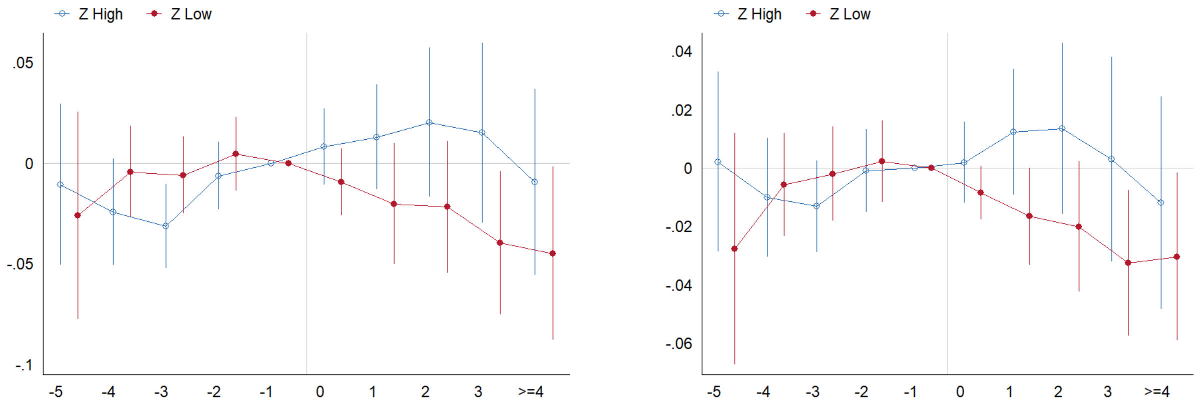
[23] http://www.restaurant.org/News-Research/News/Repeat-customers-integral-to-success,-study-finds

$$y_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \sum_k \lambda_k 1(t - d_j^f = k) + \sum_{g \in \{h,l\}} \sum_k \gamma_k^g 1(t - d_j^f = k) z_j^g + \epsilon_{jt}. \qquad (5)$$

z_j^h is defined as $z_j^h = 1(z_j^f > 4.35)$, the first-year instrument is above 67th percentile and thus less harsh; z_j^l is defined as $z_j^l = 1(z_j^l < 3.67)$, the first-year instrument is below 33th percentile and thus more harsh. d_j^f is the first year of physician j being first rated. Among physicians who are ever rated, λ_k captures the main effect in the kth year since their first ratings. Alternatively, one can interpret it as the trend of physicians who receives a medium "harshness." γ_k^h and γ_k^l test whether there are differential trends in the kth year since being first rated among physicians who have high and low first-year instruments z_j^h and z_j^l, differently from the common trend λ_k. If the exclusion restriction assumption is valid, it implies that $\gamma_k^h = \gamma_k^l = 0$ for $k < 0$, that is, physicians receiving different levels of first-year "harshness" would not differ in their pre-trend patient flow. The estimated coefficients $\hat{\gamma_k^h}$ and $\hat{\gamma_k^l}$ from (5) are displayed in Figure (3), with the log of the revenue on the left-hand side in Panel (a) and the log of the number of unique patients on the left-hand side in Panel (b). The findings are generally consistent with the exclusion restriction assumption that the pre-trend patient flow does not differ according to high or low first-year-reviewer "harshness" ($\hat{\gamma}_{k<0}^h = \hat{\gamma}_{k<0}^l = 0$). In the post-rating period, it also seems that physicians receiving higher z_j^f, thus less "harsh" reviewers, generally experience higher patient flow (blue lines) than those with lower z_j^f, thus more "harsh" reviews (red lines). The results preview the existence of a treatment effect of ratings, if reviewer "harshness" only affects patient flow through affecting ratings.

I raise three additional arguments that may strengthen the plausibility of the exclusion restriction assumption. First, in Appendix C.1.1, I construct an alternative instrument in a smaller sample that does not depend on a reviewer's ratings on healthcare businesses at all and even residualizes the impact of baseline business ratings. If a reviewer gives a 5-star to an average 4-star restaurant, conceptually I only use the 5-4=1 star residual "harshness" measurement when constructing the instruments. If one worries that a high quality physician is attracting reviewers who like to visit and rate a high quality businesses such as 4-star restaurants, the residualization removes such baseline attraction and only uses reviewers' idiosyncratic residual "harshness" that is not dependent on the average quality of the rated businesses. It is reassuring that the IV results are similar. Second, a suggestive test of the exclusion restriction assumption is to check whether the instrument z_{jt} is correlated with observable physician characteristics related to inner ability. Weak correlations would give one more confidence that the instrument is uncorrelated with unobservable characteristics on

Figure 3: Estimation Results of Equation (5)—Event Study of Patient Flow by First-Year Reviewer "Harshness"



(a) γ_k^h (Blue) and γ_k^l (Red) Using Log of Revenue as the Dependent Variable

(b) γ_k^h (Blue) and γ_k^l (Red) Using Log of # Unique Patients as the Dependent Variable

Notes: The figures above show the estimation results of equation (5). The sample combines the main rated physician sample with the Medicare physician payment data from 2012 to 2015 and include all physicians with positive payments. Panel (a) uses the log of a physician's revenue on the left-hand side and Panel (b) uses the log of a physician's number of unique patients on the left-hand side. The x-axis corresponds to k. The right-hand side pretends that physicians always receive the first-year "harshness" z_j^f after being rated. The red solid circles denote $\hat{\gamma_k^l}$, the differential trends of receiving "harsher" first-year reviewers z_j^f in each year with respect to the first year of being rated ($k = 0$) compared to a physician with a medium z_j^f; the blue hollow circles represent $\hat{\gamma_k^h}$, the differential trends of receiving less "harsher" first-year reviewers z_j^f in each year with respect to the first year of being rated ($k = 0$) compared to a physician with a medium z_j^f. The 95% confidence intervals are plotted in lines on the y-axis. γ_{-1}^h and γ_{-1}^l are normalized to 0. The regressions include physician fixed effects, physician specialty-specific time fixed effects, physician HSA-specific time fixed effects, as well as fixed effects for years since being first rated (normalizing them to be 0 for physicians without any ratings). Standard errors are two-way clustered at the physicians' HSA and specialty levels.

physicians' inner ability as well. Appendix C.1.2 explores some possible correlations between observable time-varying and unvarying physician characteristics and the instruments, such as physicians' educational and professional backgrounds and their annual adherence to the HEDIS guidelines. And I do not find significant correlations. Third, I offer some suggestive evidence that whether physician j is rated in year t, D_{jt}, is possibly exogenous as a control variable. Empirically, among physicians who are ever rated, D_{jt}, is mostly uncorrelated with both time-varying and unvarying observed physician quality characteristics, controlling for the physician fixed effects χ_j and the flexible time fixed effects $\theta_{t,s}$ and $\theta_{t,h}$, which gives one more confidence that it may also be uncorrelated with other unobserved time-varying physician quality characteristics. One interpretation is that a physician's chance of being rated is mostly determined by Yelp's regional popularity over time rather than her own quality. The details can be found in Appendix C.1.3.

I also run a series of robustness checks and heterogeneity exercises. To assess the robustness of the measurement of ratings, for Appendix C.3, I perform a similar estimation measuring ratings and instruments by the end of previous year, as they may contain fewer measurement errors since all Yelp readers will have read the previous year ratings. I find qualitatively similar and slightly larger results. In heterogeneity exercises, in Appendix C.2, I find that the response rate per star increase in ratings is larger for physicians with more reviews, smaller for physicians with older patient pools, and larger for physicians within more educated areas. All of those results are consistent with the hypothesis that patients are responding to Yelp ratings in choosing physicians, as they would trust ratings more when there are more reviews, and younger and more educated patients may use Internet tools more often.

5 Physician Responses to Being Rated

5.1 Empirical Strategy

This section empirically tests whether after being rated, physicians order more clinical services in total spending, lab and imaging spending, and opioid prescriptions and potentially cause changes of patient health outcomes. For the following reasons, I consider the date that a physician receives her first rating as the date that Yelp ratings become a significant incentive for her to change practice patterns. First, some reviewers set up the physician profiles first, reducing the costs to future reviewers who want to rate them. Second, since both Yelp's and Google's algorithms rank rated physician higher than unrated ones, being rated improves a physician's salience on the Internet, encouraging future reviewers to rate the physician and future readers to read her profile. Third, as Yelp currently only hosts ratings for a very small number of physicians, many physicians may have thought that they would never receive ratings at all. Receiving their first ratings may change their perceptions. Anecdotally, physicians are very aware of their own ratings and heated discussions exist in the profession on how to react to online ratings.[24]

The empirical strategy employed is a difference-in-differences framework. Consider a cohort m of patients: for primary care physicians first rated in year $m \leq 2015$ (treatment), versus those in 2016/2017 (control), among their "pre-existing before-m" patients, I compare health services and health outcomes before and after m. The "pre-existing before-m patients" consist of those patients who first associate with their current primary care physician before year m. I focus on those patients to isolate the effort changes of physicians rather than the new patient mix changes, since the new patients will also observe the ratings and may be different from the pre-existing patients. If $m = 2013$, for example, I compare physicians who received their first ratings in 2013 versus physicians who were first rated in 2016/2017 and test whether the pre-existing patients in the first group of physicians received more wasteful and harmful treatments before and after 2013 than the second group. Pooling all cohorts of patients $m \in 2009...2015$, I specify the following regression at the cohort(m)-patient(i)-year(t) level using Medicare claims data from 2008 to 2015:

$$y_{it}^m = \chi_{ij} + \theta_{s,t}^m + \theta_{h,t}^m + \sum_k \omega_k 1(t - m = k) T_j^m + \epsilon_{it}^m. \tag{6}$$

In the specification, y_{it}^m is the health service or health outcome of patient i in year t, who is of cohort m. j stands for $j(i,t)$ and is a patient i's primary care physician in year t. χ_{ij} represents a patent i-physician j pair fixed effects. It is a constant if i stays with physician j. However, if i switches to physician j' in some year, a new fixed effect is generated for the new relationship. The flexible patient—physician fixed effects eliminate much of the endogenous patient—physician's matching. $\theta_{s,t}^m$ and $\theta_{h,t}^m$ denote time fixed effects that are physician j's specialty $s(j)$ and practice HSA $h(j)$ specific. They are further cohort m specific and are identified from the control patients. T_j^m relates to whether physician j is in the treatment group of cohort m—being first rated in year m as opposed to 2016/2017. ω_k captures how the treatment patients differ from the control patients in the kth year after the treatment physicians receive their first rating in year m. As the primary care physicians in the control group receive their first ratings on Yelp in 2016/2017, they have not yet received any Yelp ratings during the entire estimation period 2008–2015. The identification assumptions of ω_k

[24] See https://www.wsj.com/articles/doctors-check-online-ratings-from-patients-and-make-changes-1400541227 and https://www.wsj.com/articles/what-doctors-are-doing-about-bad-reviews-online-1498442580 for two Wall Street Journal articles on physicians' awareness of their ratings.

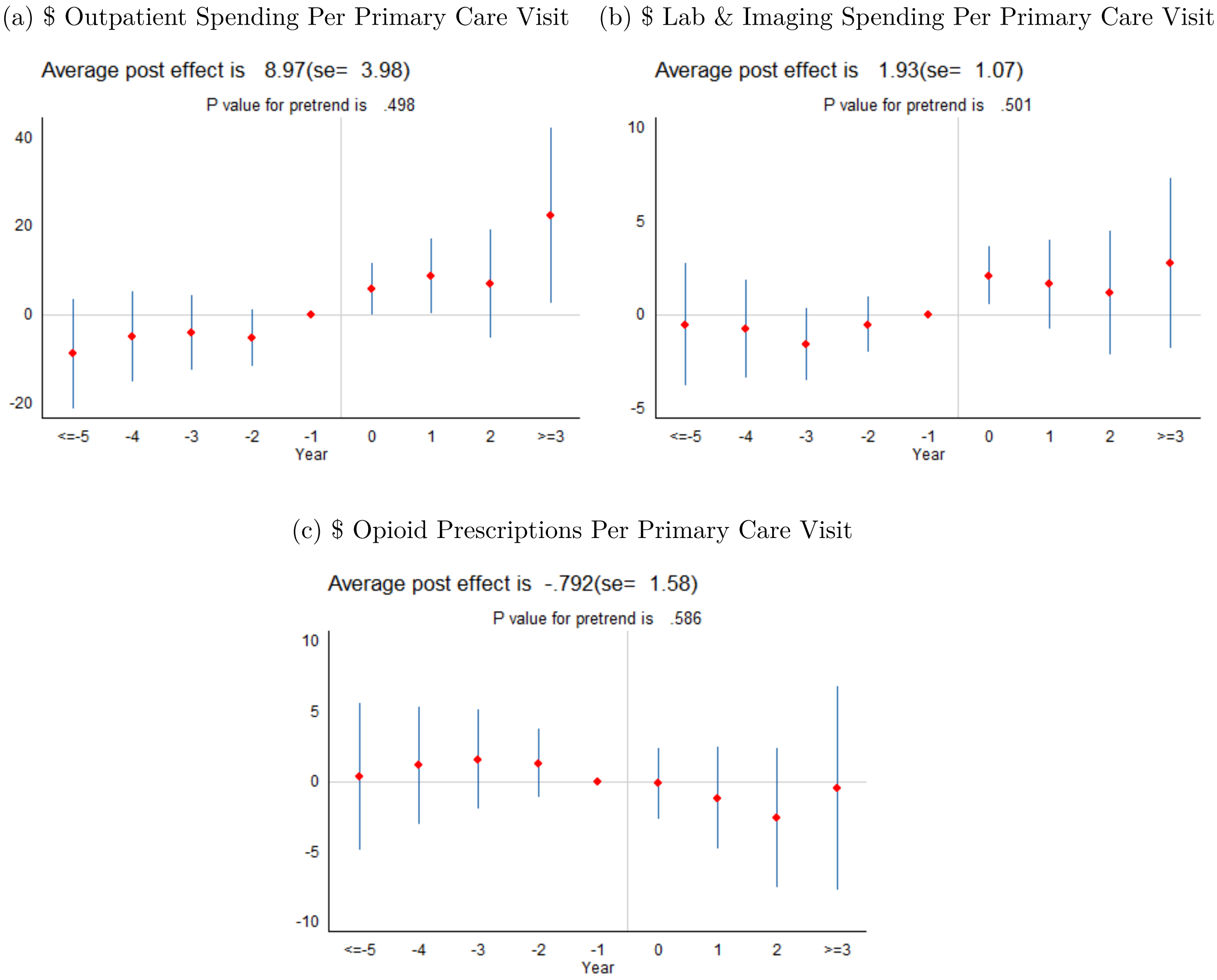
are that the timing of first reviews is exogenous to physician behaviors and patient selection, and patients do not desire more or less treatments because their physicians are rated. If both assumptions are true, the differences in the existing patients' health utilization and outcomes reflect physician efforts. ω_k should be about 0 for $k < 0$ as physicians and patients should not show differential behaviors and selection prior to the first rating date, and ω_k for $k \geq 0$ captures the causal effect of being rated on a physician's prescribing behavior. Caveats need to be applied to the interpretations of the findings due to the strong identification assumptions, and thus the results are only suggestive.

5.2 Results

Figures 4 and 5 graphically display the estimation results from equation 6. Figure 4 plots the $\hat{\omega}_k$ coefficients of patients' health services received. The dependent variables represent a patient i's total Part B outpatient services in Panel (a), total lab and imaging spending in Panel (b), and total opioid spending in Panel (c), all per primary care visit. In these figures, each dot corresponds to ω_k with k on the x-axis. ω_{-1} is normalized to 0, and $k = 0$ is the first year a treatment physician receives her rating. Prior to year 0, the treatment and control patients do not differ statistically in their amount of outpatient spending, lab and imaging spending, or opioid spending per primary care visit. After the treatment physicians receive their first rating, the treatment patients receive more in outpatient and lab and imaging spending per primary care visit than the control patients, but not differently in opioid spending. In a separate regression that only includes $\omega_{\leq -5}$, ω_{-4}, ω_{-3}, ω_{-2}, and $\omega_{\geq 0}$ instead of all flexible ω_k, The title displays the regression coefficient of $\hat{\omega}_{\geq 0}$ and the sub-title exhibits the joint test p value of $\hat{\omega}_{\leq -5}$,.., $\hat{\omega}_{-2}$. The detailed regression results are shown in Table (D.3). On average, after being rated for the first time, the treatment patients start to receive a statistically significantly \$9 more (or 1.0% compared to the mean) in outpatient spending and \$2 more (or 1.1% more compared to the mean) in lab and imaging spending per visit. Figure 5 displays the $\hat{\omega}_k$ coefficients of the health outcomes of patients using the total numbers of ER visits, CMS risk scores and Charlson comorbidity index as the dependent variables respectively in Panel (a), (b), and (c). There are generally no statistically significant trends different from zero in patient health before or after the treatment physicians receive their first ratings on Yelp.

To conclude this section, there is evidence that pre-existing patients receive slightly more total outpatient services and lab and imaging tests per primary care visit after their primary care physician receives their first Yelp rating. However, I find no evidence of changes in opioid prescriptions and patient's overall health outcomes. It seems that although physicians may

Figure 4: Estimation Results of Equation (6)—Event Study of How Patients' Health Services Received Change by Years Since Being First Reviewed

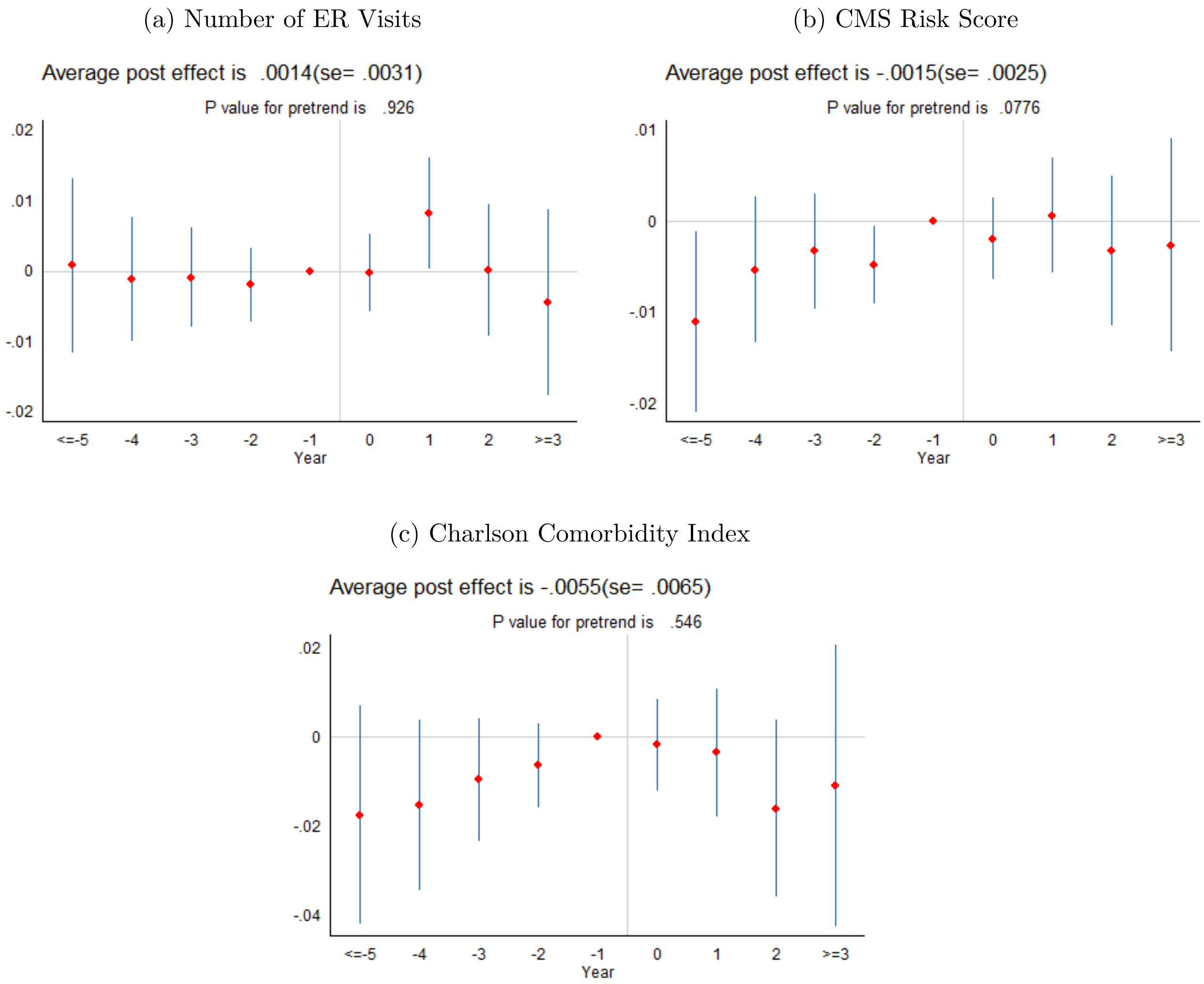

(a) $ Outpatient Spending Per Primary Care Visit

(b) $ Lab & Imaging Spending Per Primary Care Visit

(c) $ Opioid Prescriptions Per Primary Care Visit

<u>*Notes:*</u> The figures above show the estimation results of equation (6). The sample includes, among all cohorts, the pre-existing patients of the treatment physicians and control physicians from Medicare claim in the period 2008–2015. An observation is at the cohort-patient-year level. The dependent variable in Panel (a) is patient i's total Part B non-institutional outpatient spending in year t per primary care visit. The dependent variable in Panel (b) is patient i's total Part B non-institutional spending related to lab and imaging in year t per primary care visit. The dependent variable for Panel (c) is, for patient i who enrolls in both Part B and D, the total Part D spending on opioids in year t per primary care visit. k, the number of years since first ratings, is plotted on the x-axis. Each dot in the figure corresponds to ω_k on the y-axis with the 95% confidence intervals plotted on blue lines. ω_{-1} is normalized to 0. In a regression including only $\omega_{-5},...,\omega_{-2}$ and $\omega_{\geq 0}$ as the right-hand side instead of all flexible ωs, the estimated coefficient of $\omega_{\geq 0}$ and the joint p value of pre-trend coefficients are included in the titles and subtitles. Standard errors are clustered at the physicians' practice HSA levels. Specialty levels are not further clustered since the sample already only focuses on primary care physicians.

Figure 5: Estimation Results of Equation 6—Event Study of How Patients' Health Outcomes Change by Years Since Being First Reviewed



(a) Number of ER Visits

(b) CMS Risk Score

(c) Charlson Comorbidity Index

Notes: The figures above show the estimation results of equation (6). The sample includes, among all cohorts, the pre-existing patients of the treatment physicians and control physicians from Medicare claim in the period 2008–2015. An observation is at the cohort-patient-year level. The dependent variable in Panel (a) is patient i's total number of ER visits in year t. The dependent variable in Panel (b) is patient i's risk score calculated using the CMS-HCC 2015 model. The dependent variable for Panel (c) is patient i's risk score calculated using the Charlson model. k, the number of years since first ratings, is plotted on the x-axis. Each dot in the figure corresponds to ω_k on the y-axis with the 95% confidence intervals plotted on blue lines. ω_{-1} is normalized to 0. In a regression including only ω_{-5},...,ω_{-2} and $\omega_{\geq 0}$ as the right-hand side instead of all flexible ωs, the estimated coefficient of $\omega_{\geq 0}$ and the joint p value of pre-trend coefficients are included in the titles and subtitles. Standard errors are clustered at the physicians' practice HSA levels. Specialty levels are not further clustered since the sample already only focuses on primary care physicians.

very mildly order more lab services that are possibly wasteful or at least not in ways that significantly improve health, they do not seem to order more harmful opioids and harm patient health.

In the appendix, I have included a series of balance checks of the treatment and control physicians before the treatment physicians receive their first ratings on Yelp in Appendix D.1. In Appendix D.2, I test whether there is a statistically significant linear pre-trend of the outcome variables. For the most part, I cannot find any. In Appendix D.3, I also outline a robustness check restricting to physicians whose first ratings are low, as those physicians may receive their ratings more exogenously. I discover qualitatively similar findings in the exercise. In Appendix D.4, I use an approach proposed by Freyaldenhoven et al. (2018) to control for potential confounding pre-trend in patient characteristics and find similar results. In addition, I examine whether other physician behavior changes occur after being rated in Appendix D.5. I find that physicians' future ratings are higher than their first-year ratings, suggesting that physicians may have increased their efforts to get higher ratings after being rated. I also explore whether different levels of ratings have different impacts on physicians' organization sizes. I find that after receiving a high rating, a physician is more likely to bill from smaller organizations than if receiving a low rating. However, it remains unclear whether this relates to a physician's decision to work more in smaller organizations after receiving high ratings or a patient's decision to visit the high-rated physician more in her smaller practice group.

6 Who Benefits and Loses from Yelp Ratings?

6.1 Model

After analyzing the empirical evidence, I construct a stylized normative model that helps to understand the welfare implications of Yelp ratings on each player in the healthcare market.

I specify a utilitarian social planner's social welfare function. For patient surplus, I define their total consumer surplus in a world without Yelp as

$$\begin{aligned} cs &= cs^s + cs^c \\ &= \sum_j q_j^s D_j + \sum_j q_j^c D_j, \end{aligned} \tag{7}$$

where j denotes a physician, cs^s refers to patients' service quality surplus received, cs^c represents patients' clinical-quality surplus received, D_j is the demand of physician j, and

q_j^s and q_j^c denote the service and clinical quality of physician j.

In a world in which every physician receives Yelp ratings, the consumer surplus becomes

$$\begin{aligned} cs^* &= cs^{s*} + cs^{c*} \\ &= \sum_j q_j^{s*} D_j^* + \sum_j q_j^{c*} D_j^*, \end{aligned} \tag{8}$$

where q_j^{s*} and q_j^{c*} are the new quality of physician j now that she is rated on Yelp. The new demand D_j^* is defined as

$$D_j^* = (1 + \alpha_v + \beta_v r_j) D_j. \tag{9}$$

where r_j is physician j's received average cumulative rating ranging from 1 to 5 stars, α_v represents the average effect of being rated on a physician's patient volume, and β_v captures the differential effect on volume depending on the levels of ratings.

For physician surplus, the total physician revenue (denoted as supplier revenue, sr) without online ratings is

$$sr = \sum_j y_j, \tag{10}$$

where y_j denotes her revenue in a world where no physician is rated.

With every physician rated on Yelp, I assume that the total revenue would become

$$sr^* = \sum_j y_j^*, \tag{11}$$

with

$$y_j^* = y_j (1 + \alpha_R + \beta_R r_j), \tag{12}$$

where α_R represents the average effect of being rated on physician j's revenue and β_R captures the differential effect on revenue depending on the levels of ratings.

6.2 Calibration

The sample of the welfare calculation under the model above includes all physicians who billed Medicare Part B between 2012 and 2015 from the Medicare payment data but only receive their first Yelp ratings after 2015. The construction allows me to directly observe a physician's revenue without Yelp ratings.

The model is calibrated as follows. For equation (8), I calibrate β_v as 0.012 from the result of equation (3) in Table 3, as a 1-star increase in average physician ratings causally increases a physician's patient volume by 1.2%. α_v is calibrated such that the total patient visits with Yelp $\sum_j D_j^*$in the market is 101% of total visits without Yelp $\sum_j D_j$. The 101% comes from Panel (a) in Figure 4, allowing physicians to over-prescribe total outpatient services by 1% after being rated. I decide not to use $\hat{\alpha}$ from the estimation results of patient choice equation (3) and in Table (3) since that coefficient only captures the before versus after effects of being rated on a physician's own patient flow without capturing the increase in referral amounts to other physicians. D_j is measured by the observed average number of unique patients between 2012 and 2015.

For physicians' service quality q_j^s of equation (7) and (8), I use two versions of measurements. In the first one, I use a physician's first-year rating for physician service quality with and without Yelp ratings, that is, $q_j^{s*} = q_j^s = r_j^f$. This specification assumes that a physician's service quality is constant. In the second version, I use a physician's first-year rating as the quality of physician service without Yelp q_j^s. However, I define the service quality with Yelp ratings q_j^{s*} as $q_j^{s*} = q_j^s + 0.13$, where 0.13 is the average physicians' rating improvement compared to their first ratings. This is estimated from equation (25) from Appendix D.5.1. For a physician's clinical quality q^c, I assume $q_j^c = q_j^{c*}$ as I do not find health outcome changes for patients after their physicians being rated in Figure 5. I measure q_j^c as a physician's average probability of being board-certified, predicted by her Yelp cumulative average ratings in June 2017 using the matched sample between Yelp and Healthgrades data.

For equation (11), β_R is calibrated to 0.019 from the regression result of equation (3) in Table 3, as a 1-star increase in average physician ratings causally increases a physician's revenue by 1.9%. α_R is calibrated such that the total revenue with Yelp $sr^*$for all physicians amounts to 101% of total revenue without Yelp sr, using the same intuition that a physician will over-prescribe after being rated. The total revenue without Yelp sr is measured by the observed average revenue between 2012 and 2015 for the physicians the sample.

6.3 Results

Table 4 summarizes the welfare implications and their key channels. The first portion highlights the changes in patient surplus in a world without Yelp compared to to a world with every physician being rated on Yelp. If one assumes $q_j^s = q_j^{s*} = r_j^f$, that is, a physician's service quality is her first-year Yelp rating, patients will gain +2.2% in service-quality surplus with Yelp. This is because patients are now more informed and can seek out higher rated

Table 4: Welfare Implications

Players	Annual Change in Surplus	Mechanism
Patients	+2.2% to +6.6% in service quality	Ratings bring patients to higher rated physicians Physician improve ratings
	+1.1% in technical quality	Ratings bring patients to higher rated physicians Higher rated physicians are clinically better Physician do not impact patients' health after rated
Physicians	5-star physicians gain +4.1% in revenue	Ratings bring patients to higher rated physicians
	1-star physicians lose -3.4% in revenue	
	Unknown costs of physicians	Costly amenity investments Responsibility for control costs Moral standards
Taxpayers	1% more extra payment	Physicians over-prescribe after being rated

Notes: The table above shows the welfare implications of patient surplus from equations (8) and (11) compared to equations (7) and (10) for patients, physicians, and taxpayers. Column 2 displays the changes in total surplus as a percentage in a world in which every physician has Yelp ratings compared to a world without Yelp (see Section 6.1 and 6.2 for the model and related calibrations). Column 3 outlines the key channels behind the changes in surplus.

physicians who have better service quality. If one assumes that a physician's service quality improves after being rated on Yelp, patients' service-quality surplus increases by +6.0% with Yelp. For patients' clinical-quality surplus, the patients will gain +1.1% alone from finding more highly rated physicians because such physicians on Yelp also perform better clinically, as measured by higher chances of being board certified for example.

The second portion of Table 4 points out the changes in annual revenue for physicians if everyone is rated on Yelp. Physicians with 5-star ratings gain, on average, +4.1% more in revenue annually, compared to the -3.4% loss for 1-star physicians. The missing piece in physician welfare is the cost for physicians, who must exert efforts to please patients and improve their ratings. The costs may come from their investments in amenities, training office staff, cost-control responsibility for insurance firms, and moral standards after over-prescribing health services. Future studies should investigate these issues. For a risk-averse physician, the *Ex-ante* risk of receiving a noisy Yelp rating may also hurt her welfare.

Social planners should also consider the externality of online ratings. As physicians in general start to prescribe more services from Figure 4, taxpayers need to pay the extra burden of about 1% in outpatient services.

7 Conclusion

Some worry that user-generated physician ratings online could prove useless or detrimental. As consumers generally do not possess the knowledge for interpreting physicians' clinical decisions, their reviews might not reflect clinical quality and only reveal consumers' opinions on physicians' attitudes and service quality. As a result, the clinical consequence of patients visiting physicians with high ratings is ambiguous and depends on whether ratings are positively correlated with clinical abilities. It is also unclear whether these ratings actually affect patients' physician choice decisions in reality as consumers may not weigh them enough in their decision-making. Moreover, since physicians may realize that patients can more easily observe their service quality but not the clinical one, they may be distorted to over-please patients by being too nice, even if that means the treatment decisions would not be ideal clinically.

I use the empirical evidence about Yelp and Medicare to investigate these concerns and find that Yelp ratings actually help patients. Through text analysis and surveys, I discover that Yelp reviews contain much information about a physician's interpersonal skills and amenities. Using various measures of physicians' clinical quality from their educational and professional backgrounds and claims data, I find that Yelp ratings are also robustly positively correlated with these measures of clinical quality. Although the reviewers may not understand their physicians' clinical decisions, their ratings from judging service quality still provide useful information on clinical quality on average. The evidence also points out that patients' choices are significantly affected by Yelp ratings. Using the IV estimates from the plausibly exogenous variations of reviewers' "harshness," I show that physicians after receiving 5-star ratings would receive 4-8% more patient volume than if receiving 1-star. This result is striking since a physician who is rated on Yelp only has about 5 reviews on average. Last, utilizing a difference-in-differences design exploiting the different timing of first ratings among physicians, I generally do not find evidence that a physician would increase prescriptions of harmful substances for their pre-existing patients after being rated on Yelp.

Looking forward, this paper is only a starting point on the topic of online physician ratings and points out many future research areas. First, the cost of physicians in response to being rated is unclear as physicians may need to invest in staff training, amenity improvements, and etc. They also face uncertainty from potentially very noisy ratings. Second, a very useful extension would be an analysis on a different patient and physician sample, such as younger patient population and physicians in large group practices. Third, a current challenge of the Yelp rating system and online physician rating systems in general relates to the large variability of ratings due to the small number of reviews per physician. Promoting review generation in the future can improve the reliability of physician ratings. However, letting physicians collect reviews on their own will most certainly bias the reviews due to cherry-picking of good reviews. Efforts from third-parties to representatively collect patient reviews may prove more promising and requires further study. Last but not least, researchers should also think about how to better present physician information so that consumers can more easily understand the quality of a physician, for example, by showing patients objective physician performance measures along with user-generated ratings and educating them on how to interpret these measures.

References

Anderson, M. & Magruder, J. (2012). Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database*. *The Economic Journal*, 122(563), 957–989.

Bardach, N. S., Asteria-Peñaloza, R., Boscardin, W. J., & Dudley, R. A. (2013). The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ Quality & Safety*, 22(3), 194–202.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3, 993–1022.

Dafny, L. & Dranove, D. (2008). Do report cards tell consumers anything they don't already know? The case of Medicare HMOs. *The RAND Journal of Economics*, 39(3), 790–821.

Dewan, S. & Hsu, V. (2004). ADVERSE SELECTION IN ELECTRONIC MARKETS: EVIDENCE FROM ONLINE STAMP AUCTIONS. *Journal of Industrial Economics*, 52(4), 497–516.

Dranove, D. & Jin, G. Z. (2010). Quality Disclosure and Certification: Theory and Practice. *Journal of Economic Literature*, 48(4), 935–963.

Dranove, D., Kessler, D., McClellan, M., & Satterthwaite, M. (2003). Is More Information Better? The Effects of "Report Cards" on Health Care Providers. *Journal of Political Economy*, 111(3), 555–588.

Faber, M., Bosch, M., Wollersheim, H., Leatherman, S., & Grol, R. (2009). Public reporting in health care: How do consumers use quality-of-care information? A systematic review. *Medical Care*, 47(1), 1–8.

Feng Lu, S. (2012). Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes. *Journal of Economics & Management Strategy*, 21(3), 673–705.

Fenton, J. J., Jerant, A. F., Bertakis, K. D., & Franks, P. (2012). The cost of satisfaction: A national study of patient satisfaction, health care utilization, expenditures, and mortality. *Archives of Internal Medicine*, 172(5), 405–411.

Freyaldenhoven, S., Hansen, C., & Shapiro, J. (2018). *Pre-event Trends in the Panel Event-study Design.* Technical report, National Bureau of Economic Research, Cambridge, MA.

Fung, C. H., Lim, Y.-W., Mattke, S., Damberg, C., & Shekelle, P. G. (2008). Systematic review: the evidence that publishing patient care performance data improves quality of care. *Annals of internal medicine*, 148(2), 111–23.

Gentzkow, M., Kelly, B., & Taddy, M. (2017). *Text as Data.* Technical report, National Bureau of Economic Research, Cambridge, MA.

Hibbard, J. H., Greene, J., & Daniel, D. (2010). What is quality anyway? Performance reports that clearly communicate to consumers the meaning of quality of care. *Medical Care Research and Review*, 67(3), 275–293.

Hoffmann, T. C. & Del Mar, C. (2015). Patients' Expectations of the Benefits and Harms of Treatments, Screening, and Tests. *JAMA Internal Medicine*, 175(2), 274.

Holmstrom, B. & Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, 7, 24–52.

Howard, P., Fellow, S., & Feyman, Y. (2017). Yelp for Health Using the Wisdom of Crowds To Find High-Quality Hospitals.

Jacob, B. A. & Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 118(3), 843–877.

Jin, G. Z. & Leslie, P. (2003). The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards. *The Quarterly Journal of Economics*, 118(2), 409–451.

Kolstad, J. T. (2013). Information and Quality When Motivation is Intrinsic: Evidence from Surgeon Report Cards. *American Economic Review*, 103(7), 2875–2910.

Kolstad, J. T. & Chernew, M. E. (2009). Quality and consumer decision making in the market for health insurance and health care services. *Medical care research and review : MCRR*, 66(1 Suppl), 28S–52S.

Lembke, A. (2012). Why Doctors Prescribe Opioids to Known Opioid Abusers. *New England Journal of Medicine*, 367(17), 1580–1581.

Lu, S. F. & Rui, H. (2017). Can We Trust Online Physician Ratings? Evidence from Cardiac Surgeons in Florida. *Management Science*, 2015-March(November), mnsc.2017.2741.

Luca, M. (2015). User-Generated Content and Social Media. *SSRN Electronic Journal*, (pp. 1–49).

Luca, M. (2016). Reviews, Reputation, and Revenue: The case of Yelp.com. *Working Paper*.

Luca, M. & Smith, J. (2013). Salience in Quality Disclosure: Evidence from the U.S. News College Rankings. *Journal of Economics & Management Strategy*, 22(1), 58–77.

Mullen, K. J., Frank, R. G., & Rosenthal, M. B. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *The RAND Journal of Economics*, 41(1), 64–91.

Pham, H. H., Landon, B. E., Reschovsky, J. D., Wu, B., & Schrag, D. (2009). Rapidity and Modality of Imaging for Acute Low Back Pain in Elderly Patients. *Archives of Internal Medicine*, 169(10), 972.

Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., Asch, D. A., Ungar, L. H., & Merchant, R. M. (2016). Yelp Reviews Of Hospital Care Can Supplement And Inform Traditional Surveys Of The Patient Experience Of Care. *Health Affairs*, 35(4), 697–705.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, 102(7), 3184–3213.

Scanlon, D. P., Chernew, M., McLaughlin, C., & Solon, G. (2002). The impact of health plan report cards on managed care enrollment. *Journal of Health Economics*, 21(1), 19–41.

Schneider, E. C. & Epstein, A. M. (1998). Use of public performance reports: A survey of patients undergoing cardiac surgery. *Journal of the American Medical Association*, 279(20), 1638–1642.

Sievert, C. & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, (pp. 63–70).

Sinaiko, A. D., Eastman, D., & Rosenthal, M. B. (2012). How report cards on physicians, physician groups, and hospitals can have greater impact on consumer choices. *Health Affairs*, 31(3), 602–611.

Torke, A. M., Sachs, G. A., Helft, P. R., Montz, K., Hui, S. L., Slaven, J. E., & Callahan, C. M. (2014). Scope and Outcomes of Surrogate Decision Making Among Hospitalized Older Adults. *JAMA Internal Medicine*, 174(3), 370.

Werner, R. M., Konetzka, R. T., Stuart, E. A., Norton, E. C., Polsky, D., & Park, J. (2009). Impact of Public Reporting on Quality of Postacute Care. *Health Services Research*, 44(4), 1169–1187.

Werner, R. M., Norton, E. C., Konetzka, R. T., & Polsky, D. (2012). Do consumers respond to publicly reported quality information? Evidence from nursing homes. *Journal of Health Economics*, 31(1), 50–61.

Appendix

A Detailed Yelp Data Collection and Matching Algorithms

A.1 Algorithm of Yelp Rating Collection

The following algorithm is used to collect individual review for physicians on Yelp in the U.S.

1. Go through each city in each state within the U.S.

2. Go through "Doctors" under "Health and Medical."

3. Collect non-empty business listing for the top 1000 pages.

4. If a city has sub-neighborhoods, go through top 1000 pages under each sub-neighborhood.

5. Within each listing, go through each individual review. Collect the date of review, review stars, review text, and the rating distribution of all the reviewer's reviews in all businesses.

This process results in the collection of 95,030 business listings.

A.2 Algorithm of Matching Yelp with NPI Directory

Next, I merge the data with the 2016 National Provider Identifier (NPI) directory.[25] NPI is "a Health Insurance Portability and Accountability Act (HIPAA) Administrative Simplification Standard. The NPI is a unique identification number for covered health care providers. Covered health care providers and all health plans and health care clearinghouses must use the NPIs in the administrative and financial transactions adopted under HIPAA."[26] To match the Yelp rating data with NPI directory, I proceed to use an algorithm with the following steps:

1. Remove common words from Yelp ratings (e.g. Dr., Jr., etc.). Define the first word of a business listing as the first name and the second word as the last name.

[25] Downloaded from http://download.cms.gov/nppes/NPI_Files.html.

[26] Cited from https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand/.

2. Merge the rating data with the NPI directory using exact matches for the last name, the first name, and the relevant Health Service AREA (HSA).

3. Merge the rating data with the NPI directory using reversed matching of the last name, the first name, and the relevant HSA location. That is, use the second word in step 1 as the first name and the first word as the last name. If there is a conflict in matching from step 2, use the matching result from step 2.

Through the algorithm, 39% of the Yelp listings can be found a match. Out of the 39%, 90% of them can be found a unique match from the NPI directory and the rest 10% that have duplicated matches are discarded. In total, this process ends up matching 36,787 individual physician listings.

B Details of Text Analysis

B.1 Details of the LDA Text Analysis

The algorithm uses a latent Dirichlet allocation model and involves the following steps.

1. Importing documents
   Raw Yelp reviews with suffixes "MD," "DO," "OD" in the listing names are read into a Python program.

2. Tokenization
   Break a review into words split by any non-word character, e.g., spaces and punctuations. Transform all resulting words into lowercase letters.

3. Removing stop-words
   Remove stop-words in English from the documents such as "for," "or," "the." The list comes from the Python package "stop_words.get_stop_words," with additional stop words "dr," "s," "m," "z," "d," "ve," and "t." "Dr" is commonly mentioned in review texts, and the others are commonly used after apostrophes and other characters.

4. Stemming words
   Reduce similar words such as "stemming," "stemmer," "stemmed" into "stem." This comes from the Python package "nltk.stem.porter.PorterStemmer."

5. Applying LDA model
   Feed the pre-cleaned data into the LDA model using the Python package "gensim.models.ldamodel," pre-specifying 20 topics in total and running through the model 100

times. When finished, the model classifies reviews into topics, and estimate the probability distribution of words within each topic.

The technical details of the LDA model are documented in Blei et al. (2003). The basic idea is that a review i is considered an unordered set of words. It has a probability distribution of generating topics governed byθ_i. A topic k is a probability distribution that generates each key word in the entire vocabulary (all possible words in Yelp reviews), governed by ψ_k. A particular word w in review i is then generated from the following steps. First, drawing a topic z from review i's topic distribution θ_i; then, given topic z, drawing a keyword w from the keyword distribution ψ_k. After the algorithm reads in all the realized keyword distributions from all reviews, it employs a Bayesian algorithm to infer the posterior distribution of topic distribution per review θ_i and keyword distribution per topic ψ_k.

6. Rank topics and keywords

To rank topics, I compute the probability of any review belonging to each topic and rank all topics in reverse order by that probability. To rank keywords within a topic, I adopt the approach in Sievert & Shirley (2014) that penalizes a keyword if its overall probability from all reviews is high and rank each keyword w by its adjusted within-topic(k) probability, $r(w, k)$. It is calculated using the following formula, in which ϕ_{kw} is the probability of keyword w within topic k, P_w is the overall probability of keyword w across all reviews, and λ, the relative weight between the probability of a keyword within topic ϕ_{kw} and the penalization term if the keyword has high overall probability $\frac{\phi_{kw}}{p_w}$, is set to $\frac{2}{3}$.

$$r(w, k) = \exp[\lambda \log(\phi_{kw}) + (1 - \lambda) \log(\frac{\phi_{kw}}{p_w})]$$

B.2 Details of Amazon Mechanical Turk Survey

I consider all Yelp reviews with suffixes "MD," "DO," "OD" in the listing names and draw a random sample of 1,500 reviews from them. I then design the questionnaire in Figure B.1, including 4 categories—"service quality related," "clinical quality related," "both of the above," and "other." I send the survey to "Amazon Mechanical Turk" and request that each review must have two respondents answering it to ensure data quality. A survey respondent can examine multiple reviews. In total, 50 survey respondents from "Amazon Mechanical Turk" answer the $(1500 * 2 = 3000$ reviews).

Among the 1,500 reviews, 781 prompt the same answer from two reviewers. Among them, 359 are classified as "service quality related," 70 as "clinical quality related," 277 as "both of the above," and 75 as "other." This result translates to that 78% of reviews are "service quality related," and 36% are "clinical quality related," with "both of the above" counting for both categories.

The analysis above perhaps has higher quality since two reviewers agree on the interpretation of a review. However, if instead, I consider the two answers for each of the 1,500 reviews as independent answers for 3,000 reviews, 77% of reviews would be considered as "service quality related" and 48% would fall under the category "clinical quality related."

C Robustness and Heterogeneity of Patient Choices

C.1 Robustness Checks of Instruments

C.1.1 Alternative Construction of Instruments

I construct an alternative version of the instruments that may alleviate some concerns regarding the potential endogeneity. The intuition is to use only non-medical businesses to construct a reviewer's "harshness" and remove the impacts of baseline rating of that business. The sample only includes physicians from the default list of popular cities on Yelp.[27] For every reviewer k of a physician j, I collect the average leave-reviewer-k-out ratings of the first three businesses that reviewer k has reviewed, denoted as a_{k1}, a_{k2}, and a_{k3}, and denote k's ratings for those three businesses, denoted as g_{k1}, g_{k2}, g_{k3}. Then among all reviewers k and their three reviews, excluding medical businesses, I estimate the following equation:

$$g_{ki} = \beta_0 + \beta_1 a_{ki} + \eta_{ki}.$$

I then construct the alternative "harshness" for reviewer k as

$$h_k = \frac{1}{3}\sum_i \hat{\eta}_{ki}.$$

The new instrument z^{alt} at the physician year level is constructed in a similar fashion as the main instrument in equation (4):

$$z^{alt}_{jt} = \frac{1}{n_{jt}} \sum_{k \in K(jt)} h_k. \tag{13}$$

This construction is perhaps more exogenous, since it removes the baseline attraction of "harsh" reviewers into good or bad businesses, and removes a reviewer's likelihood to rate an overall good or bad business. However, such a method requires extensive data collection

[27] From https://www.yelp.com/locations

at the reviewer level. I therefore only collect the data for physicians from the popular city list and for only three business listings per reviewer. The results are shown in Table C.5. Reassuringly, the results very closely reflect those of the main estimation results from Table 3.

C.1.2 Correlating Instruments with Observables

I test whether the instruments are correlated with observable quality characteristics. If they are not correlated, one may have more confidence that they are not correlated with unobservable physician quality. For time-varying characteristics, I merge the main rated physician sample with Medicare claims data in the period 2008–2015 and estimate the following equation (14). The dependent variables are a primary care physicians' annual average adherence to HEDIS guidelines among a physician's patients. The sample includes only physicians who have ever received ratings (I ignore the physicians who are never rated since both λ and β in equation (3) are only identified off physicians who have changed from being non-rated to being rated).

$$x_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \alpha D_{jt} + \gamma z_{jt} D_{jt} + \epsilon_{j,t}. \tag{14}$$

The estimation results from Table C.1 indicate that the correlations are weak and insignificant.

Many physician quality characteristics are also time unvarying. I conduct a cross-sectional regression to test for the existence of a correlation between the instruments and these characteristics. The sample merges the estimation sample in equation (3) with either Healthgrades or Physician Compare data. Among the physicians ever rated and only including t during which physician is being first rated, I estimate:

$$x_{jt} = \theta_{s,t} + \theta_{h,t} + \gamma z_j^f + \epsilon_{j,t}. \tag{15}$$

Here x_j includes a physician's board certification, education ranking, and self-reported number of accreditations from Physician Compare. z_j^f refers to physicians' "harshness" index during their first year of receiving ratings. The estimation results from Table C.2 suggest there are no statistically significant relationships even between the first-year instruments and observable time-unvarying physician quality.

C.1.3 Correlating Being Rated with Observables

I test whether a physician j being rated in year t D_{jt} is correlated with observable quality characteristics controlling the flexible fixed effects. If they are not correlated, one may have more confidence that they are not correlated with unobservable physician quality as well. For time-varying characteristics, I merge the main rated physician sample with Medicare claims data in 2008-2015 and estimate the following equation (16) using only the physicians ever rated (I ignore the physicians who are never rated since both λ and β in equation (3) are only identified off physicians who changed from being non-rated to being rated):

$$D_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \gamma x_{jt} + \epsilon_{j,t}, \tag{16}$$

where x_{jt} are physicians' annual average adherence to HEDIS guidelines among her patients. The estimation results from Table C.3 indicate that the correlations are weak and insignificant.

Many physician quality characteristics are also time-unvarying. I estimate a cross-sectional regression to test whether the instruments are correlated with these characteristics. The sample merge the estimation sample in equation (3) with either Healthgrades data or Physician Compare data. Among the physicians ever rated, I estimate:

$$D_{jt} = \theta_{s,t} + \theta_{h,t} + \gamma x_j + \epsilon_{j,t}. \tag{17}$$

Here x_j includes a physician's board certification, education ranking, and self-reported number of accreditations from Physician Compare. The estimation results from Table C.4 suggest that mostly (other than Usnews medical school rankings), there are no statistically significant relationships even between being rated and observable time-unvarying physician quality.

C.2 Heterogeneity of the Impact of Yelp ratings on Patients' Physician Choices

C.2.1 Are Patient Responses Stronger with More Reviews?

Bayesian learning suggests that, with more reviews, an average Yelp rating signals more information about a physician's quality. This suggests that patient response β should be larger among physicians with more reviews. I test such prediction using the specification below, among the same sample from equation (3):

$$y_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \alpha D_{jt} + \beta R_{jt} \cdot D_{jt} + \gamma R_{jt} \cdot n_{jt}^{0.5} \cdot D_{jt} + \kappa_{n_{jt}^{0.5}} + \epsilon_{j,t}. \tag{18}$$

$n_{jt}^{0.5}$ is the squared root of the number of reviews of physician j by year t. $\kappa_{n_{jt}^{0.5}}$ represents the fixed effect of the number of reviews squared $n_{jt}^{0.5}$ to control for the potential endogeneity of number of reviews. The interaction term between rating R_{jt} and the number of reviews captures whether ratings affect the left-hand side more strongly when there are higher number of reviews. The square root specification allows the effects to be diminishing. The estimated results using OLS are displayed in columns 1 and 2 of Table C.6. Patient volume responses are stronger for physicians with more reviews ($\hat{\gamma} > 0$), consistent with a Bayesian learning story of patient response. In addition, I provide instrumental variable estimations in columns 7 and 8, instrumenting $R_{jt} \cdot D_{jt}$ and $R_{jt} \cdot n_{jt}^{0.5} \cdot D_{jt}$ with $z_{jt} \cdot D_{jt}$ and $z_{jt} \cdot n_{jt}^{0.5} \cdot D_{jt}$. The results are similar to the OLS estimations.

C.2.2 Are Patient Responses Weaker among Physicians with Older Patients?

Online physician ratings may also be used less for older patients who tend to lack Internet access. I construct an age index a_j of a physician j's patient pool by computing the average share of patients aged 85+ among physician j's total number of Medicare patients from 2012 to 2015. I specify the following equation among the same sample from equation (3).

$$y_{jt} = \chi_j + \theta_{s,j} + \theta_{h,j} + \alpha D_{jt} + \beta R_{jt} \cdot D_{jt} + \gamma R_{jt} \cdot (a_j - \mu_a) \cdot D_{jt} + \epsilon_{j,t}. \quad (19)$$

μ_a is the mean of a_j in the sample. I expect γ_j to be negative because older patients may respond less strongly to online ratings. The empirical estimation results using OLS estimations are shown columns 3 and 4 in Table C.6. The estimated $\hat{\gamma}$ is indeed negative. If instrumenting $R_{jt} \cdot D_{jt}$ and $R_{jt} \cdot a_j \cdot D_{jt}$ with $z_{jt} \cdot D_{jt}$ and $z_{jt} \cdot a_j \cdot D_{jt}$, the results in columns 9 and 10 are also similar.

C.2.3 Are Patient Responses Stronger in Areas with More Educated Elderly?

Education also positively correlates with Internet and Yelp usage.[28] I measure elderly education levels using national percentiles of (% of elderly with a bachelor degree or up in each county) in 2016 from the American Community Survey, denoted as e_{jt} in each physician j's county in year t. I specify the following estimation:

$$y_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \alpha D_{jt} + \beta R_{jt} \cdot D_{jt} + \gamma R_{jt} \cdot (e_{jt} - \mu_e) \cdot D_{jt} + \epsilon_{j,t}. \quad (20)$$

μ_e is the mean of e_{jt} in the sample. The intuition predicts that the more educated elderly would respond more strongly to Yelp ratings, and therefore $\gamma > 0$. In columns 5 and 6,

[28] Cited from https://www.yelp.com/factsheet downloaed on 4/20/2018, 66% of Yelp users are college educated compared to 19% who are below college degrees.

Table C.6 reveals positive although insignificant $\hat{\gamma}$s using OLS estimations. The findings suggests that in the most educated area ($e_j = 100$) compared to the least ($e_j = 1$), Yelp ratings have a +1% impact on a physician's revenue per star increase. In columns 11 and 12, the instrumented versions show similar findings, $\hat{\gamma}$ are positive but insignificant.

The heterogeneity estimations above use different variations of review and patient characteristics from the main regression (3) and yet are all consistent with patients' physician choice theory due to online ratings. The interaction findings strengthen our confidence that patients' physician choice drives the association between Yelp ratings and physician revenue. Also interesting is the fact that the findings suggest that online physician ratings have some regressive distributional outcomes. Younger and more educated patients use Yelp ratings more often. Policymakers should consider how to promote the usage of online ratings among the older and less-educated members of the population.

C.3 Alternative Timing of Rating and Instrument Measurements

Readers who use Yelp reviews must have observed the ratings up to the previous year. One may argue that measuring ratings and instruments up to the end of the previous year instead of the current year may alleviate some of the measurement errors in the timing of when readers see and visit a physician. The downside of this measurement is that one may lose the effects of new ratings during the current year on patients' physician choices. I specify the estimation as follows similar to equation (3) and instrument $R_{jt-1}D_{jt-1}$ with $z_{jt-1}D_{jt-1}$:

$$y_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \alpha D_{jt-1} + \beta R_{jt-1}D_{jt-1} + \epsilon_{j,t}. \tag{21}$$

The results are displayed in Table C.7. Using the IV estimate, a star increase in the previous year average rating affects today's physician revenue by about +3% and patient volume by about +2%.

D Robustness and Heterogeneity of Physician Responses to Being Rated

D.1 Balance Check

I access the balance of the treatment and control group physicians and patients in equation (22). The estimation includes all cohorts of pre-existing patients defined in equation (6) and is at the cohort-patient-year level. The estimation sample is further restricted to observations before the treatment date for each cohort ($t < m$). The specification is as follows:

$$x_{it}^m = \alpha T_j^m + \theta_{s,t}^m + \theta_{h,t}^m + \eta_{it}^m, \tag{22}$$

where j is patient i's primary care physician in year t; $T_{j(i,t)}^m$ indicates whether patient i's primary care physician in year t, physician j, is in the treatment group of cohort m; θ_t are flexible cohort-specialty $(s(j), m)$ and cohort-HSA $(h(j), m)$ level time fixed effects; α captures the treatment and control group level differences in observed patient and physician characteristics x_{it}^m. Although the difference-in-differences estimation assumption from equation (6) does not necessarily require similar levels of patient and physician characteristics—only parallel trends are required—it will be more comfortable if the levels of x_{it}^m are also similar.

The estimation results are displayed in columns 1–9 of Table D.2. As seen in the table, before the treatment physicians receive their first ratings on Yelp, the two group of physicians do not differ statistically in their annual revenue, numbers of patients, organization sizes, and their pre-existing patients' total outpatient bill per primary care visit, total lab and imaging amount per primary care visit, and total opioid prescriptions per primary care visit. However, it does seem that the treatment patients are slightly healthier than the control patients.

D.2 Regression Results with Linear Pre-trends

To systematically test whether there is a pre-trend in the differences between the treatment and control group, using the sample from equation (6), I estimate the following specification to test whether there is a linear pre-trend among the treatment group:

$$y_{it}^m = \chi_{ij} + \theta_{s,t}^m + \theta_{h,t}^m + \alpha \cdot 1(t < m) \cdot (t - m) T_j^m + \beta \cdot 1(t \geq m) \cdot T_j^m + \epsilon_{it}^m, \tag{23}$$

where αcaptures the linear pre-trend between the treatment and control patients, and β captures the post effect of the treatment and control patients after their primary care physicians receive their first ratings on Yelp. The regression results are displayed in Table D.4 and they typically do not detect a statistically significant linear pre-trend $\hat{\alpha}$.

D.3 Low First Rating Only

One may argue that physicians with low first ratings are being first rated more exogenously as they are not trying to get rated. I perform a similar difference-in-differences estimation as in equation (6) restricting the sample to physicians whose first ratings are lower than or equal to 2 stars.

The balance check results using equation (22) are listed in columns 10–18 of Table D.2. Prior to the treatment physicians receiving their first ratings when focusing only on physicians with low first ratings, the results suggest that indeed the treatment and control physicians do look more similar than those from the previous exercise in D.1. Other than numbers of ER visits, the treatment and control patients and physicians do not differ significantly in other dimensions.

The results of the difference-in-differences estimations are displayed in Figures D.1 and D.2 and they reveal very similar patterns to those of the main estimations in Figures 4 and 5. Treatment patients seem to receive slightly more services in labs and imaging but not statistically different amount of opioids or change health outcomes after the treatment physicians receive their first Yelp ratings.

D.4 Estimator from Freyaldenhoven et al. (2018)

From the balance checks of columns 7-9 in Table D.2, patients in the treatment group seem slightly healthier than the control group before the treatment physicians are rated. In the main results from Figure 5, the treatment patients also seem to display a slight pre-trend in their CMS risk scores compared to the control group. Perhaps patients' different health characteristics would affect the timing of being first rated or they may demand different amount of health services over time, which would potentially confound the interpretation that it is being rated on Yelp that causes physicians to over-prescribe total outpatient services and lab and imaging tests.

In an attempt to control for this concern, I use an approach from Freyaldenhoven et al. (2018). If there is a covariate x_{it}that measures the confounding patient characteristics, one can add in this variable to control for the potential confounder. In addition, if x_{it} is unaffected by the event of being rated itself and nonetheless displays a pre-trend, one can use the lead of the event as an instrument for x_{it} to resolve the measurement error issue of x_{it}, that is, x_{it} does not measure the confounder perfectly. The IV strategy works since the lead of the event would affect x_{it} as x_{it} has a pre-trend (inclusion restriction). And the lead of the event would only correlate with x_{it} through correlating with the confounder but would not affect x_{it} otherwise by assumption (exclusion restriction). In my application, if one further assumes that patients' CMS risk scores are themselves unaffected by the event of being rated, and they measure patients' underlying health characteristics, which potentially affect the event of being rated, one can use these CMS risk scores as the additional control and instrument them with the lead of physician being rated. Using this method to remove the impacts of patients' health characteristics that are possibly confounding, I estimate the following regression at the cohort(m)-patient(i)-year(t) level using Medicare claims data from 2008 to 2015 to detect whether there are physician responses in ordering health services:

$$y_{it}^{m} = \chi_{ij} + \theta_{s,t}^{m} + \theta_{h,t}^{m} + \sum_{k} \omega_k 1(t - m = k) T_j^m + \tau x_{it} + \epsilon_{it}^{m}, \tag{24}$$

where y_{it}^{m} is patient i's health spending in year t who is also of cohort m and x_{it} is patient i's CMS health risk score in year t. In the estimation, I instrument x_{it} with the lead of the event $1(t+1 \geq m)T_{j(i,t)}^{m}$, that is, in the next year $t+1$, i's physician j has been rated in the treatment group.

The estimation results are displayed in Figure D.3. They are very similar as the main results from Figure 4 and even seem to have flatter pre-trends.

D.5 Other Changes after Physicians Are Rated

D.5.1 Rating Improvements

This section examines two other outcomes that physicians may change after being first rated. First, I test whether after receiving her first rating, a physician's future rating improves. This test may serve as suggestive evidence that physicians try harder to improve ratings. Among the physicians rated since 2008 in the main rated physician sample, starting from their first rated year and up until the end of rating data June 2017, I estimate the following physician(j)-year(t) panel regression:

$$r_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \alpha \cdot 1(t > d_j^f) + \epsilon_{jt}. \tag{25}$$

r_{jt} is physician j's rating in year t using various measurements. θ_t are flexible time fixed effects that are physician j's specialty $s(j)$ and HSA $h(j)$ specific. d_j^f denotes the year physician j receives her first rating. The coefficient of interest is α, which captures how physician ratings differ in the subsequent years from their first-year ratings. One should note that $\hat{\alpha}$ is only suggestive evidence of physicians' efforts to improve their ratings after observing that they are rated on Yelp, as the first raters can differ from future raters as well.

Table D.1 contains the estimation results. Column 1 uses the average new ratings received in year t (flow) as the dependent variable. Column 2 uses the cumulative average ratings received up to year t (stock) as the dependent variable. Columns 3 and 4 are similar to columns 1 and 2, but they use the reviewer "harshness" instruments as the dependent variables. From the table, subsequent ratings are on average +0.13 stars higher than the first-year ratings in flow and +0.05 higher in stock. On the other hand, the "harshness" instruments in subsequent years hardly change or are even a little lower (thus the reviewers are more “harsh”). The findings suggest that physician ratings do improve over time and not through physicians’ cherry-picking of “easier” reviewers. As mentioned earlier, the first reviewers may differ from consequent reviewers; thus, the results are not necessarily causal. However, a calculation not displayed in the table indicates that after the first year of ratings, the average new ratings during a year are still +0.08 stars higher in stars than the second ratings.

D.5.2 Implications on Organization Sizes

Second, different levels of ratings may affect where a physician sees a patient after being rated. High-rated physicians would work in small organizations to reap extra revenue, whereas low-rated physicians may switch to big organizations to bury their bad reputations. Alternatively, for physicians working in multiple groups, a patient may visit a highly rated physician more in her solo-practice rather than in her shared-practice group with other physicians because the former organization directly benefits from the high rating profile. If such effects exist, they will diminish the power of the rating system to discipline low-rated physicians, who will see more patients in large groups than the highly rated physicians. This section examines whether high- and low-rated physicians differentially bill from organizations of different sizes.

The main specification is an impulse response estimation. I consider all physicians who are rated between 2008 and 2017 in the main rated physician sample and estimate a panel regression at the physician j year t level during the period 2008–2015:

$$org_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + \sum_k (\mu_k D_{jt+k} + \nu_k D_{jt+k} R_{jt+k}). \tag{26}$$

org_{jt} is the organization size of physician j in year t as defined in Section 2.2. χ_j is a physician fixed effect. θ are flexible time trends for a physician’s specialty $s(j)$ and practice HSA $h(j)$. D_{jt+k} is an indicator whether physician j is rated by year $t+k$. μ_k captures the main effect of how having a rating in year $t+k$ affects a physician’s organization size in year t. ν_k is the coefficient of interest and captures how differentially high and low ratings in year $t+k$ affect the organization size. I predict that past ratings ($k<0$) will affect a physician’s current organization size negatively but not the future ratings ($k>0$).

The estimation results are presented in Table D.5. As one can see, only the past and current ratings are associated negatively and significantly with a physician’s organization size today. A 1-star increase in the cumulative average rating in year $t-1$, a stock measure, is associated with -0.2% in $Prob(org_{jt} \geq 10)$ and -0.7% in $log(org_{jt})$. A 1-star increase in average new ratings in year $t-1$, a flow measure, is associated with -0.3% in $Prob(org_{jt} \geq 10)$ and -0.9% in $log(org_{jt})$.

To visualize the organization size response, I further consider a triple-differences estimation. Among all physicians from Medicare Part B between 2008 and 2015 whether rated or not, I compare physicians who are ever rated versus never rated, as well as high- versus low-rated physicians, through the following physician j year t level regression:

$$org_{jt} = \chi_j + \theta_{s,t} + \theta_{h,t} + D_j \sum_k \psi_k 1(t-k=d_j^f) + D_j H_j \sum_k \omega_k 1(t-k=d_j^f) + \epsilon_{jt}. \tag{27}$$

In the specification, χ_j represents a physician fixed effect. θ are flexible time trends for a physician's specialty $s(j)$ and practice HSA $h(j)$, which capture the baseline time trends of the never rated physicians. ψ_k reflects the differential trends of physicians that have ever received a rating, in kth year relative to their first rated year d_j^f. H_j is an indicator of physicians being "high" rated, measured by their cumulative average ratings in June 2017 greater than or equal to 4 stars. ω_k captures the triple difference coefficient of interest, how high-rated physicians differ in organization sizes from physicians with other ratings in kth year since their first rated year. The advantage of this specification is that the triple-differences estimation visualizes how organization sizes change with respect to the first year a physician receives a rating. The disadvantages are that there are two coarse measures. First, the H_j indicator makes a binary segment of the ratings and loses the power of the continuous measure. Second, the measure of H_j is at the final year 2017, which fails to account for the intertemporal changes of the ratings before 2017. The identification assumption is that the physicians who are never rated provide a baseline for how the ever rated physicians will change their organization sizes in different calendar years. The physicians with lower ratings provide a counterfactual trend regarding how the high-rated physicians would evolve over time should they receive low or medium ratings.

The estimation results are plotted in Figure D.4. Before a physician's first rating ($k < 0$), there is no discernible difference between high-rated physicians and other rated physicians in their organization sizes. However, after the first ratings, high-rated physicians are on average less likely to bill big organizations than low-rated ones, -1.2% in $1(org_{jt} \geq 10)$ and -2.1% smaller in $\log(org_{jt})$. There are caveats to the causal interpretation of the findings. I cannot rule out the alternative story that the physicians decide to work harder in their small groups first and then receive higher ratings.

Additional Tables and Figures

Table A.1: Summary Statistics for Matched and Unmatched Yelp Physician Listings Containing “MD”, “DO”, or “OD”

	Matched Listing	Unmatched Listing
Yelp Number of Reviews by 2017	5.47 (med=2)	4.44 (med=2)
Avg Yelp Ratings By 2017	3.59 (med=4)	3.53 (med=4)
First Year of Rating	2013 (med=2013)	2013 (med=2014)
N	30,729	13,179

Notes: The tables above display the summary statistics among the matched physicians versus the unmatched ones. I consider Yelp physician listings with suffixes “MD”, “DO”, or “DO” as individual physicians. The first column displays the mean and median (in parentheses) of each variable for the listings that are matched with the NPI directory. The second column displays the statistics for those that are unmatched.

Table B.1: Top 10 Topics of Reviews from Yelp for High Rating Reviews

Topic Number	Probability	Top Relevant Keywords	Subjective Interpretation
1	10.6%	doctor, know, like, get, can, go, good, really, just, thing	generic
2	10.4%	care, doctor, best, patient, ever, physician, recommend, one, knowledge, family	generic
3	8.7%	question, answer, time, feel, concern, make, rush, take, always, listen	office amenities, interpersonal skills
4	8.6%	call, appoint, wait, office, day, minute, schedule, time, hour, get	office amenities, interpersonal skills
5	7.5%	staff, office, friendly, great, nice, front, help, always, profession,	office amenities interpersonal skills
6	6.4%	feel, thank, made, amazing, comfort, make, say, enough, life, team	generic
7	6.2%	recommend, highly, manner, bedside, profession, staff, great, excel, procedure, explain	office amenities, interpersonal skills
8	5.9%	problem, pain, help, diagnose, life, issue, treatment, condition, symptom, prescribe	clinic related
9	5.3%	result, look, breast, surgeon, procedure, consult, plastic, done, botox, happy	clinic related
10	5.0%	year, now, see, ago, move, since, go, 10, mother, drive	generic

Notes: The sample includes 155,993 reviews for all physicians with suffixes “MD,” “OD,” or “DO” with ratings higher than or equal to 4 stars. The model assumes that there are in total 20 topics and runs through the LDA algorithms 200 times. Topic numbers are ranked by the probability that a Yelp review in the sample is classified according to each topic. “Avg rating” refers to the weighted average of each review rating over all reviews, weighted by the probability of the focal topic belonging to each review. Top relevant words are derived using the formula and modules provided by Sievert & Shirley (2014), setting $\lambda = 2/3$, which is the weight balancing a keyword’s probability in a topic, and its probability in a topic divided by the overall probability of the keyword in all usage. See Appendix B for details. Subjective interpretation consists of my personal interpretation of the keywords of each topic.

Table B.2: Top 10 Topics of Reviews from Yelp for Low Rating Reviews

Topic Number	Probability	Top Relevant Keywords	Subjective Interpretation
1	13.7%	said, ask, told, go, went, just, doctor, want, back, see	generic
2	13.0%	doctor, patient, go, people, will, like, care, can, know, money	generic
3	9.3%	appoint, see, schedule, time, doctor, new, month, year, cancel, patient	office amenities, interpersonal skills
4	9.2%	wait, minute, hour, time, room, appoint, min, late, 15, arrive	office amenities, interpersonal skills
5	8.4%	call, phone, office, back, get, day, answer, message, said, told	office amenities, interpersonal skills
6	7.4%	medic, care, physician, patient, issue, treatment, condition, health, doctor, year	clinic related
7	7.0%	rude, staff, office, front, ever, worst, desk, service, unprofessional, horrible	office amenities, interpersonal skills
8	4.8%	insurance, bill, pay, charge, company, cover, paid, visit, office, fee	office amenities, interpersonal skills
9	4.3%	manner, seem, bedside, question, like, nice, feel, rush, good, friendly	office amenities, interpersonal skills
10	4.2%	review, yelp, write, read, experience, neg, base, post, bad, posit	generic

Notes: The sample includes 69,571 reviews for all physicians with suffixes "MD," "OD," or "DO" with ratings lower than or equal to 2 stars. The model assumes that there are in total 20 topics and runs through the LDA algorithms 200 times. Topic numbers are ranked by the probability that a Yelp review in the sample is classified according to each topic. "Avg rating" refers to the weighted average of each review rating over all reviews, weighted by the probability of the focal topic belonging to each review. Top relevant words are derived using the formula and modules provided by Sievert & Shirley (2014), setting $\lambda = 2/3$, which is the weight balancing a keyword's probability in a topic, and its probability in a topic divided by the overall probability of the keyword in all usage. See Appendix B for details. Subjective interpretation consists of my personal interpretation of the keywords of each topic.

Table B.3: Correlations between Yelp Ratings and Physician Clinical Ability, Controlling for Medical School Rankings

	(1)	(2)
Dep Variables:	Board	Log(#Accreditations)
Ratings	0.0270***	0.0280***
	(0.00480)	(0.00834)
Startclass	0.00101***	6.86e-05
	(8.01e-05)	(0.000243)
Observations	8,755	4,405
R-squared	0.211	0.293

	(3)	(4)	(5)	(6)	(7)
Dep Variables:	Eye Exam	Mammogram	PQI	Risk Score (Charlson)	Risk Score (CMS)
Ratings	0.00142*	0.00560***	-0.000769***	-0.0124***	-0.00747***
	(0.000776)	(0.00109)	(0.000243)	(0.00279)	(0.00173)
Startclass	0.000200***	0.000288***	-4.29e-05***	-0.000570***	-0.000352***
	(2.24e-05)	(2.52e-05)	(4.72e-06)	(5.99e-05)	(3.82e-05)
Observations	810,464	751,746	3,013,423	2,891,537	2,891,537
R-squared	0.058	0.067	0.056	0.411	0.386

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation results of equations (1) in columns 1–2 and (2) in columns 3–7, including StartClass medical school rankings as an additional regressor, which is from the Physician Compare data. The estimation sample in column 1 includes the Healthgrades data for primary care physicians matched with the main rated physician sample using a physician's last name, first name, and HSA. The dependent variable in column 1 is an indicator variable for whether or not physician j is board certified. For column 2, the sample includes Physician Compare performance data from 2015 matched with the main rated physician sample using NPI numbers. The dependent variable is the log of the number of self-reported quality indicators of physician j. Columns 1–2 are estimated at the physician level, include physician specialty and HSA fixed effects, and are two-way clustered at the specialty and HSA levels. Columns 3–7 include the main rated physician sample linked with Medicare Part B non-institutional claims data between 2008 and 2015. The dependent variable for columns 3 and 4 relate to whether a patient receives eye exams and mammograms among eligible diabetic patients and female patients younger than 74. The dependent variable for column 5 is an indicator of whether patient i receives preventable inpatient admissions in year t, defined by the numerators of the PQI index from AHRQ. The dependent variables for columns 6 and 7 are the computed Charlson and CMS-HCC-2015 risk scores using a patient's medical history up to year t. Columns 5–9 are estimated at the patient-year level, include physician specialty and HSA fixed effects, and are two-way clustered at the physicians' HSA and specialty levels.

Table C.1: Correlations between Instruments and Time-Varying Observable Physician Characteristics (Panel)

	(1)	(2)
Dep Variables:	Eye Exam	Mammogram
z_{jt}	-0.00122	-0.000335
	(0.00277)	(0.00279)
Observations	85,042	88,091
R-squared	0.385	0.475

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation result of equation (14). The estimation sample includes the main rated physician sample merged with Medicare claims data between 2008 and 2015 only for physicians who are ever rated on Yelp. The dependent variable for column 1 relates to whether a diabetic patient i receives an eye exam, documented at http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2017-table-of-contents/diabetes-care. The dependent variable for column 2 indicates whether an eligible female receives breast cancer mammograms in the past two years, documented at http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2017-table-of-contents/breast-cancer. Physician, physician practice HSA fixed effects are included. Standard errors are clustered at the physicians' HSA levels. Specialty levels are not further included as fixed effects and clustered since the sample already only focuses on primary care physicians.

Table C.2: Correlations between Instruments and Time-Unvarying Observable Physician Characteristics (Cross-Sectional)

	(1)	(2)	(3)	(4)
Dep Variables:	Board	Usnews	Startclass	Log(#Accreditations)
z_{jt}	0.00666	-0.371	-0.0289	0.0344
	(0.0116)	(0.284)	(0.610)	(0.0316)
Observations	2,988	12,669	12,669	1,421
R-squared	0.302	0.265	0.291	0.488

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation result of equation (15). The sample restricts to the estimation sample in equation (3) merged with Healthgrades primary care physician data in column 1 and Physician Compare data in columns 2–4. It includes observations among physicians who are ever rated and during their first rating year. The dependent variable for column 1 is an indicator variable related to whether physician j is board certified. In columns 2 and 3, a physician's ranking of medical school is obtained by manually matching her medical school information from Physician Compare with rankings from Usnews or StartClass. I reverse the order so that the best schools receive the highest number and unranked schools are input as 0. For column 4, the dependent variable is the log of the number of self-reported quality indicators of physician j. Physician specialty and practice HSA fixed effects are included. Standard errors are two-way clustered at the physicians' practice HSA and specialty levels.

Table C.3: Correlations between Being Rated and Time Varying Observable Physician Characteristics (Panel)

	(1)	(2)
Dep Variables:	D_{jt}	D_{jt}
EyeExam	-0.00465	
	(0.00495)	
Mammogram		0.00113
		(0.00671)
Observations	85,042	88,091
R-squared	0.698	0.697

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation result of equation (16). The estimation sample includes the main rated physician sample merged with Medicare claims data 2008-2015 and only estimate on physicians who are ever rated. The independent variable for column 1 is whether a diabetic patient i receives an eye exam, documented in http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2017-table-of-contents/diabetes-care. The independent variable for column 2 is whether an eligible female receives breast cancer mammograms in the past two years, documented in http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2017-table-of-contents/breast-cancer. Physician, physician practice HSA fixed effects are included. Standard errors are clustered at the physicians' practice HSA levels. Specialty levels are not further included as fixed effects and clustered since the sample already only focuses on primary care physicians.

Table C.4: Correlations between Being Rated and Time Unvarying Observable Physician Characteristics(Cross-Sectional)

	(1)	(2)	(3)	(4)
Dep Variables:	D_{jt}	D_{jt}	D_{jt}	D_{jt}
Board	0.00781			
	(0.00769)			
Usnews		0.000191**		
		(9.47e-05)		
Startclass			7.56e-05	
			(5.59e-05)	
log(# accreditations)				-0.00887
				(0.00844)
Observations	31,301	113,037	113,037	17,171
R-squared	0.267	0.229	0.229	0.323

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation result of equation (17). The sample restricts to the estimation sample in equation (3) merged with Healthgrades primary care physician data in column 1 and Physician Compare data in columns 2–4. It includes observations among physicians who are ever rated and during their first rating year. The independent variable for column 1 an indicator variable whether physician j is board-certified. In columns 2–3, a physician's ranking of medical school is obtained by manually matching her medical school information from Physician Compare with rankings from Usnews or StartClass. I reverse the order so that the best schools receive the highest number and unranked schools are input as 0. For column 4, the independent variable is the log of the number of self-reported quality indicators of physician j. Physician specialty and practice HSA fixed effects are included. Standard errors are two-way clustered at the physicians' practice HSA and specialty levels.

Table C.5: Regression Results for Equation (3)—Effects of Yelp Ratings on Patient Flow Using Alternative Instruments

	(1)	(2)	(3)	(4)
Dep Variables:	Log(Revenue)		Log(#Unq Patients)	
Method	OLS	IV	OLS	IV
βR_{jt}	0.0134***	0.0303***	0.00781***	0.0193**
	(0.00346)	(0.0103)	(0.00253)	(0.00880)
D_{jt}	-0.00653	-0.00664	-0.00368	-0.00375
	(0.00780)	(0.00807)	(0.00782)	(0.00609)
Observations	3,415,214	3,415,214	3,415,214	3,415,214
R-squared	0.913	0.913	0.922	0.922

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation result of equation (3). The estimation sample combines the main rated physician sample with the Medicare physician payment data from 2012 to 2015 and include all physicians with positive payments. The sample is restricted to physicians from the Yelp default popular city list where the data of the alternative instrument defined in equation (13) is collected. Columns 1 and 2 use the log of a physician's total revenue in a year as the dependent variable. Columns 3 and 4 use the log of a physician's number of unique patients as the dependent variable. Columns 1 and 3 use OLS estimates. Columns 2 and 4 use IV estimates according to the alternative instrument defined in equation (13). The regressions include physician fixed effects, physician specialty-specific time fixed effects, and physician HSA-specific time fixed effects. Standard errors are two-way clustered at the physicians' HSA and specialty levels.

Table C.6: Regression Results for Equation (18), (19), and (20)—Heterogeneity Effects of Yelp Ratings on Patient Flow

	(1)	(2)	(3)	(4)	(5)	(6)
Dep Variables:	Log(Revenue)	Log(#Unq Patients)	Log(Revenue)	Log(#Unq Patients)	Log(Revenue)	Log(#Unq Patients)
Method	OLS	OLS	OLS	OLS	OLS	OLS
$R_{jt}D_{jt}$	-0.0192***	-0.0162***	0.0124***	0.00740***	0.0109***	0.00679***
	(0.00357)	(0.00299)	(0.00228)	(0.00171)	(0.00251)	(0.00198)
$R_{jt}D_{jt}n_{jt}^{0.5}$	0.0277***	0.0208***				
	(0.00422)	(0.00309)				
$R_{jt}D_{jt}(a_j-\mu_a)$			-0.0235	-0.0218**		
			(0.0148)	(0.00979)		
$\frac{1}{100}R_{jt}D_{jt}(e_j$-$\mu_e)$					0.0127	0.00540
					(0.00973)	(0.00665)
Observations	3,474,061	3,474,061	3,221,427	3,324,994	3,221,427	3,221,427
R-squared	0.914	0.923	0.907	0.910	0.907	0.911
	(7)	(8)	(9)	(10)	(11)	(12)
Dep Variables:	Log(Revenue)	Log(#Unq Patients)	Log(Revenue)	Log(#Unq Patients)	Log(Revenue)	Log(#Unq Patients)
Method	IV	IV	IV	IV	IV	IV
$R_{jt}D_{jt}$	-0.0398***	-0.0333***	0.0190**	0.0123**	0.0174**	0.0114**
	(0.0149)	(0.0109)	(0.00740)	(0.00487)	(0.00734)	(0.00504)
$R_{jt}D_{jt}n_{jt}^{0.5}$	0.0506***	0.0396***				
	(0.0109)	(0.00920)				
$R_{jt}D_{jt}(a_j-\mu_a)$			-0.0407**	-0.0354***		
			(0.0171)	(0.0125)		
$\frac{1}{100}R_{jt}D_{jt}(e_j-\mu_e)$					0.0137	0.00657
					(0.00893)	(0.00682)
Observations	3,474,061	3,474,061	3,474,061	3,323,682	3,221,427	3,221,427
R-squared	0.914	0.914	0.923	0.910	0.907	0.911

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation result of equations (18), (19), and (20). The estimation sample combines the main rated physician sample with the Medicare physician payment data from 2012 to 2015 and include all physicians with positive payments. Columns 1, 3, 5, 7, 9 and 11 use the log a physician's total revenue in a year as the dependent variable. Columns 2, 4, 6, 8, 10, and 12 use the log of a physician's number of unique patients as the dependent variable. Columns 1, 2, 7, 8 correspond to regression 18. Columns 3 4, 9, and 10 correspond to regression 19. a_j, the average share of 85+ patients in physician j's Medicare patient pool between 2012 and 2015, is de-meaned. Columns 5, 6, 11, and 12 correspond to regression 20. e_j, the national percentile of the percentage of elderly with bachelor's degrees in physician j's county in 2016 ranging from 1 to 100, is de-meaned. Columns 1–6 use OLS estimations and columns 7–12 use IV estimations with $z_{it}D_{jt}$ and $z_{it}D_{jt}$ interacted with the additional variable $n_{jt}^{0.5}$, a_j, or e_j as the instruments. The regressions include physician fixed effects, physician specialty-specific time fixed effects, and physician HSA-specific time fixed effects. Columns 1 and 2 also include the number of reviews fixed effects. Standard errors are two-way clustered at the physicians' HSA and specialty levels.

Table C.7: Regression Results for Equation (21)—Yelp Ratings on Patient Flow Using Previous Year Ratings

	(1)	(2)	(3)	(4)
Dep Variables:	Log(Revenue)		Log(#Unq Patients)	
Method	OLS	IV	OLS	IV
R_{jt-1}	0.0171***	0.0315***	0.0128***	0.0242***
	(0.00206)	(.00850)	(0.00183)	(0.00797)
D_{jt-1}	-0.0340***	-0.0321***	-0.0217***	-0.0201***
	(0.00559)	(0.00570)	(0.00370)	(0.00386)
Observations	3,473,803	3,473,803	3,473,803	3,473,803
R-squared	0.914	0.914	0.923	0.923

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation results of equation (21). The sample combines the main Yelp-NPI sample with the Medicare physician payment data from 2012 to 2015 and include all physicians with positive payments. Columns 1 and 2 use the log of a physician's total revenue in a year as the dependent variable. Columns 3 and 4 use the log of a physician's number of unique patients as the dependent variable. Columns 1 and 3 use OLS estimation. Columns 2 and 4 use instrumental variable estimation. The regressions include physician fixed effects, physician specialty-specific time fixed effects, and physician HSA-specific time fixed effects. Standard errors are two-way clustered at the physicians' HSA and specialty levels.

Table D.1: Regression Results for Equation (25)—How Future Ratings Compare to Initial Ratings

	(1)	(2)	(3)	(4)
Dep Variables:	New Ratings	Cum Ratings	New “Harshness”	Cum “Harshness”
$1(t > d_j^f)$	0.128***	0.0464***	-0.0265*	0.00574
	(0.0198)	(0.00464)	(0.0159)	(0.00382)
Observations	67,351	162,843	64,328	159,378
R-Squared	0.514	0.909	0.386	0.890

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation result of equation 25. The estimation sample is the main rated physician sample in which an observation is at the physician year level ranging from a physician's first year with ratings until June 2017. Column 1 uses the average new ratings received in year t as the dependent variable. Column 2 uses the cumulative average ratings up to year t. Column 3 uses the average new “harshness” instrument received in year t. Column 4 uses the cumulative average “harshness” instrument received up to year t. In columns 1 and 3, only years in which a physician obtains a new rating/“harshness” instrument are included in the estimation. All columns use OLS estimation. The regressions include physician fixed effects, physician specialty-specific time fixed effects, and physician HSA-specific time fixed effects. Standard errors are two-way clustered at the physicians' HSA and specialty levels.

Table D.2: Regression Results of Equation (22)—Balance Check of Physicians First Rated at Different Years

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dep Variables:	Log (Org Size)	Log (Revenue)	Log (# Visits)	$Outpatient PV	$Lab PV	$Opioid PV	#ER	Charlson	CMS
T^m	-0.0715	-0.0223	-0.0396	-4.703	-0.273	-2.151	-0.0356***	-0.0794***	-0.0435***
	(0.0488)	(0.0242)	(0.0286)	(6.462)	(1.830)	(1.445)	(0.00758)	(0.0210)	(0.0107)
Observations	3,923,327	3,923,327	3,923,327	3,923,327	3,923,327	2,106,819	3,923,327	3,484,182	3,484,182
R-squared	0.441	0.397	0.372	0.056	0.082	0.044	0.027	0.054	0.062
Sample	All	All	All	All	All	All	All	All	All
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Dep Variables:	Log (Org Size)	Log (Revenue)	Log (# Visits)	$Outpatient PV	$Lab PV	$Opioid PV	#ER	Charlson	CMS
T^m	-0.128	-0.0407	-0.0331	6.925	2.589	0.767	-0.0326**	-0.0293	-0.0282
	(0.0854)	(0.0435)	(0.0487)	(12.28)	(3.676)	(2.628)	(0.0129)	(0.0372)	(0.0187)
Observations	1,638,946	1,638,946	1,638,946	1,638,946	1,638,946	902,092	1,638,946	1,444,303	1,444,303
R-squared	0.569	0.527	0.524	0.072	0.102	0.055	0.042	0.079	0.093
Sample	First Low	First Low	First Low	First Low	First Low	First Low	First Low	First Low	First Low

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation result of equation (22). In columns 1–9, the sample includes, among all cohorts m, the pre-existing before-m patients of treatment physicians and control physicians from Medicare claim in the period 2008–2015, before the treatment year m of each cohort. In columns 10–18, the sample is further restricted to physicians whose first ratings are lower than or equal to two stars. An observation is at the cohort-patient-year level. The dependent variables for columns 1 and 10 are the log of the organization size of physician j.The dependent variables of columns 2 and 11 are the log of the total revenue from Part B non-institutional claims of physician j from the Medicare claims data between 2008 and 2015. The dependent variables for columns 3 and 12 is the log of the number of visits from Part B non-institutional claims, where a visit is defined as a patient-physician encounter at the day level. The dependent variables for 4 and 13 are patient i's total Part B non-institutional outpatient spending in year t per primary care visit. The dependent variables for 5 and 14 are patient i's total Part B lab and imaging spending in year t per primary care visit. The dependent variables for 6 and 15 are patient i's total opioid prescription amount in year t per primary care visit.The dependent variables for 7 and 16 are patient i's total number of ER visits in year t. The dependent variables for 8 and 17 are patient i's Charlson comorbidity index in year t. The dependent variables for 9 and 18 are patient i's CMS-HCC-2015 risk scores in year t. Standard errors are clustered at the physicians' practice HSA levels. Specialty levels are not further clustered since the sample already only focuses on primary care physicians.

Table D.3: Regression Results of Equation (6)—How Patients' Health Services Received and Outcomes Differ by Years from Being First Rated

	(1)	(2)	(3)	(4)	(5)	(6)
Dep Variables:	\$Outpatient PV	\$Lab PV	\$Opioid PV	#ER	Charlson	CMS
$1(\text{t-m}\leq\text{-5})T_j^m$	-7.951	-0.493	0.304	0.000471	-0.0182	-0.0110**
	(6.087)	(1.619)	(2.617)	(0.00616)	(0.0122)	(0.00493)
$1(\text{t-m=-4})T_j^m$	-4.418	-0.696	1.150	-0.00148	-0.0157	-0.00534
	(5.047)	(1.307)	(2.093)	(0.00442)	(0.00960)	(0.00398)
$1(\text{t-m=-3})T_j^m$	-3.778	-1.541	1.582	-0.00110	-0.00983	-0.00332
	(4.212)	(0.970)	(1.773)	(0.00354)	(0.00697)	(0.00316)
$1(\text{t-m=-2})T_j^m$	-5.188	-0.525	1.289	-0.00208	-0.00649	-0.00479**
	(3.304)	(0.758)	(1.231)	(0.00267)	(0.00475)	(0.00213)
$1(\text{t-m}\geq 0)T_j^m$	8.973**	1.925*	-0.792	0.00137	-0.00550	-0.00152
	(3.979)	(1.071)	(1.578)	(0.00306)	(0.00649)	(0.00254)
Observations	5,532,809	5,532,809	3,115,249	5,532,809	5,031,196	5,031,196
R-squared	0.602	0.608	0.741	0.619	0.777	0.779
Pre-P	0.498	0.501	0.586	0.926	0.546	0.0776

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above contains the estimation results of equation (6) including $\omega_{\leq -5}$,...,ω_{-2} and $\omega_{\geq 0}$ on the right-hand side. The sample consists of, among all cohorts, the pre-existing patients of the treatment physicians and control physicians from Medicare claim in the period 2008–2015. The dependent variable for column 1 is patient i's total Part B non-institutional outpatient spending in year t per primary care visit. The dependent variable for column 2 is patient i's total Part B non-institutional spending related to lab and imaging in year t per primary care visit. The dependent variable for column 3 is, for patient i who enrolls in both Part B and D, the total Part D spending on opioids in year t per primary care visit. The dependent variable in column 4 is patient i's total number of ER visits in year t. The dependent variable in column 5 is patient i's risk score calculated using the Charlson model. The dependent variable in column 6 is patient i's risk score calculated using the CMS-HCC 2015 model. ω_{-1} is normalized to 0. the joint p value of pre-trend coefficients $\omega_{\leq -5}$,...,ω_{-2} are showed in the last row. Standard errors are clustered at the physicians' practice HSA levels. Specialty levels are not further clustered since the sample already only focuses on primary care physicians.

Table D.4: Regression Results of Equation (23)—How Patients' Health Services Received and Outcomes Differ by Years from Being First Rated, Assuming Linear Pre-Trend

	(1)	(2)	(3)	(4)	(5)	(6)
Dep Variables:	$Outpatient PV	$Lab PV	$Opioid PV	#ER	Charlson	CMS
$1(t<m)\cdot(t-m)T_j^m$	1.313	0.00102	-0.126	-0.000638	0.00380	0.00199*
	(1.356)	(0.366)	(0.568)	(0.00135)	(0.00265)	(0.00110)
$1(t\geq m)T_j^m$	9.095**	2.426**	-1.241	0.00363	-0.00767	-0.00276
	(3.649)	(0.994)	(1.574)	(0.00321)	(0.00664)	(0.00275)
Observations	5,532,809	5,532,809	3,115,249	5,532,809	5,031,196	5,031,196
R-squared	0.602	0.608	0.741	0.619	0.777	0.779

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above shows the estimation results of equation (23). The sample includes, among all cohorts, the pre-existing patients of the treatment physicians and control physicians from Medicare claim in the period 2008–2015. The dependent variable for column 1 is patient i's total Part B non-institutional outpatient spending in year t per primary care visit. The dependent variable for column 2 is patient i's total Part B non-institutional spending related to lab and imaging in year t per primary care visit. The dependent variable for column 3 is, for patient i who enrolls in both Part B and D, the total Part D spending on opioids in year t per primary care visit. The dependent variable in column 4 is patient i's total number of ER visits in year t. The dependent variable in column 5 is patient i's risk score calculated using the Charlson model. The dependent variable in column 6 is patient i's risk score calculated using the CMS-HCC 2015 model. Standard errors are clustered at the physicians' practice HSA levels. Specialty levels are not further clustered since the sample already only focuses on primary care physicians.

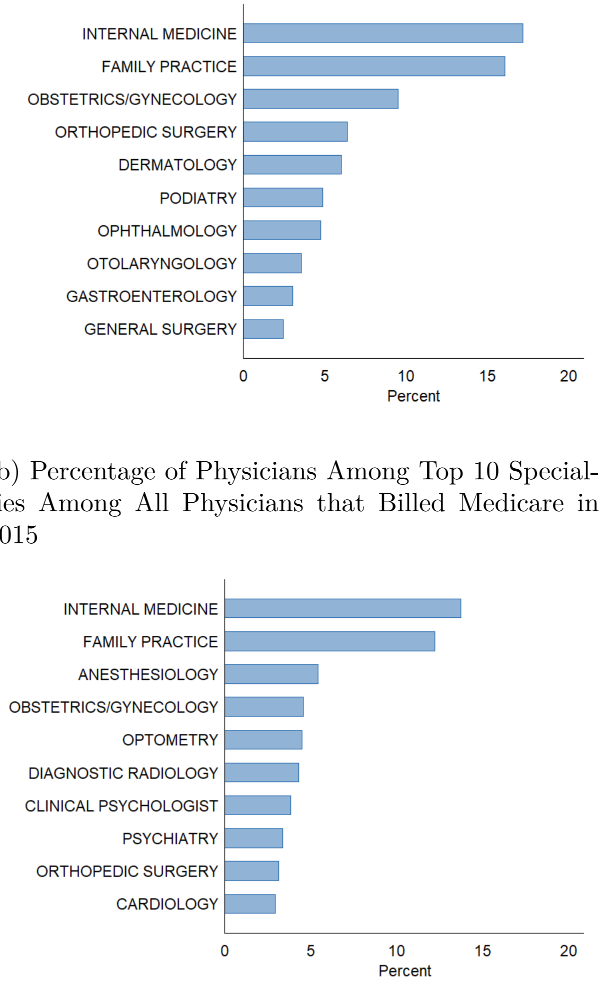
Table D.5: Regression Results for Equation (26)—How Physicians Organization Sizes Associate With Ratings Received of Different Years

	(1)	(2)	(3)	(4)
Dep Variables:	$1(org_{jt} \geq 10)$	$1(org_{jt} \geq 10)$	$\log(org_{jt})$	$\log(org_{jt})$
$\frac{1}{100}D_{jt-2}R_{jt-2}$	-0.000989	-0.286***	0.202	-1.04***
	(0.121)	(0.0773)	(0.494)	(0.278)
$\frac{1}{100}D_{jt-1}R_{jt-1}$	-0.184**	-0.324***	-0.723**	-0.936***
	(0.0581)	(0.0573)	(0.297)	(0.357)
$\frac{1}{100}D_{jt}R_{jt}$	-0.261**	-0.161**	-0.595**	-0.427***
	(0.0794)	(0.0651)	(0.288)	(0.441)
$\frac{1}{100}D_{jt+1}R_{jt+1}$	0.0174	-0.0249	-0.222	-0.0638
	(0.0491)	(0.0478)	(0.265)	(0.318)
$\frac{1}{100}D_{jt+2}R_{jt+2}$	-0.0434	-0.0466	.170	-0.00495
	(0.0657)	(0.0402)	(0.347)	(0.272)
Observations	226,489	226,489	226,489	226,489
R-Squared	0.878	0.878	0.898	0.898
Measures	Stock	Flow	Stock	Flow

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The table above contains the estimation results of equation (26). The estimation sample is all physicians from Medicare Part B non-institutional claims between 2008 and 2015 whether they are ever rated or never rated. An observation is at the physician-year level. org_{jt} is a measure of a physician's organization size. Columns 1 and 2 use $1(org_{jt}) \geq 10$ as the dependent variable. Columns 3 and 4 use $\log(org_{jt})$ as the dependent variable. Columns 1 and 3 are the stock measures of ratings, using cumulative average ratings of physician j up to the end of year t. D_{jt} denotes whether physician j has a cumulative average rating by year t. Columns 2 and 4 are the flow measures, using average ratings of physician j in year t. D_{jt} denotes whether physician j has an average new rating in year t. All columns use OLS estimation. The regressions include physician fixed effects, physician specialty-specific time fixed effects, and physician HSA-specific time fixed effects. Standard errors are two-way clustered at the physicians' HSA and specialty levels.

Figure A.1: Top 10 Specialties of Physicians that Bill Medicare

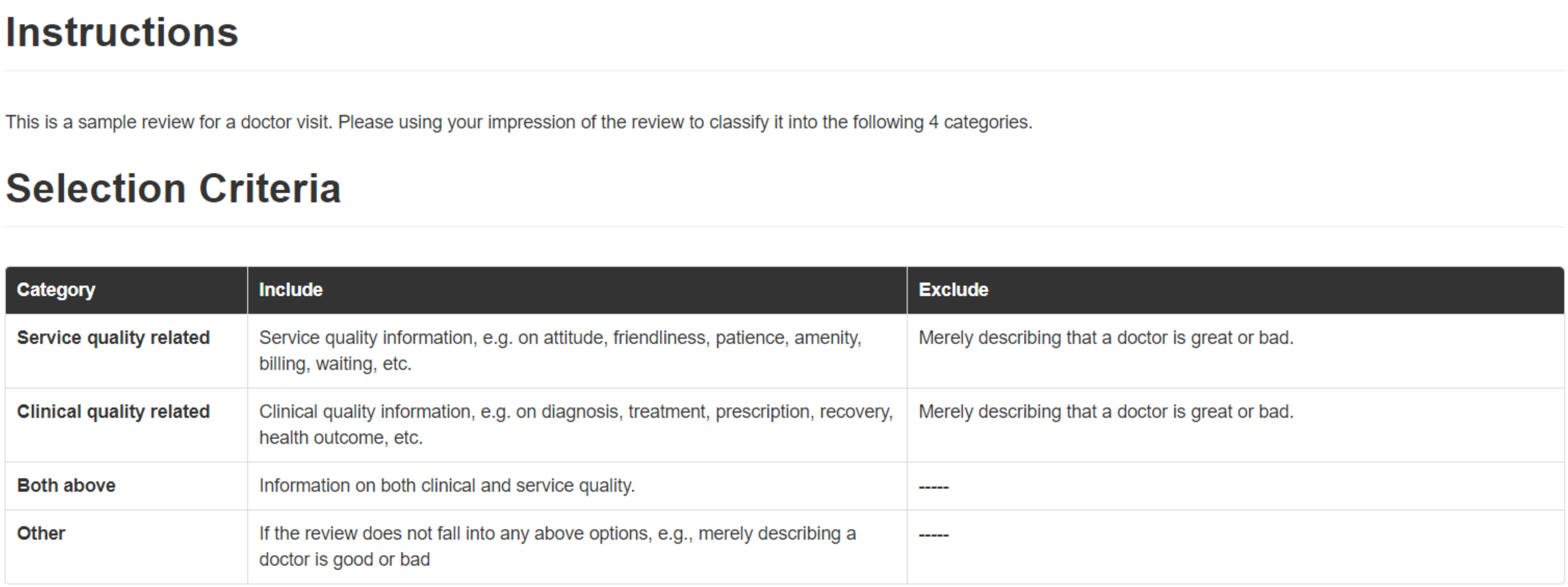


(a) Percentage of Physicians Among Top 10 Specialties Among Physicians that Billed Medicare in 2015 in the Matched Sample

(b) Percentage of Physicians Among Top 10 Specialties Among All Physicians that Billed Medicare in 2015

Notes: The figures above display the percentages of physicians of the top 10 specialties among the matched physicians that billed Medicare from the main sample in Panel (a) and among all physicians that billed Medicare in 2015 in Panel (b). In Panel (a), the main sample is matched with Physician Compare national demographics data in 2015, which contains demographic information of physicians that billed Medicare in 2015. In Panel (b), the specialties directly come from Physician Compare national demographics data in 2015. In both panels, the analysis excludes health workers whose specialties are listed as nurses, physician assistants, or social workers.

Figure B.1: Amazon Mechanical Turk Survey Questionnaire



Instructions

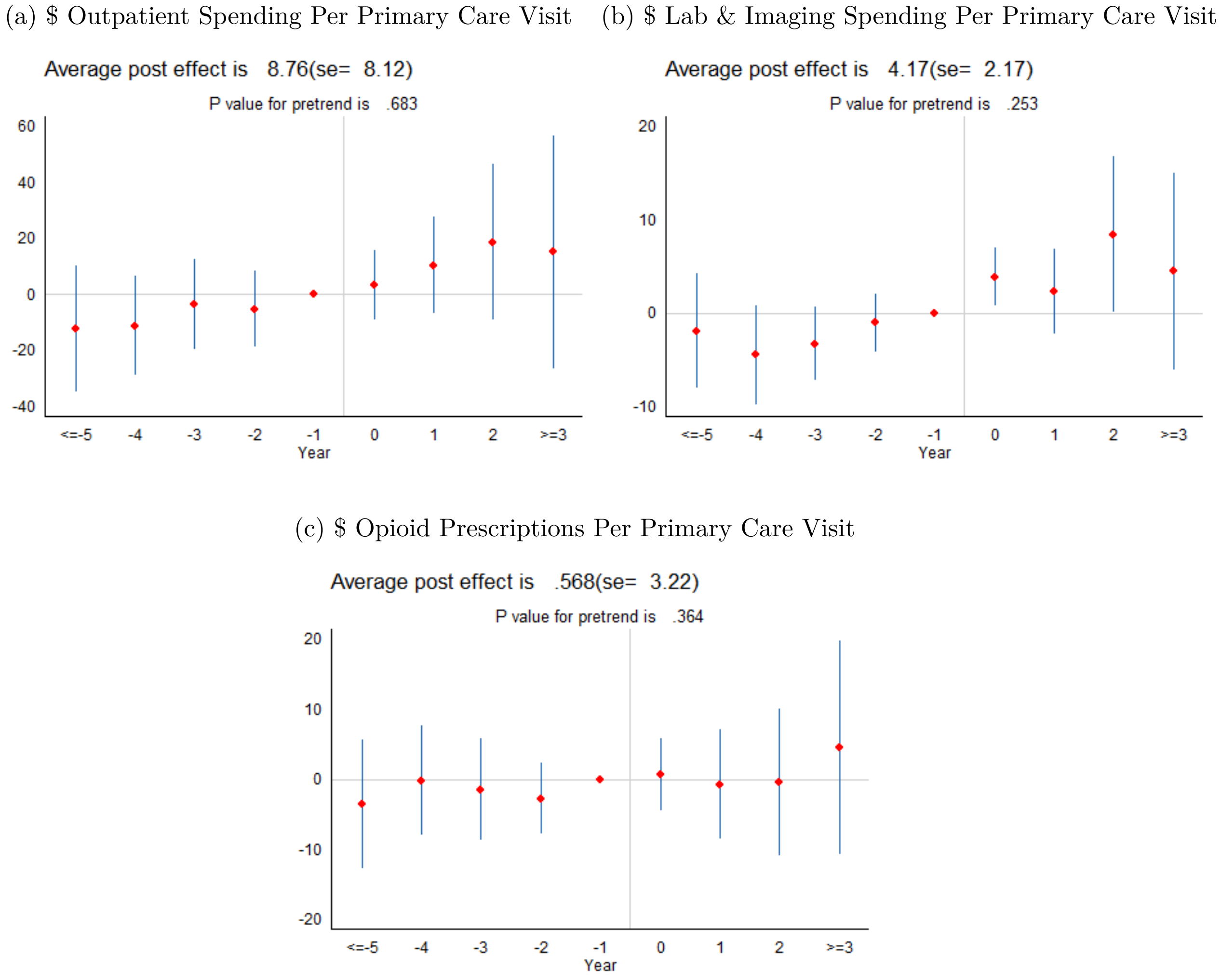
This is a sample review for a doctor visit. Please using your impression of the review to classify it into the following 4 categories.

Selection Criteria

Category	Include	Exclude
Service quality related	Service quality information, e.g. on attitude, friendliness, patience, amenity, billing, waiting, etc.	Merely describing that a doctor is great or bad.
Clinical quality related	Clinical quality information, e.g. on diagnosis, treatment, prescription, recovery, health outcome, etc.	Merely describing that a doctor is great or bad.
Both above	Information on both clinical and service quality.	-----
Other	If the review does not fall into any above options, e.g., merely describing a doctor is good or bad	-----

Notes: The figure above contains the survey of the questionnaire used in Amazon Mechanical Turk.

Figure D.1: Estimation Results of Equation (6)—Event Study of How Patient Health Services Received Change by Years Since Being First Reviewed, for Physicians with Low First Ratings

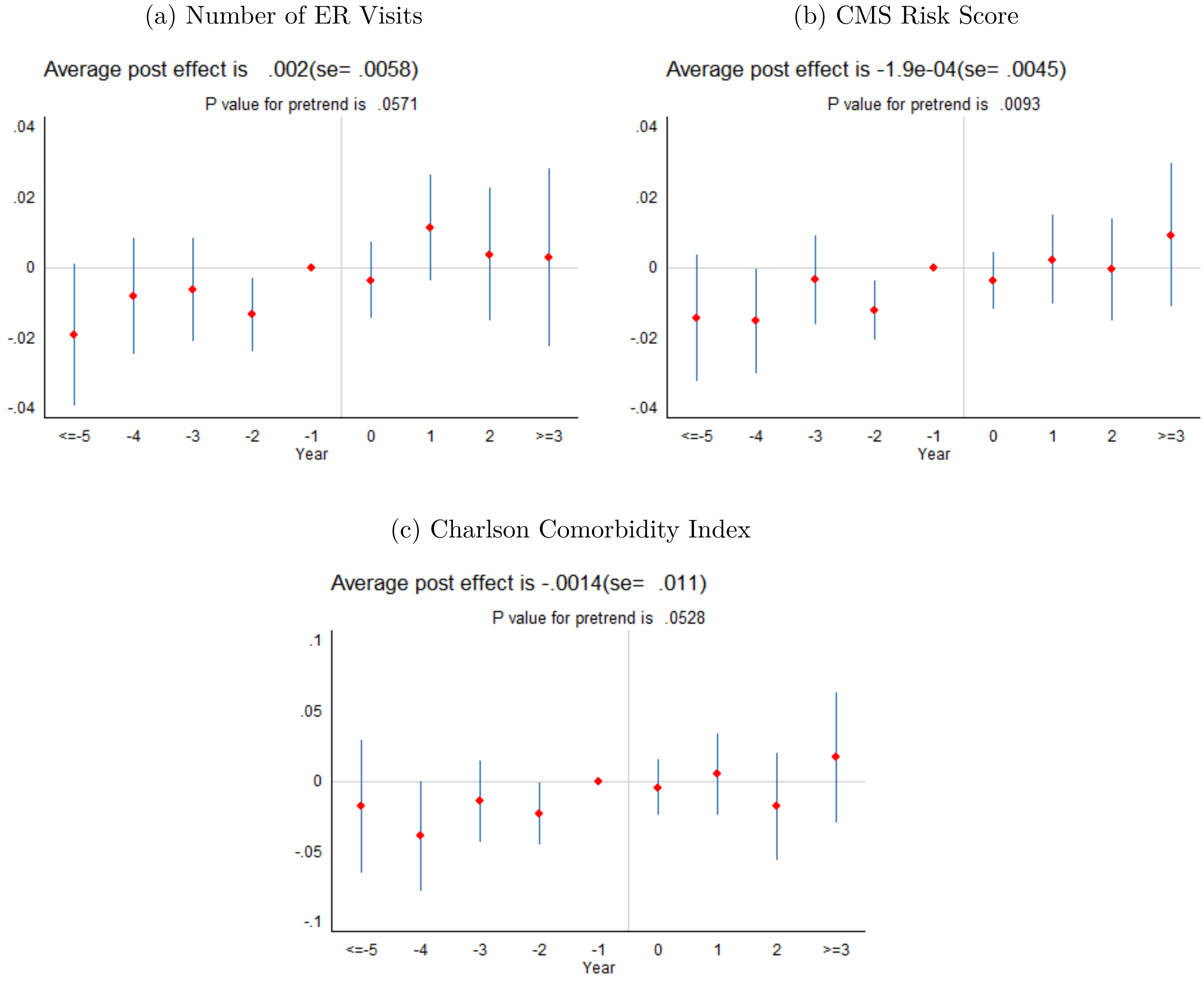

(a) $ Outpatient Spending Per Primary Care Visit

(b) $ Lab & Imaging Spending Per Primary Care Visit

(c) $ Opioid Prescriptions Per Primary Care Visit

Notes: The figures above contain the estimation results of equation (6). The sample includes, among all cohorts, the pre-existing patients of the treatment physicians and control physicians from Medicare claim in the period 2008–2015. It is further restricted to physicians whose first ratings are lower than or equal to 2 stars. An observation is at the cohort-patient-year level. The dependent variable in Panel (a) is patient i's total Part B non-institutional outpatient spending in year t per primary care visit. The dependent variable in Panel (b) is patient i's total Part B non-institutional spending related to lab and imaging in year t per primary care visit. The dependent variable for Panel (c) is, for patient i who enrolls in both Part B and D, the total Part D spending on opioids in year t per primary care visit. k, the number of years since first ratings, is plotted on the x-axis. Each dot in the figure corresponds to ω_k on the y-axis with the 95% confidence intervals plotted on blue lines. ω_{-1} is normalized to 0. In a regression including only ω_{-5},...,ω_{-2} and $\omega_{\geq 0}$ as the right-hand side instead of all flexible ωs, the estimated coefficient of $\omega_{\geq 0}$ and the joint p value of pre-trend coefficients are included in the titles and subtitles. Standard errors are clustered at the physicians' practice HSA levels. Specialty levels are not further clustered since the sample already only focuses on primary care physicians.

Figure D.2: Estimation Results of Equation (6)—Event Study of How Patient Health Outcomes Change by Years Since Being First Reviewed, for Physicians with Low First Ratings



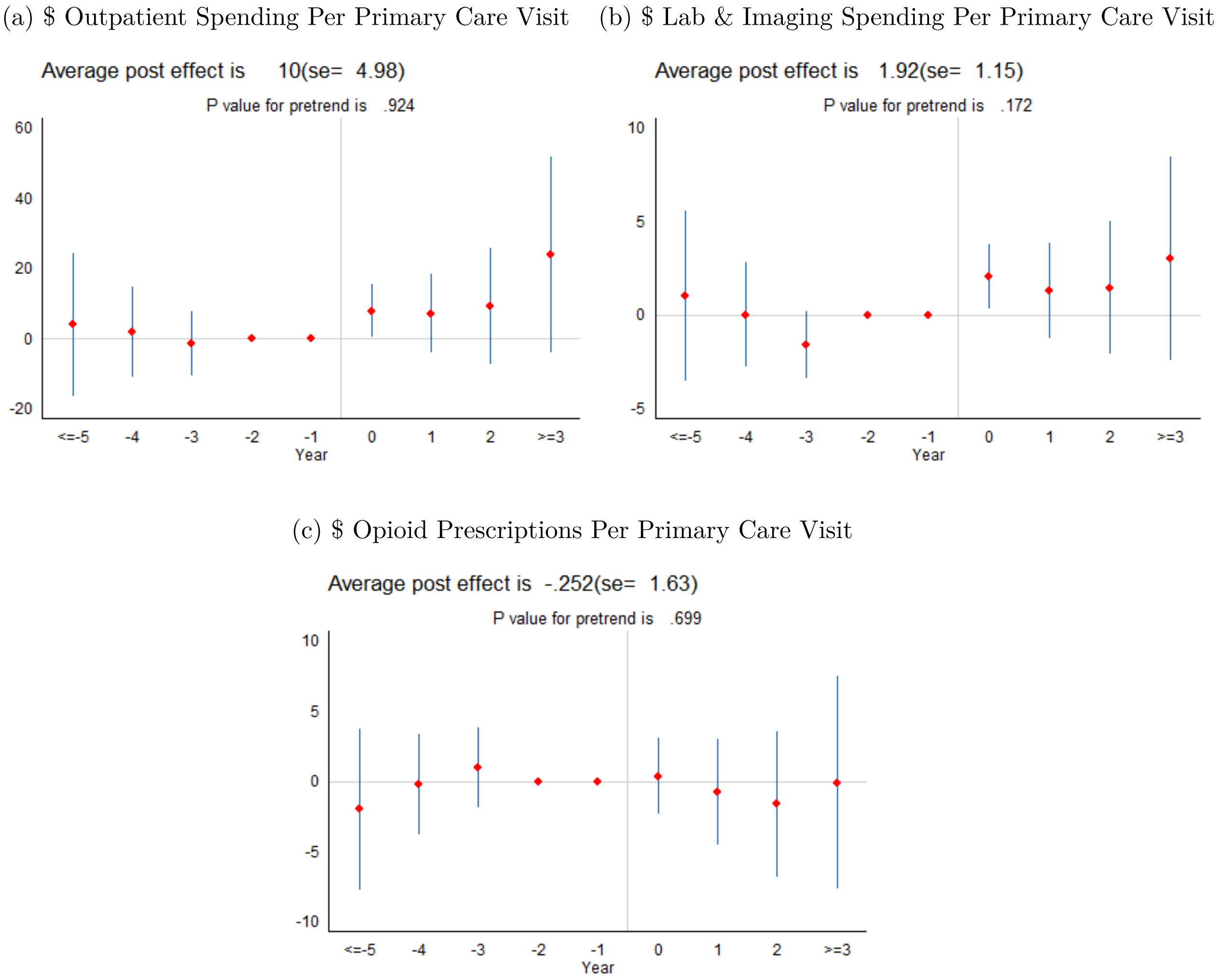
(a) Number of ER Visits

(b) CMS Risk Score

(c) Charlson Comorbidity Index

Notes: The figures above display the estimation results of equation (6). The sample includes, among all cohorts, the pre-existing patients of the treatment physicians and control physicians from Medicare claim in the period 2008–2015. It is further restricted to physicians whose first ratings are lower than or equal to 2 stars. An observation is at the cohort-patient-year level. The dependent variable in Panel (a) is patient i's total number of ER visits in year t. The dependent variable in Panel (b) is patient i's risk score calculated using the CMS-HCC 2015 model. The dependent variable for Panel (c) is patient i's risk score calculated using the Charlson model. k, the number of years since first ratings, is plotted on the x-axis. Each dot in the figure corresponds to ω_k on the y-axis with the 95% confidence intervals plotted on blue lines. ω_{-1} is normalized to 0. In a regression including only ω_{-5},...,ω_{-2} and $\omega_{\geq 0}$ as the right-hand side instead of all flexible ωs, the estimated coefficient of $\omega_{\geq 0}$ and the joint p value of pre-trend coefficients are included in the titles and subtitles. Standard errors are clustered at the physicians' practice HSA levels. Specialty levels are not further clustered since the sample already only focuses on primary care physicians.

Figure D.3: Estimation Results of Equation (6)—Event Study of How Patient Health Services Received Change by Years Since Being First Reviewed, Correcting Pre-trend from Patient Health Characteristics



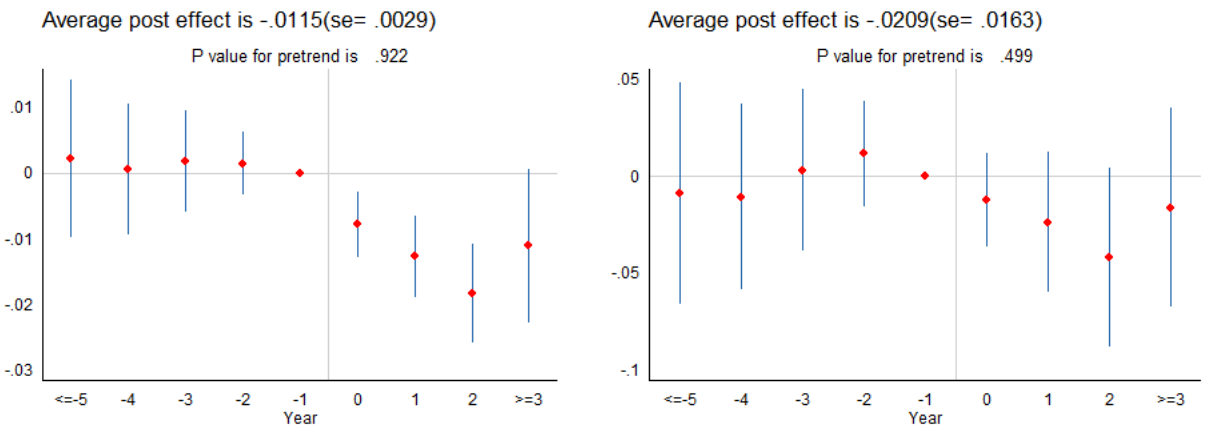
(a) $ Outpatient Spending Per Primary Care Visit

(b) $ Lab & Imaging Spending Per Primary Care Visit

(c) $ Opioid Prescriptions Per Primary Care Visit

Notes: The figures above contain the estimation results of equation (24) using method from Freyaldenhoven et al. (2018). The sample includes, among all cohorts, the pre-existing patients of the treatment physicians and control physicians from Medicare claim in the period 2008–2015. An observation is at the cohort-patient-year level. The dependent variable in Panel (a) is patient i's total Part B non-institutional outpatient spending in year t per primary care visit. The dependent variable in Panel (b) is patient i's total Part B non-institutional spending related to lab and imaging in year t per primary care visit. The dependent variable for Panel (c) is, for patient i who enrolls in both Part B and D, the total Part D spending on opioids in year t per primary care visit. k, the number of years since first ratings, is plotted on the x-axis. Each dot in the figure corresponds to ω_k on the y-axis with the 95% confidence intervals plotted on blue lines. ω_{-1} and additionally ω_{-2} are normalized to 0 since the lead of the event is included as an excluded instrument. In a regression including only ω_{-5},...,ω_{-2} and $\omega_{\geq 0}$ as the right-hand side instead of all flexible ωs, the estimated coefficient of $\omega_{\geq 0}$ and the joint p value of pre-trend coefficients are included in the titles and subtitles. Standard errors are clustered at the physicians' practice HSA levels.

Figure D.4: Estimation Results of Equation (27)—Event Study of How Physician Organization Size Changes by Years Since Being First Reviewed



(a) $1(org_{jt}) \geq 10$

(b) $\log(org_{jt})$

Notes: The figures above display the estimation results of equation (27). The estimation sample includes all physicians from Medicare Part B non-institutional claims between 2008 and 2015 whether rated or not. An observation is at the cohort-physician-year level. org_{jt} is a measure of physician's organization size. The left Panel (a) uses $1(org_{jt}) \geq 10$ as the dependent variable. The right Panel (b) uses $\log(org_{jt})$ as the dependent variable. The x-axis plots k, the kth year with respect to the first rated year. Each dot in the figure corresponds to ω_k on the y-axis with the 95% confidence intervals plotted on blue lines. ω_{-1} is normalized to 0. In a regression including only ω_{-5},...,ω_{-2} and $\omega_{\geq 0}$ as the right-hand side instead of all flexible ωs, the estimated coefficient of $\omega_{\geq 0}$ and the joint p value of pre-trend coefficients are showed in the title and subtitle. Standard errors are two-way clustered at the physicians' HSA and specialty levels.