McCarthy, Ian M.; Sanbower, Kaylyn R.; Sánchez Aragón, Leonardo

Working Paper

Online reviews and hospital choice

EAG Discussion Paper, No. EAG 22-1

Provided in Cooperation with:
Economic Analysis Group (EAG), Antitrust Division, United States Department of Justice

Suggested Citation: McCarthy, Ian M.; Sanbower, Kaylyn R.; Sánchez Aragón, Leonardo (2022) : Online reviews and hospital choice, EAG Discussion Paper, No. EAG 22-1, U.S. Department of Justice, Antitrust Division, Economic Analysis Group (EAG), Washington, DC

This Version is available at:
https://hdl.handle.net/10419/284001

ECONOMIC ANALYSIS GROUP DISCUSSION PAPER

Online Reviews and Hospital Choices

Ian McCarthy,[1] Kaylyn Sanbower,[2] and Leonardo Sánchez Aragón[3]

EAG 22-1 December 2022

[1]Department of Economics, Emory University, Rich Building 319, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: immccar@emory.edu

[2]Economic Analysis Group, Antitrust Division, U.S. Department of Justice. E-mail: kaylyn.sanbower@usdoj.gov

[3]Facultad de Ciencias Sociales y Humanísticas, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador. E-mail: lfsanche@espol.edu.ec.

Online Reviews and Hospital Choice[§]

Ian McCarthy[¶] Kaylyn R. Sanbower[‖] Leonardo Sánchez Aragón[**]

December 5, 2022

Abstract

Information problems in health care and the multifaceted nature of hospital quality complicate hospital choice. Online reviews provide an accessible, salient means through which researchers and health care decision-makers can gather information about a hospital's quality of care, and given their increasing popularity, these measures may affect hospital choice and may have implications for hospital prices. Using the universe of hospital Yelp reviews and inpatient claims data for elective procedures in Florida from 2012 through 2017, we exploit exogenous variation in online hospital ratings over time to identify the effect of online reviews on hospital choice. We find that among admissions for elective, inpatient procedures, patients are willing to travel between 5 and 30 percent further to receive care from a hospital with a higher Yelp rating, relative to other hospitals in the market. We also find evidence that higher ratings translate into higher commercial payments from insurers, albeit with relatively modest magnitudes. Our results indicate that novel, accessible sources of quality information have the potential to affect health care decisions, with potential downstream effects on health care prices.

Keywords: quality disclosure; hospital choice; online reviews; hospital prices

[§]This research greatly benefited from feedback from Martin Gaynor, Amanda Starc, and participants at the Carolina Region Empirical Economics Day and the ASHEcon Annual Conference. We are grateful to Pablo Estrada for excellent research assistance.

[¶]Department of Economics, Emory University, Rich Building 319, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: immccar@emory.edu

[‖]Economic Analysis Group, Antitrust Division, U.S. Department of Justice. The views expressed herein are not purported to reflect those of the U.S. Department of Justice. E-mail: kaylyn.sanbower@gmail.com

[**]Facultad de Ciencias Sociales y Humanísticas, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador. E-mail: lfsanche@espol.edu.ec.

1 Introduction

Information problems complicate decision making processes, particularly in the context of complex products and services. One such service is hospital care, where difficulties in measuring and communicating quality and other characteristics result in incomplete information among patients and even health care providers. Government initiatives including websites and provider report cards attempt to improve upon the paucity of quality information, but they may not encompass the breadth of features of care relevant to hospital choice (Dafny and Dranove, 2008; Dranove and Sfekas, 2008; Jin and Sorensen, 2006). When selecting a hospital, decision makers may value clinical and non-clinical aspects of care, and if existing measures are incomplete, then metrics that provide new and relevant information or present existing information in a more accessible way may affect hospital choice (Dranove and Jin, 2010; Cutler, 2011; Garthwaite et al., 2020). Online review platforms offer one such source of information (Ranard et al., 2016).

In this paper, we examine the effects of online reviews on hospital choice. We do so using rich inpatient claims data and online Yelp reviews from 2012 through 2017. During that time, Yelp was a particularly popular site for hospital reviews, with hospital profiles on the platform growing ten-fold from 2010 to 2018, representing nearly 50% of all general acute care hospitals by year-end 2018. Yelp also provides a compelling source of exogenous variation for causal inference. The star ratings presented on the platform are an average of the prior user reviews, rounded to the nearest half-star. We use this rounding to construct an instrument for the percentile rank of a hospital's star rating in its respective market. We capture hospital choice using Florida inpatient claims data for planned or elective procedures—namely labor

and delivery, and orthopedic surgery—because these patients are presumably better able to shop for a hospital than those with urgent or emergency admissions. We then estimate a discrete choice model using a control function approach, where patient utility depends on a hospital's star rating relative to others in the market, among many other factors. We discuss details and the motivation of our identification strategy in more detail in Section 3.

We find that online reviews meaningfully affect hospital choice. Specifically, we find that labor and delivery patients are willing to travel an additional 0.68 miles for a standard deviation increase in percentile rank of hospital star ratings, which is approximately a half-star increase. With an average distance of 8.8 miles among these patients, this represents a 7.7% increase in willingness to travel. We similarly find that orthopedic surgery patients are willing to travel 4.4 more miles, which is 33.9% further for a standard deviation increase in rating. The different magnitudes between these two results correspond to the nature of labor and delivery in contrast to orthopedic surgery, with orthopedic surgery patients travelling farther distances for inpatient stays on average. The results for both procedures are robust to alternative specifications and different market definitions. Falsification analyses that estimate the model in the context of emergency department admissions find null results, which indicates that the main findings are not simply spurious and lends further confidence to the conclusion that online reviews affect hospital choice.

We then extend our analysis to consider the effect of online reviews on hospital prices. In a bilateral negotiation between insurers and hospitals, it follows that any influence of online reviews on demand should also influence hospital pricing. We evaluate this theoretical prediction by pairing hospital Yelp ratings with estimated hospital prices from the

Healthcare Cost Report Information System (HCRIS).[1] Due to the inherent limitations of our pricing and claims data, we cannot estimate the effects of ratings on hospital choice and prices as part of a single bargaining model. Instead, we examine the effects of ratings on prices separately from the analysis of ratings and choice, where we again exploit the nature of the Yelp rating algorithm to construct an instrument for a hospital's reported star rating in each year. With this instrumental variable strategy, we find that higher-rated hospitals are subsequently able to increase negotiated payments with commercial insurers. The magnitudes of our estimates are relatively small but reasonable given our demand estimates, empirical setting, and the existing literature. Our findings are also robust to several empirical concerns, including violations of the exclusion restriction; the presence of price outliers; and alternative specifications that include or exclude hospital fixed effects and controls for other measures of hospital quality from Hospital Compare.

In identifying these causal relationships, our analysis contributes to the growing literature on the effect of information disclosure in health care.[2] We find empirical evidence that online reviews affect hospital choice, which supports the notion that health care decision-makers are responsive to accessible, aggregate, patient-driven measures of quality (Chandra et al., 2016; Varkevisser et al., 2012; Dranove and Jin, 2010; Dafny and Dranove, 2008; Dranove and Sfekas, 2008; Jin and Sorensen, 2006). Studies of health care report cards find that providers with higher reported quality have increased market share and that this form of quality disclosure is informative to consumers (Cutler et al., 2004; Jin and Sorensen, 2006;

[1] We follow Dafny (2009) in our derivation of a price estimate from the HCRIS data. This estimate approximates the average non-Medicare insurance payment to a hospital for an inpatient stay. For brevity, we refer to this measure simply as "price" throughout.

[2] For a detailed summary of this literature across various industries, including health care, refer to Dranove and Jin (2010); this article covers both theoretical predictions and empirical analyses of quality disclosure.

Dafny and Dranove, 2008; Bundorf et al., 2009). Our analysis both lends support to those existing findings and provides new evidence on the effects of a novel source of information—online reviews—on hospital choice and subsequently hospital pricing.

Other studies highlight the difficulty in measuring and communicating hospital quality, and note that people are more responsive to overall ratings and measures of patient satisfaction as opposed to granular clinical measures (Dranove and Sfekas, 2008; Romley and Goldman, 2011; Scanlon et al., 2002; Pope, 2009; Chandra et al., 2016). For example, existing hospital quality measures consist primarily of process, outcome, and patient experience of care metrics. Process of care measures capture the extent to which the hospital treats its patients based on the best-known standards of care, whereas outcome of care measures communicate the results. The Centers for Medicare and Medicaid Services (CMS) collect these data for hospitals that receive Medicare payments. CMS also measures patient experience of care through the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, which collects information from hospital patients following an inpatient stay. Relatively recently, CMS sought to improve hospital quality disclosure through its website, "Hospital Compare," which made its various hospital quality measures publicly available and aggregated the measures into overall star ratings. However, industry concerns about the methodology behind the aggregate star ratings thwarted these efforts, and as such, it was not a consistently viable source of information over the sample period for this study (American Hospital Association, 2016, 2017).[3] The inconsistent availability of this information limits

[3]In November 2020, CMS announced that they would aggregate each of their independent "Compare" platforms into a single quality compare site titled "Care Compare." During the transition, CMS noted that they would not update the star ratings for hospitals until July 2021. See https://www.aha.org/news/headline/2020-11-12-cms-will-not-update-hospital-star-ratings-quality-data-january.

its potential to inform hospital choices.

Other efforts, such as hospital report cards and U.S. News and World Report ratings, work to synthesize and communicate clinical quality and HCAHPS patient satisfaction measures more accessibly. Pope (2009) shows that the U.S. News and World Report ratings do affect hospital choice; however, such overall ratings may not address all of the relevant features of hospital care because they are limited to clinical quality measures and structured survey instruments.[4] In the case of hospital report cards, Dranove and Sfekas (2008) argues that other studies find mixed results on the effect of report cards because they do not always disclose novel information. Romley and Goldman (2011) finds that non-clinical dimensions of the hospital experience impact hospital demand, but those features are not included in the existing metrics.[5] Additionally, these metrics report lagged quality information, which creates a notable barrier for decisions-makers who want to include timely information in their choice criteria.

Ultimately, our understanding of hospital quality disclosure is based on the prevailing measures that paint a fragmented picture of the hospital experience, thereby perpetuating the information problems that plague this market. This speaks to the potential importance of measures that provide novel, accessible, and relevant information on hospital care. To the extent that online reviews embody these characteristics, they are positioned to provide valuable information to the hospital selection process. Indeed, existing research finds that online

[4]See https://health.usnews.com/health-care/best-hospitals/articles/faq-how-and-why-we-rank-and-rate-hospitals.

[5]Conversely, there is a growing literature on the response to increased cost sharing in health care, which finds that patients do not shop for lower-cost providers (Brot-Goldberg et al., 2017; Desai et al., 2017; Mehrotra et al., 2017; Chernew et al., 2018). This indicates that forms of disclosure need to address relevant dimensions of care to affect choice and that demand responses in health care will likely be limited to novel quality information on non-clinical aspects of care.

reviews provide new information, addressing numerous dimensions of care not captured in the HCAHPS survey and highlight the potential for online reviews to speak to aspects of care that are relevant to decision-makers (Ranard et al., 2016).[6] Even if hospital reviews do not provide new information, they could affect choice if they make information more accessible. In health care, we can think of accessibility as both the disclosure itself and the ability of decision-makers to interpret the information once disclosed. Hospital reviews may make information more accessible in multiple ways. They disclose quality on a platform that is likely more familiar than the outlets used to communicate formal quality metrics, they can be updated in real time, and they mirror traditional "word-of-mouth" communication, which has been shown to affect hospital choice (Dellarocas, 2003; Moscone et al., 2012). Further, online reviews—and more specifically Yelp reviews—make information more accessible through aggregate star ratings and narratives of the patient perspective of care, which are poised to affect hospital choice.

A natural extension of our hypothesized effect on demand is an effect on hospital prices, yet we are not aware of any studies that directly examine the effect of quality information on prices for health care services.[7] One likely explanation for this gap in the literature is the aforementioned difficulty in defining and measuring hospital quality information which would tend to depress any estimated effects of such measures on price (Dranove and Jin, 2010; Romley and Goldman, 2011; Ranard et al., 2016). Our analysis makes a novel contribution in this area. Further, in examining the effects of ratings on prices, our study contributes

[6]Online reviews are imperfectly correlated with HCAHPS ratings and show little to no correlation with clinical quality measures, which emphasizes that these metrics likely communicate novel information (Bardach et al., 2013; Howard and Feyman, 2017; Campbell and Li, 2018; Perez and Freedman, 2018).

[7]The supplemental material provides a formal theoretical framework in which to analyze the relationship between information disclosure and prices.

to the broader literature on hospital pricing, much of which focuses on the role of hospital mergers or hospital acquisitions of other providers (Lin et al., 2020; Schmitt, 2018; Lewis and Pflum, 2017; Dafny, 2009; Capps et al., 2003; Gaynor and Vogt, 2003). Our contribution is to highlight another likely avenue by which hospitals can increase prices.

While economic theory predicts that decision makers will respond to information disclosure, in an opaque market such as health care, this is ultimately an empirical question. Online reviews possess characteristics that, in theory, should allow them to provide clarity to the decision making process, but in the absence of empirical evidence, that relationship is uncertain. Our analysis informs this open question, demonstrating that online reviews affect hospital choice, and subsequently, hospital prices. Given the significant efforts from the Centers for Medicare and Medicaid Services to provide relevant and accessible quality information on health care providers, understanding the potential effects of such information on hospital demand and the affordability of health care services is critical for future policy in this area.

2 Data

Our analysis on the effect of online reviews on hospital choices relies on two main sources of data. Online review data come from the rating platform, Yelp, which is well-suited for this analysis due to its popularity over the study period. Data on hospital choices come from Florida inpatient claims, which we limit to elective admissions for specific medical needs (namely, labor and delivery, and orthopedic surgery). The study also incorporates data from the American Hospital Association (AHA) Annual Survey—the most comprehensive

source of data on hospital characteristics—along with additional hospital features and quality measures from CMS. The following subsections first describe the data sources independently and then detail the final combined datasets used in our examination of hospital choice.

2.1 Online Reviews

Yelp has been a popular outlet for crowd-sourced information on various services and businesses over the past decade. Yelp launched in 2004 and within three years amassed one million reviews.[8] The platform continued to gain popularity through the 2010s, and while Google is now the most commonly used review site, Yelp appears to have been the most prominent source of online review information during the study period, which begins in 2012 and ends in 2017.[9]

A hospital appears on Yelp once its profile has been established, which can be done by a user, or a hospital, or a Yelp employee. Users may then review the hospital by leaving a star rating of 1 through 5 and a narrative comment. Both a star rating and a comment are required to post a review. Only users registered on Yelp may leave a review, but anyone can view them either through a search engine or looking directly on the site. When a visitor arrives to the site, they first see a summary of the hospital, which includes the number of reviews, an aggregate star rating, and other location and contact information. They can then click on the hospital to go beyond the summary and view each review that the hospital received. Yelp's algorithm determines the order in which reviews are presented, and it only presents reviews that are not deemed fraudulent. Any reviews identified as spam or

[8] See https://www.theatlantic.com/technology/archive/2011/07/infographic-the-incredible-six-year-history-of-yelp-reviews/242072/.

[9] See https://www.reviewtrackers.com/reports/online-reviews-survey/.

inauthentic are not included in the aggregate star rating and are available separately under the link "other reviews that are not currently recommended." The reviews were collected using web scraping methods to compile a dataset that consists of the date, the star rating, the review narrative, and the user ID for all of the hospital reviews on Yelp through year-end 2018.[10]

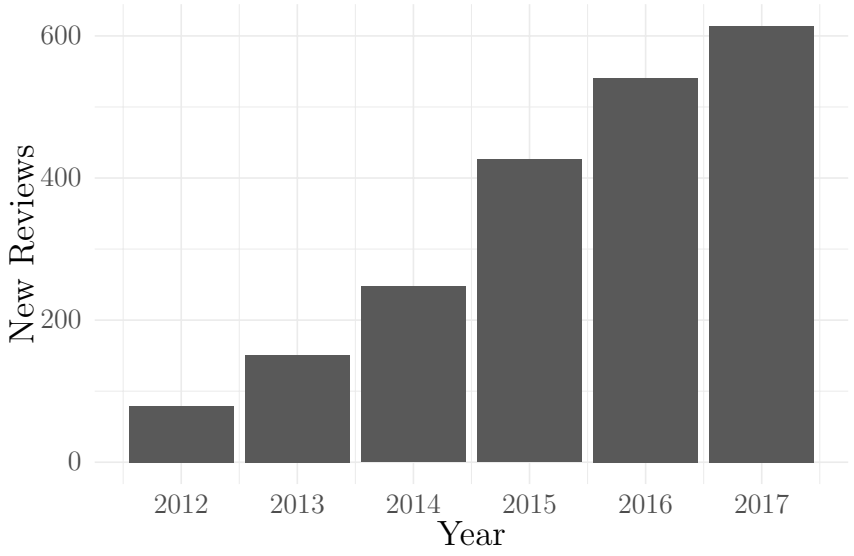
Using the star ratings from each of these reviews, we construct the aggregate rating for a hospital at any given point in time. The aggregate rating that a visitor would see is the average of a hospital's star ratings rounded to the nearest half-star. We refer to this value as the "observed" rating, and it takes on values between 1 and 5 in half-star increments. For example, a hospital with three ratings (1, 4, and 5 stars) has an average rating of 3.33, but the observed rating is 3.5 stars. Note that the underlying average rating is not directly observable to the visitor to the site, and therefore, the transformation of this average value to the nearest half-star provides plausibly exogenous variation to observed hospital ratings.

We limit the Yelp reviews to hospitals in Florida to correspond with the hospitals available in the claims data. Similar to the national level trends, less than 20% of Florida hospitals had a Yelp profile prior to 2012. By year-end 2012, nearly 25% of hospitals in the state were on the platform, and by the end of the study period (2017), this figure surpasses 50%. Figure 1 shows the number of new ratings posted to the platform by year. The growth in this figure can be explained both by the increase in the number of hospitals on the platform and the frequency with which hospitals are reviewed. For example, in 2012 there was an average of two new reviews per year for every hospital on the platform, but by 2017 there were eight new reviews per hospital. This speaks to the popularity of the platform and

[10]The supplemental material provides details on the data collection and cleaning processes.

provides evidence that people use it to share their experiences.

Figure 1: New Hospital Reviews on Yelp by Year



NOTES: The figure depicts the number of new reviews on Yelp for hospitals in Florida by year.

2.2 Inpatient Claims Data

Hospital choice data come from the Florida Agency for Health Care Administration (AHCA), which maintains claims data for the state of Florida. The data comprise the population of inpatient discharges from 2012 through 2017, including patient characteristics and diagnosis and procedure codes relevant to the admission. To maintain patient confidentiality, the AHCA omits patient identification, social security, and medical record numbers. They also withhold the patient's date of birth, and instead provide their age at the time of admission. Lastly, in lieu of admission and discharge dates, the agency discloses quarter of admission.[11]

We limit the data to elective admissions at general acute care hospitals. This eliminates

[11]Due to these confidentiality measures, we cannot identify repeat or first-time patients and the timing in the analysis can be no more granular than quarter level.

urgent or emergency room admissions, in order to isolate patients whose hospital choices were not impaired by the circumstances of their admission. We drop any admissions where the patient is discharged to court, law enforcement, or a psychiatric facility, as this may indicate that the patient had limited agency in selecting the hospital. Additionally, we eliminate observations with missing or foreign zip codes, along with any patients who are not from Florida and any admissions to hospitals that are not included in the AHA data.

To analyze the effect of online reviews on hospital choice, we focus on common procedures for which the patients have at least some freedom to select their hospital, as is the case for labor and delivery, and orthopedic surgery admissions.[12] Existing research on hospital choice has studied labor and delivery admissions because of the potential for patients to seek out information and scrutinize their options (Avdic et al., 2019). Additionally, text analysis of the Yelp data used for this study finds that labor and delivery is frequently discussed in the review narratives, indicating that it may be an influential information source for those planning child birth. For similar reasons, researchers have analyzed elective orthopedic surgery admissions to understand hospital choice (Gutacker et al., 2016). In the U.S. context, this procedure is particularly useful because it is common among Medicare fee-for-service patients, where insurance restrictions do not limit hospital choice. The following subsections discuss procedure-specific limitations on the data and summarize the online review data, claims data, and hospital characteristics that comprise the final dataset for the given procedure.

[12]The supplemental material details the ICD-9 and ICD-10 codes used to identify the admissions for the respective procedures.

2.2.1 Labor and Delivery Data

We limit the labor and delivery data to admissions for patients with ages between the 5th and 95th percentiles, which results in an age range of 20 to 38. The data include Medicaid, Medicaid HMO, and privately insured patients, where 49% have private insurance. We include each of these payer types because the data show that it is unlikely that hospitals are turning away Medicaid patients for this type of admission. The supplemental material provides details on the public insurance options in Florida. Lastly, we drop hospitals that average fewer than five labor and delivery admissions per quarter to limit the data to hospitals that are viable choices for this procedure. This leaves 361,040 labor and delivery admissions across 86 hospitals.

Table 1 summarizes the resulting dataset. The average patient in the data is nearly 29 years old and is equally likely to have Medicaid or private insurance. The majority of the patients are white, and nearly one third identify as Latina. The second panel in Table 1 includes hospital characteristics from AHA and CMS data. There are 86 hospitals in the sample, and they are more likely to be private, system hospitals, but are equally likely to be for-profit or non-profit. Further, 7% are major teaching hospitals and 51% satisfy a broader definition of teaching hospital, which, for example, includes hospitals with residency training approval and medical school affiliations. Of those hospitals, 46 of them have a Yelp presence at some point during the sample period. The average end-of-quarter observed rating is 2.9, and on average, a hospital receives 1.3 new reviews each quarter. While we can calculate a hospital's observed rating at any point in time, we present end-of-quarter observed ratings and number of new ratings per quarter to correspond with the unit of time in the claims

data, and subsequently, the unit of time used in the choice model.

Table 1: Summary Statistics: Labor and Delivery

	Mean	Median	St. Dev.	5th Pct.	95th Pct.
Patient Characteristics					
Age	28.81	29.00	4.82	21.00	37.00
Private Insurance	0.49	0.00	0.50	0.00	1.00
Black	0.19	0.00	0.39	0.00	1.00
Latina	0.29	0.00	0.45	0.00	1.00
Asian	0.02	0.00	0.14	0.00	0.00
Hospital Characteristics and Quality					
Total Beds	400.16	298.00	374.11	119.90	1007.75
Physicians	26.59	8.00	55.03	0.00	117.00
Nurses	719.35	450.00	878.14	177.65	2310.95
Government	0.11	0.00	0.32	0.00	1.00
Non-profit	0.50	0.00	0.50	0.00	1.00
Major Teaching Hospital	0.07	0.00	0.26	0.00	1.00
Any Teaching Hospital	0.51	1.00	0.50	0.00	1.00
System Member	0.83	1.00	0.38	0.00	1.00
Payer Mix	0.57	0.56	0.11	0.39	0.75
Case Mix Index	1.59	1.59	0.17	1.33	1.87
Hospital Wide Readmission Rate	15.99	15.90	1.06	14.40	17.80
Yelp Reviews					
Observed Rating	2.91	3.00	0.84	1.50	4.50
New Reviews	1.28	1.00	1.53	0.00	4.00

Notes: Patient characteristics are from inpatient claims data for labor and delivery admissions and are measured at the annual level. Over the sample period, there are 361,040 admissions. The hospital characteristics are measured annually. The Yelp review data are measured at the quarterly level to correspond with the unit of time available in the inpatient claims data.

2.2.2 Orthopedic Surgery Data

For orthopedic surgery admissions, we limit the data to Medicare fee-for-service (FFS) claims for knee or hip replacement among beneficiaries aged 65 and above.[13] By limiting the data to FFS admissions, we focus on patients whose choices are not restricted by specific insurance

[13]Information on the relevant diagnosis related group (DRG) codes is in the supplemental material.

networks.[14] We then omit any admissions in the top 5th percentile of age, which results in an age range of 65 to 85. Lastly, we limit the analysis to hospitals that average at least one orthopedic surgery admission per quarter, to eliminate any hospitals that are not viable choices.[15] The resulting sample comprises 128,862 admissions at 132 hospitals.

Table 2 summarizes these data. The average patient in the data is around 73 years old and more likely to be female. The patients are overwhelmingly white. The second panel shows that of the 132 hospitals in the sample, the majority are private, members of a hospital system, and about half of the hospitals have some teaching capacity. Over the sample period, 73 hospitals had a Yelp profile, with an average aggregate rating of 2.88, and an average of 1.22 new reviews per quarter. The data sources used for this table are analogous to those used for Table 1.

2.3 Hospital Markets

To analyze hospital choice, we also need to determine the relevant market for each respective patient.[16] Traditionally, hospital choice models rely on hospital referral regions (HRR), health service areas (HSA), and counties to define hospital markets. While these definitions are useful and suitable to certain analyses, they possess certain characteristics that make them less useful in this context. HRRs, for instance, are based on referral patterns for tertiary surgery and have not been updated since 1993. They are, therefore, unlikely

[14] Additional information on Medicare eligibility can be found here: https://www.cms.gov/Medicare/Eligibility-and-Enrollment/OrigMedicarePartABEligEnrol.

[15] This differs from the requirement of at least five admissions per quarter that we impose upon the labor and delivery data, because there are fewer orthopedic surgery admissions overall and that restriction would eliminate hospitals that appear to be viable options for this procedure.

[16] The use of the term "markets" here and throughout the duration of this paper is not meant to be interpreted as relevant antitrust markets.

Table 2: Summary Statistics: Orthopedic Surgery

	Mean	Median	St. Dev.	5th Pct.	95th Pct.
Patient Characteristics					
Age	73.26	73.00	5.49	65.00	83.00
Black	0.036	0.00	0.19	0.00	0.00
Latino	0.043	0.00	0.20	0.00	0.00
Asian	0.005	0.00	0.07	0.00	0.00
Male	0.394	0.00	0.49	0.00	1.00
Hospital Characteristics and Quality Measures					
Total Beds	332.97	249.00	324.30	84.00	835.00
Physicians	24.52	6.00	62.45	0.00	112.50
Nurses	571.14	373.00	750.23	113.90	1613.40
Government	0.10	0.00	0.30	0.00	1.00
Non-profit	0.42	0.00	0.49	0.00	1.00
Major Teaching Hospital	0.05	0.00	0.23	0.00	1.00
Any Teaching Hospital	0.47	0.00	0.50	0.00	1.00
System Member	0.86	1.00	0.35	0.00	1.00
Payer Mix	0.56	0.56	0.11	0.38	0.73
Case Mix Index	1.56	1.55	0.19	1.27	1.87
Hospital Wide Readmission Rate	16.03	15.90	1.04	14.40	17.87
Hip and Knee Replacement Readm. Rate	4.92	4.90	0.74	3.80	6.30
Yelp Reviews					
Observed Rating	2.88	3.00	0.96	1.00	5.00
New Reviews	1.22	1.00	1.60	0.00	4.00

Notes: Patient characteristics are from inpatient claims data for orthopedic surgery admissions and are measured at the quarter level. Over the sample period, there are 128,862 admissions. The hospital characteristics are measured annually. The Yelp review data are measured at the quarterly level to correspond with the unit of time available in the inpatient claims data.

to accurately capture hospital markets for the specific sets of procedures we study (Everson et al., 2019). Unlike HRRs, HSAs are based on annual Medicare inpatient hospital fee-for-service claims, which makes them better suited to capture current markets but less relevant for our understanding of hospital choices among expectant mothers, for example. Lastly, commonly used geographic boundaries such as counties or zip codes may not reflect a patient's procedure-specific choice set.[17]

Community detection (CD) algorithms provide a novel way for researchers to define hospital markets, tailored to a procedure-specific analysis (Everson et al., 2019). In the hospital context, community detection leverages patterns of patient flows to identify groups of hospitals that draw patients from common zip codes.[18] Note that this is essentially the same process that is used to determine HRRs and HSAs, but by using these methods to define markets instead of relying on existing definitions, researchers can gain valuable flexibility. CD methods allow the researcher to more precisely determine relevant markets and update these market definitions as often as their data allow, which may better reflect the competitive landscape and the observed choices of the hospitals in their analyses.

We employ these methods to define procedure-specific hospital markets using zip code level patient flows.[19] Separately for each procedure, we aggregate the claims data to determine the number of patients from a given zip code admitted to each hospital. We then use these aggregate values to implement the CD method and determine the relevant hospital

[17] These boundaries may be more appropriate when considering general inpatient acute care due to patients' interest in proximity to the hospital.

[18] The supplemental material describes the community detection method in greater detail.

[19] Please see https://github.com/graveja0/health-care-markets for an excellent resource that explains how to construct these markets. For more detail on our adaption of his code, please see the github repository: https://github.com/kaylynsanbower/hospital-marketshares. Note that this does not include the data used for this analysis per the terms of our data use agreement.

markets.[20]

Regardless of procedure, there are additional features of the data and the empirical setting that must be considered. First, note that these algorithms can define markets that consist of only a single hospital, and in the context of a hospital choice analysis, monopoly markets are uninformative. Further, markets with too few rated hospitals are of limited use in an analysis of reviews and choice, given that this effect is likely to depend on a hospital's rating *relative* to other hospitals in its market. Therefore, for each procedure, we use the broadest market definition from the community detection methods and then layer in additional restrictions. We limit the final sample to choice sets that have at least three hospitals on Yelp, which ensures that there are sufficient nearby hospitals on the platform for it to be a viable source of information. Additionally, we require that at least one hospital in the market has three or more reviews, because in order for a hospital to have an average rating that is not equal to a half-star increment, it must have at least three reviews.[21] Ultimately, these criteria serve to limit the analysis to markets where Yelp is a sufficiently popular platform.

The final consideration pertains to the distance to each hospital in a patient's choice set. Markets consist of groups of hospitals that draw patients from a common set of contiguous zip codes. This means that the market may include hospitals that are further away than the patient would realistically travel. Therefore, for the main results, we limit the patient's choice set to hospitals whose centroid distance from the patient's home zip code falls within

[20]The supplemental material details the CD methodology in the context of this analysis.

[21]In the market definitions used for the main specifications for both labor and delivery and orthopedic surgery, there are no rounded hospitals that are eliminated based on this criteria. This is important because it dispels concerns that the main source of exogeneity—the rounding—might be correlated with other variables that are relevant for limiting the sample.

a certain radius. Sections 4.1 and 4.2 detail their respective choice sets used for estimation, which include these restrictions and procedure-specific caveats.[22]

3 Empirical Approach

We investigate the effect of online reviews on hospital choice by modeling patient utility as a function of hospital Yelp ratings, relative to other hospitals in the market. These ratings can directly affect utility, meaning that the patient herself incorporates the star ratings into her decisions, or indirectly, through family, friends, and physicians, who gleaned information from this platform.[23] We estimate the following model where patient i's utility from receiving care at hospital j at time t is defined as:

$$\begin{aligned} u_{ijt} &= v_{ijt} + \epsilon_{ijt} \\ &= \beta_1 P_{j,t-1} + \beta_2 NR_{j,t-1} + D'_{ij}\alpha_d + H'_{jt}\alpha_h + Q'_{jt}\alpha_q + X'_{ijt}\alpha_x + \varepsilon_{ijt}. \end{aligned} \tag{1}$$

The first two terms of v_{ijt} capture a hospital's Yelp presence in $t-1$. Recall that the most granular unit of time for the hospital admissions data is quarter-year. Hence, t refers to the quarter of admission. If online reviews affect hospital choice, then this information must enter into the decision prior to admission. As such, the utility function captures a hospital's rating at the end of the prior quarter, i.e., $t-1$.[24] Specifically, $P_{j,t-1}$ is the percentile rank

[22] Given the importance of the market definition in this analysis, we also implement the econometric approach using the market definitions from other CD algorithms, FIPS codes, and HSAs in the supplemental material.

[23] The supplemental material presents information showing that people do in fact engage with the platform, which indicates that this information is likely relevant to some people involved in selecting a hospital.

[24] This allows up to three months for a patient to internalize these ratings, but one might be concerned that this timeline does not leave sufficient time for patients admitted at the beginning of the quarter to use

of a hospital's star rating among the hospitals in its market. By using the percentile rank instead of raw star ratings, we capture a hospital's quality information relative to other hospitals on the platform.[25] Further, because some hospitals in a patient's choice set are not on the platform, $NR_{j,t-1}$ is an indicator for whether or not the hospital is rated.

The utility function also includes D_{ij}, which is a vector of linear and squared centroid distances between the patient's home and the hospitals in her choice set. Hospital characteristics are represented by H_{jt}, which includes counts of total beds, physicians, nurses; indicators for government, for profit status, system members, and teaching hospitals; and payer mix.[26] The vector Q_{jt} controls for clinical quality using hospital readmission rates. For labor and delivery, Q_{jt} refers to hospital wide 30-day readmission rates, and in the case of orthopedic surgery, Q_{jt} also includes 30-day readmission rates for hip and knee replacement.

Lastly, to allow for a rich substitution pattern, X_{ijt} is a vector of hospital-level variables interacted with patient-level variables. These hospital variables include distance, total beds, case mix index, payer mix, and readmission rates. The individual-level variables applicable for both procedures are age, and race and ethnicity. For labor and delivery, we also include an indicator for public insurance because the data have a mix of privately and publicly insured patients. For orthopedics, we include an indicator for the sex of the patient. The error term, ϵ_{ijt}, is assumed to be i.i.d. Type I extreme value, which yields the common logit form for the probability of patient i selecting hospital j. Note that because patients must choose a hospital in order to appear in the data, there is no outside option. We estimate the

this information. The following section and the supplemental material include results with further lags. The results are unchanged, which may also reflect the limited flow of new reviews per quarter (1.3 on average).

[25]The percentile rank is calculated among rated hospitals, where the lowest rated hospital's percentile rank is $1/n$, and the highest ranked is 1. Non-rated hospitals have a zero percentile rank and an indicator to designate that they are not rated.

[26]Payer mix is the proportion of total discharges that do not come from Medicare or Medicaid.

underlying utility parameters using maximum likelihood.

We are interested in identifying the effect of Yelp star ratings on hospital choice. One concern, however, is that hospital star ratings—and percentile rankings based on these ratings—are endogenous. Hospital reputation, for example, is likely to affect choice and also likely correlated with Yelp ratings. The current model, therefore, will suffer from omitted variable bias.[27] Moreover, we are interested in the direct effect of star ratings on choice, not the relationship between underlying quality of various dimensions and hospital selection. To identify this effect, we implement an instrumental variable strategy. Recall that the star rating that a visitor to Yelp would see for a given hospital is the average of each of its individual reviews rounded to the nearest half star. Therefore, at the midpoint between each half-star increment, hospitals above are rounded up and those below are rounded down. This transformation generates exogenous variation in hospital ratings that we use to instrument for the percentile rank of a hospital's star rating. Specifically, we construct as an instrument an indicator for being rounded into a higher rating, where a hospital is considered rounded if it is within the range of the midpoint and 0.1 above the midpoint.

Using this instrument, we estimate the model using a control function approach, which conditions on the part of the observed rating that is correlated with other unobserved hospital characteristics relevant to hospital choice. The control function isolates exogenous variation in the percentile rank variable due to rounding and then controls for the remaining endogenous variation in the observed percentile rank by including the first stage residuals as an additional covariate, which enables consistent estimation of the percentile rank coefficient

[27] The supplemental material includes these results, which for both procedures are still positive and significant but have different magnitudes.

(Petrin and Train, 2010). To implement, we first estimate the following equation,

$$P_{j,t-1} = \gamma R_{j,t-1} + \zeta NR_{j,t-1} + D'_{ij}\psi_d + H'_{jt}\psi_h + X'_{ijt}\psi_x + \epsilon_{ijt}, \tag{2}$$

where a hospital's percentile rank is the function of the instrument—an indicator, R, for whether or not the hospital is rounded up into the next star rating—and all of the right-hand side variables included in Equation 1. We then include the residuals, $\widehat{\epsilon}_{ijt}$, in the following equation to recover a consistent estimate of β_1. The second stage, therefore, is

$$u_{ijt} = \beta_1 P_{j,t-1} + \beta_2 NR_{j,t-1} + D'_{ij}\alpha_d + H'_{jt}\alpha_h + X'_{ijt}\alpha_x + \widetilde{\varepsilon}_{ijt}, \tag{3}$$

where $\varepsilon_{ijt} = \eta\epsilon_{ijt} + \widetilde{\varepsilon}_{ijt}$, and $\widehat{\epsilon}_{ijt}$ is an estimate for ϵ_{ijt}. We use this specification for the main results. Additionally, to provide a more readily interpretable result, we use the coefficients from this model to calculate a patient's willingness to travel (WTT) for a standard deviation increase in percentile rank. This is defined by the negative marginal rate of substitution between percentile rank and the measures of distance multiplied by the average standard deviation of percentile rank across choice sets. For the main results, this is

$$WTT = -\frac{\partial U_{ij}}{\partial P_{ij}} \Bigg/ \frac{\partial U_{ij}}{\partial D_j} \times SD(P). \tag{4}$$

For Column (1) of our tables of results in Section 4, this is simply: $WTT = \frac{-\beta_1}{\alpha_d} \times SD(P)$. For Columns (2) through (4), i.e. the columns that present results with interactions between

individual characteristics and distance and distance squared, the measure is

$$WTT = \frac{-\beta_1}{\alpha_d + 2\alpha_{d^2} D + A + B} \times SD(P), \tag{5}$$

where A represents the terms of $\frac{\partial U_{ij}}{\partial D_j}$ that correspond to the interactions between distance and patient characteristics and B represents the terms of $\frac{\partial U_{ij}}{\partial D_j}$ that correspond to the interactions between distance squared and patient characteristics.[28] We calculate standard errors using the delta method. The following section details the choice sets, model results, and willingness to travel estimates by procedure.

4 Online Reviews and Hospital Choice

As discussed previously, we examine the effects of online reviews and hospital choice in three distinct clinical settings. First, in subsection 4.1, we consider hospital choice among labor and delivery patients. This analysis includes patients across different types of insurers, including Medicaid and private insurance. As such, while labor and delivery is widely perceived as an area where patients can exercise agency and discretion in hospital choice, this area is also one in which hospital choice may be restricted (and in unobserved ways) due to a patient's insurance network.

Second, in subsection 4.2, we consider hospital choice in elective orthopedic surgery. Like

[28] More specifically, $A = \psi_{d\times a}x^a + \psi_{d\times b}x^b + \psi_{d\times l}x^l + \psi_{d\times pm}x^{pm}$ and $B = 2\psi_{d^2\times a}Dx^a + 2\psi_{d^2\times b}Dx^b + 2\psi_{d^2\times l}Dx^l + 2\psi_{d^2\times pm}Dx^{pm}$. The subscripts on the ψ terms signify to which interaction term the coefficient corresponds. These terms consist of distance (d), distance-squared (d^2), age (a), a Black indicator (b), and a Latino indicator (l). For the labor and delivery analysis, this also includes a Medicaid indicator (pm), which is not applicable to the orthopedic surgery analysis. Similarly, for orthopedic surgery, there is a male indicator (m), which is not applicable for labor and delivery.

labor and delivery, this is an area in which patients likely exercise some discretion in hospital choice. Moreover, our analysis of orthopedic surgery is limited to Medicare fee-for-service patients, so that hospital choice is not restricted by insurance networks.

Finally, we consider online reviews and choice among emergency department visits in subsection 4.3. Given that non-deferrable emergency visits are dictated by acute need for immediate hospital care, this area serves as a falsification analysis in which we do not expect online ratings to influence hospital choice.

4.1 Choice in Labor and Delivery

We use patient flows to the 86 hospitals in the sample to identify labor and delivery-specific markets. The community detection algorithm identified 12 total markets, all of which contained more than one hospital.[29] While the data consist of elective admissions, given the possibility that an expectant mother may need to get to the hospital quickly, a hospital closer to the patient's home is likely preferable. This bears out in the data, which show that the average distance traveled to the chosen hospital is 11.2 miles (standard deviation 12.8). Based on the nature of this procedure, we drop hospitals from a patient's choice set if they are over 30 miles away. Lastly, as detailed in Section 2, we limit the sample to choice sets that have at least three rated hospitals, where one of which must have at least three reviews. This results in a final sample of 176,587 admissions to 49 hospitals across five markets. On average, a choice set in this data has between 9 and 10 hospitals, where about half of those hospitals have Yelp profiles. The average star rating among these hospitals is just under

[29]Note that over the sample period, there were 35 counties in Florida with hospitals that had admissions for labor and delivery, meaning that these markets are less granular than county definitions.

three stars, with an average of 17 reviews (median 13).

The first panel of Table 3 presents the estimation results for the first stage as shown in Equation 2. As suspected, the instrument for being rounded into a higher rating is positively and significantly related to percentile rank. In addition to the variables shown, each column consists of hospital characteristics (total bed, nurse, and physician counts; indicators for teaching hospital status and system membership; payer mix, and case mix index), hospital-wide readmission rates, and the distance between the patient's home zip code and the hospital. Interactions between these terms and individual characteristics (age, and indicators to capture if the patient is Black, Hispanic or Latina, or insured through Medicaid) are layered in as indicated. Note that when interactions with "distance variables" are included, the specification consists of distance, distance squared, and interactions between distance and individual characteristics.

We then implement the control function approach by including these residuals in Equation 3 (Petrin and Train, 2010). The second panel of Table 3 presents the results. The coefficients represent the marginal utilities for the average patient and can be informative about the direction of the effect of these characteristics on a patient's utility. Across each specification, utility is increasing in percentile rank. The interpretation for the distance coefficient requires more nuance. In column (1), where there are no interaction terms, the results show that utility is decreasing in distance, which corresponds with prior findings in the literature. In the subsequent columns, the inclusion of interaction terms between distance and individual characteristics inhibits the ability to interpret this directly. Instead, the bottom row of Table 3 presents the WTT results, where, in the most saturated specification, we find that patients are willing to travel an additional 0.68 miles to receive care from a hospital with

Table 3: Labor & Delivery Results

	(1)	(2)	(3)	(4)
Panel 1: First Stage Results				
Rounded Indicator	0.1425***	0.1404***	0.1394***	0.1391***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Not Rated	-0.5563***	-0.5575***	-0.5658***	-0.5650***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
F-Statistic	314934	159167	112143	101722
R^2	0.7470	0.7491	0.7545	0.7550
Panel 2: Discrete Choice Estimates				
Percentile Rank	0.2812***	0.2631***	0.2789***	0.3018***
	(0.0705)	(0.0719)	(0.0728)	(0.0735)
Not Rated	0.0565	0.0502	0.0831*	0.0833*
	(0.0420)	(0.0429)	(0.0441)	(0.0444)
Distance	-0.1559***	-0.1446***	-0.1325***	-0.1542***
	(0.0005)	(0.0104)	(0.0104)	(0.0105)
$Distance^2$		-0.0015***	-0.0018***	-0.0012***
		(0.0004)	(0.0004)	(0.0004)
Willingness to Travel	0.662***	0.611***	0.655***	0.680***
	(0.166)	(0.167)	(0.172)	(0.166)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X

NOTES: The first panel presents the first stage results, where the dependent variable across all specifications is percentile rank. The second panel presents the discrete choice estimates. In both panels, all specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

a one standard deviation higher percentile rank. For this procedure, a standard deviation is 0.37. Given that the average distance to the chosen hospital for patients in this sample is 8.8 miles, the WTT estimate represents an 7.7% increase in travel distance. On average, a standard deviation increase in percentile rank translates to a 0.58 increase in stars, meaning that a patient is willing to travel nearly 8% further for around a half-star increase in star ratings.

Note that the main results presented in Table 3 model choices based on ratings in the prior quarter, i.e. $t-1$. One consideration in the labor and delivery context, however, is that expectant mothers are likely choosing a hospital *earlier* in their pregnancy than the quarter prior to delivery. Table 4, therefore, presents results based on ratings in $t-2$ and $t-3$. Here we see that these decisions are most sensitive to ratings in $t-3$, which corresponds to the time during which expectant mothers are likely finding out that they are pregnant and are beginning to make care plans for their pregnancies. The higher responsiveness to this information in the quarters earlier in gestation lend confidence to the hypothesis that this information affects hospital choice.

This result furthers our understanding of how expectant mothers value the trade-off between distance and quality and is commensurate with existing estimates. Avdic et al. (2019) assesses the responses of mothers to clinical quality and patient satisfaction scores in Germany. For the patient satisfaction scores—which are most comparable to the quality measure used in this analysis—they find that expectant mothers are willing to travel an average of 0.55 km for a standard deviation increase in higher reported subjective quality. Compared to the average distance to the chosen hospital (10.76 km), this represents a 5.11% increase.

Table 4: Labor & Delivery Discrete Choice Estimates at $t-2$ and $t-3$

	(1)	(2)	(3)	(4)
Panel 1: Ratings at $t-2$				
Percentile Rank	0.4805***	0.4746***	0.4463***	0.4782***
	(0.0663)	(0.0675)	(0.0680)	(0.0686)
Not Rated	0.1660***	0.1674***	0.1759***	0.1810***
	(0.0387)	(0.0395)	(0.0403)	(0.0406)
Distance	-0.1562***	-0.1445***	-0.1322***	-0.1540***
	(0.0005)	(0.0104)	(0.0104)	(0.0105)
Distance^2		-0.0015***	-0.0018***	-0.0013***
		(0.0004)	(0.0004)	(0.0004)
Willingness to Travel	1.127***	1.101***	1.048***	1.076***
	(0.155)	(0.157)	(0.160)	(0.155)
Panel 2: Ratings at $t-3$				
Percentile Rank	0.6513***	0.6324***	0.6280***	0.6413***
	(0.0639)	(0.0650)	(0.0663)	(0.0667)
Not Rated	0.2531***	0.2474***	0.2713***	0.2650***
	(0.0365)	(0.0372)	(0.0384)	(0.0386)
Distance	-0.1561***	-0.1448***	-0.1323***	-0.1543***
	(0.0005)	(0.0104)	(0.0104)	(0.0105)
Distance(2)		-0.0015***	-0.0018***	-0.0013***
		(0.0004)	(0.0004)	(0.0004)
Willingness to Travel	1.52***	1.461***	1.470***	1.438***
	(0.149)	(0.151)	(0.156)	(0.150)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X

NOTES: The table presents the results corresponding to equation 3, but replaces percentile rank in $t-1$ with percentile rank in $t-2$ and then $t-3$. The first stage includes an indicator for being rounded up in the given quarter instead of $t-1$. Each column includes all of the same covariates as Table 3. All specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Standard errors on the willingness to travel measures are calculated using the Delta Method. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

These results are robust to a variety of modifications and alternative specifications, all of which are detailed in the supplemental material. We estimate the model where we replace distance with differential distance—i.e., the distance to a given hospital minus the distance to the closest hospital in the patient's choice set—and the estimates are nearly identical. Further, to assess the sensitivity of our results to the chosen market definition, we estimate the main specification with other market definitions based on community detection algorithms, FIPS codes, and radii around the patient's home zip code. The findings generally hold across various market definitions, insofar as the markets reflect patient flows, which is untrue of the markets defined solely on radius.

4.2 Choices in Orthopedic Surgery

We identify markets for orthopedic surgery using patient flows to the 132 hospitals in the sample. The community detection algorithm identified 14 markets, all of which contain more than one hospital. In comparison to labor and delivery, the data show that these patients are willing to travel much further for this procedure. The average travel distance to the chosen hospital is 14.9 miles with a standard deviation of 21.4, and the data for these choices are right skewed. To reflect this, we limit a patient's choice set to hospitals within a 100-mile radius. Limiting the sample to markets with at least three rated hospitals, one of which must have at least three reviews, we arrive at the final sample which contains 84,981 hospital admissions to 122 hospitals across 12 markets. The choice sets for orthopedic surgery have, on average, just under 11 hospitals, where around 5 of those are on Yelp. The rated hospitals have an average star rating of 2.8, with 15.3 reviews on average.

The results for the first stage are shown in the first panel of Table 5. They indicate that the instrument for being rounded into a higher rating is positively and significantly related to percentile rank. The covariates included in Table 5 correspond to those described for Table 3 with two exceptions. First, our orthopedic surgery admissions include both male and female patients, so we include an indicator for sex. Second, we limit the data to Medicare fee-for-service patients, eliminating the indicator for insurer type.

As previously described, we then use the residuals from the results in Panel 1 of Table 5 to implement the control function approach. The results are shown in the second panel of Table 5, where we see that the patient's marginal utility is increasing in percentile rank across each specification. As suspected, the effect of distance on utility in the first column is negative and significant. In the following columns, distance-squared and interaction terms preclude direct interpretation. However, the WTT estimates indicate that orthopedic surgery patients are willing to travel 4.4 additional miles (in the most saturated specification) for a standard deviation (0.37) increase in percentile rank. This is compared to an average distance of 12.97 miles to the chosen hospital, which translates to a 33.9% increase in travel distance. For this procedure, a standard deviation increase in percentile rank translates to an average increase of 0.69 stars. Given that Yelp presents ratings in half-star increments, this means that for between a half and full-star increase, a patient is willing to travel around 30% farther.

To our knowledge, much of the existing literature on hospital choice for elective, inpatient surgery focuses on clinical quality metrics and provides limited insight on the effects of measures that are geared toward non-clinical, subjective quality. For example, Moscelli et al. (2016) explores hospital choice for elective hip replacements in the English National Health

Table 5: Orthopedic Surgery Results

	(1)	(2)	(3)	(4)
Panel 1: First Stage Results				
Rounded Indicator	0.1278***	0.1279***	0.1278***	0.1278***
	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Not Rated	-0.5611***	-0.5609***	-0.5611***	-0.5610***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
F-Statistic	187027	97087	67312	55883
R^2	0.7414	0.7416	0.7419	0.7420
Panel 2: Discrete Choice Estimates				
Percentile Rank	1.6848***	1.7032***	1.6696***	1.6604***
	(0.1111)	(0.1120)	(0.1121)	(0.1122)
Not Rated	0.9028***	0.9174***	0.8969***	0.8901***
	(0.0662)	(0.0668)	(0.0668)	(0.0669)
Distance	-0.1170***	0.0075	-0.0026	-0.0029
	(0.0005)	(0.0119)	(0.0121)	(0.0121)
Distance^2		-0.0007***	-0.0006***	-0.0006***
		(0.0002)	(0.0002)	(0.0002)
Willingness to Travel	5.28***	4.534***	4.436***	4.41***
	(0.349)	(0.298)	(0.297)	(0.298)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X

NOTES: The first panel presents the first stage results, where the dependent variable across all specifications is percentile rank. The second panel presents the discrete choice estimates. In both panels, all specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Service (NHS), and finds that patients are willing to travel 4% further to avoid a standard deviation increase in emergency room admissions. Similarly, Gutacker et al. (2016) finds that patients in the NHS are willing to travel 6% further for a standard deviation improvement in procedure-specific clinical quality but finds insignificant effects for other, more general quality measures. The magnitudes of these effects appear relatively small but difficult to compare with our context. An arguably more comparable study is Romley and Goldman (2011), which analyzes how revealed quality—an index of hospital features both known to and valued by patients—affects hospital choice for Medicare FFS pneumonia admissions. The study finds that patients are willing to travel between 2.41 and 3.94 additional miles for revealed quality at the 75th percentile as opposed to the 25th percentile. Given that the mean distance to the patient's chosen hospital is 2.8 miles, this indicates that patients are willing to approximately double their travel distance for higher revealed quality. Our estimates therefore appear reasonable relative to other studies from Romley and Goldman (2011) and Gutacker et al. (2016), as well as the labor and deliver results in 4.1, wherein we would expect a larger WTT estimate for orthopedic surgery relative to labor and delivery.

Our main results hold up to various alternative specifications, all of which are detailed in the supplemental material. We estimate the model where we replace distance with differential distance and find qualitatively similar results. We also estimate hospital choice at time t based on ratings in $t-2$ instead of $t-1$. These results are quite similar to the main results. Lastly, we assess the sensitivity of our estimates to the chosen market definition by estimating the main specification with other market definitions based on community detection algorithms, FIPS codes, and radii around the patient's home zip code. The estimates hold across market definitions of similar sample sizes but are attenuated in market definitions

that limit the data to a smaller subset of admissions and are larger when limiting the data to urban centers.

4.3 Falsification Analysis

To assess the credibility of the results in Sections 4.1 and 4.2, we estimate the same model, but among patients for whom online reviews should not affect their chosen hospital. We limit the data to patients admitted through the hospital's emergency department whose priority of admission was classified as "emergency," which is defined as patients that require "immediate medical intervention as a result of severe, life threatening or potentially disabling conditions." The data do not indicate which patients arrived via ambulance. Beyond the substantive change in the nature of admission, we limit the sample using the same criteria outlined in Section 2.[30]

We do not place any additional age or insurer restrictions on these patients. The data include Medicare, Medicare HMO, Medicaid, Medicaid HMO, and privately insured patients, where 19% have private insurance. This results in 7,321,225 emergency admissions across 139 hospitals. We use all of these admissions to construct hospital markets, but due to computational limitations, we use a random sample of 250,000 admissions to estimate the discrete choice model. Table 6 summarizes the entire sample. The second and third panel in Table 6 show that the hospital characteristics and their Yelp presence are similar to the labor and delivery and orthopedic surgery contexts.

We then use these data to define hospital markets. Analogous to the approach for labor

[30]Specifically, we drop admissions for patients that are discharged to court, law enforcement, or a psychiatric facility. We also limit to patients who have valid Florida zip codes and were admitted to hospitals whose information is contained in the AHA data.

Table 6: Summary Statistics: Emergency Admissions

	Mean	Median	St. Dev.	5th Pct.	95th Pct.
Patient Characteristics					
Age	62.12	66.00	22.07	20.00	90.00
Black	0.18	0.00	0.38	0.00	1.00
Latino	0.17	0.00	0.37	0.00	1.00
Asian	0.01	0.00	0.09	0.00	0.00
Male	0.45	0.00	0.50	0.00	1.00
Privately Insured	0.19	0.00	0.39	0.00	1.00
Hospital Characteristics and Quality Measures					
Total Beds	335.05	245.00	325.62	84.05	856.60
Physicians	20.22	6.00	47.92	0.00	99.95
Nurses	578.16	364.00	758.60	119.00	1850.35
Government	0.09	0.00	0.29	0.00	1.00
Non-profit	0.41	0.00	0.49	0.00	1.00
Major Teaching Hospital	0.05	0.00	0.23	0.00	1.00
Any Teaching Hospital	0.48	0.00	0.50	0.00	1.00
System Member	0.87	1.00	0.33	0.00	1.00
Payer Mix	0.57	0.57	0.11	0.39	0.75
Case Mix Index	1.56	1.55	0.20	1.25	1.89
Hospital Wide Readmission Rate	16.05	16.00	1.06	14.40	17.80
Yelp Reviews					
Observed Rating	2.84	3.00	0.98	1.00	4.50
New Reviews	1.29	1.00	1.66	0.00	5.00

Notes: Patient characteristics are from inpatient claims data for emergency department admissions and are measured at the quarter level. This table summarizes all emergency admissions, with a total of 7,321,225 observations. We use a random sample of 250,000 observations to estimate the model. The hospital characteristics are measured annually. The Yelp review data are measured at the quarterly level to correspond with the unit of time available in the inpatient claims data.

and delivery and orthopedic surgery, we start with the most broad market definition from the community detection algorithms and then layer in additional restrictions. This means we start with 14 markets, then limit the sample to markets with at least three rated hospitals, one of which must have at least three reviews. We also require that a patient's choice set is limited to hospitals within a 25-mile radius of her home zip code, because the 95th percentile for travel distance is just under 25 miles, and on average, these patients travel 8.5 miles. This leaves 118,599 admissions to 105 hospitals across 10 markets.

Using these admissions and the corresponding choice sets, we estimate the model starting with the first stage as outlined in Equation 2. The first panel of Table 7 presents the results. The instrument for being rounded into a higher rating is positively and significantly related to percentile rank across each specification. The covariates in each column are analogous to those in the first panels of Tables 3 and 5. Interactions with individual characteristics—which consist of age, sex, and indicators to capture if the patient is Black, Hispanic or Latino—are layered in as outlined.

Panel 2 of Table 7 presents the main results, where each column includes the residuals from the corresponding column in Panel 1. Across each column, the coefficient on percentile rank is small and insignificant. Following Sections 4.1 and 4.2, we present the WTT estimates in the bottom row of the table. These estimates are all small and insignificant, indicating that emergency patients are not willing to travel further for higher star ratings. This lends confidence to the positive and significant results found in the labor and delivery and orthopedic surgery contexts.

Table 7: Emergency Admissions Results

	(1)	(2)	(3)	(4)
Panel 1: First Stage Results				
Rounded Indicator	0.1363***	0.1362***	0.1366***	0.1363***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Not Rated	-0.5016***	-0.5012***	-0.5000***	-0.4993***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
F-Statistic	150523	80995	57694	45189
R^2	0.6762	0.6771	0.6809	0.6848
Panel 2: Discrete Choice Estimates				
Percentile Rank	-0.0007	0.0764	0.0347	-0.0087
	(0.0904)	(0.0922)	(0.0926)	(0.0929)
Not Rated	0.0662	0.0961*	0.0812	0.0480
	(0.0486)	(0.0495)	(0.0497)	(0.0498)
Distance	-0.2373***	-0.1684***	-0.1780***	-0.1771***
	(0.0009)	(0.0076)	(0.0078)	(0.0078)
$Distance^2$		0.0000	0.0001	0.0001
		(0.0003)	(0.0004)	(0.0004)
Willingness to Travel	-0.001	0.092	0.042	-0.010
	(0.133)	(0.112)	(0.112)	(0.112)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X

NOTES: The first panel presents the first stage results, where the dependent variable across all specification is percentile rank. The second panel presents the discrete choice estimates. In both panels, all specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

5 Implications for Pricing

Our analysis provides compelling evidence that online reviews affect hospital choice for patients that have some ability to choose. Given this effect on demand, there is reason to suspect that online reviews may also have subsequent effects on other dimensions of hospital markets—namely, hospital prices. Simply put, in a bilateral negotiation between hospitals and insurers, an increase in demand will improve a hospital's bargaining position, enabling them to negotiate higher prices. To analyze this relationship more formally, we revisit the bargaining model presented in Ho and Lee (2017) (henceforth, "HL"). As shown in HL and following the notation used in McCarthy and Huang (2018), we define the negotiated price between hospital i and insurer j as

$$p_{ij} = \arg\max_{p_{ij}} \left(\Delta\pi_{ij}^{H}\right)^{b_{ij}} \times \left(\Delta\pi_{ij}^{M}\right)^{1-b_{ij}}. \tag{6}$$

Here $\Delta\pi_{ij}^{H}$ represents the difference between hospital i's profits from reaching an agreement with insurer j or not. Analogously, $\Delta\pi_{ij}^{M}$ represents insurer j's change in profits from reaching an agreement with hospital i. In the case where the two parties do not come to an agreement, the model assumes that hospital i is excluded from insurer j's network. Additionally, b_{ij} represents hospital i's bargaining power in negotiations over prices with insurer j. The profit functions for hospital i and insurer j are

$$\pi_{i}^{H}(\boldsymbol{p}, \boldsymbol{\theta}) = \sum_{n} D_{in}^{H}(p_{in} - c_i), \text{ and} \tag{7}$$

$$\pi_{j}^{M}(\boldsymbol{p}, \boldsymbol{\theta}) = D_{j}^{M}(\theta_j - \eta_j) - \sum_{h} D_{hj}^{H} p_{hj}. \tag{8}$$

Note that D_{in}^H represents the demand for hospital i across patients enrolled with insurer n, and c_i is the average cost per admission. Further, D_j^M denotes the demand for insurer j, θ_j is the insurer's premium, and η_j is non-inpatient hospital costs. As derived in HL, the resulting negotiated price between hospital i and insurer j is:

$$p_{ij}^* D_{ij}^H = b_{ij}\left[\Delta D_j^M(\theta_j - \eta_j) - \sum_{h \neq i} p_{hj}^* \Delta D_{hj}^H\right] + (1 - b_{ij})\left[c_i D_{ij}^H - \sum_{n \neq j} \Delta D_{in}^H (p_{in}^* - c_i)\right]. \quad (9)$$

The first term on the right side of Equation 9 denotes the change in net revenues to insurer j due to potential loss in enrollment minus the change in payments to the hospitals in insurer j's network, excluding hospital i. Within the brackets, $\Delta D_j^M(\theta_j - \eta_j)$ represents the effect of hospital i's inclusion in insurer j's network on premium revenue. The second term on the right-hand side captures hospital i's costs less its change in profits from other insurers. Within the brackets, the term $c_i D_{ij}^H$ is the hospital cost effect, and $\sum_{n \neq j} \Delta D_{in}^H (p_{in}^* - c_i)$ represents the recapture effect, i.e. the changes in hospital i's reimbursements from other insurers when hospital i is not included in insurer j's network. This captures i's outside option: what would hospital i be paid by other insurers if not included in insurer j's network?

As evident from Equation (9), if higher quality ratings increase the probability of selecting that hospital, then the insurer's willingness to pay to keep that hospital in their network also increases. Given our findings in Section 4, we predict that the effect of ratings on hospital choice should also create upward pricing pressure in a hospital's negotiation with private insurers.

It is worth noting that the term of a negotiated payment contract between a hospital and insurer is typically more than one year, although there is considerable variation in contract

length across insurers and hospitals. In any given year, some subset of contracts will be up for renegotiation while others will remain under an existing contract. Since our analysis considers an overall average hospital price, there exists yearly variation in our pricing measure due to this churn of contracts over time. Even among contracts that are not renegotiated in a given year, there is an opportunity for variation in observed prices if those contracts are based on a "percentage of charges."[31] While we do not have access to the schedules or terms of any contracts, the local average treatment effect from our estimation strategy will tend to consist of relatively larger hospitals with more reviews. As examined in Cooper et al. (2019), such hospitals are also more likely to negotiate prices as a percentage of charges. These details highlight the opportunity for meaningful variation in observed hospital-level prices from year to year, even as the term of each individual contract extends beyond a single year.

5.1 Incorporating Hospital Pricing Data

We combine the hospital Yelp data with annual cost report data from the Healthcare Cost Report Information System (HCRIS) to analyze the effect of reviews at time t on prices in $t+1$.[32] The annual reporting of the HCRIS data requires us to conduct this analysis at the hospital-year level. We therefore use year-end star ratings to construct our explanatory variables of interest, meaning that, in this context, t indicates the end of the given year. Recall that hospitals can have an aggregate rating between 1 and 5 stars in half-star increments. This translates to a total of nine rating groups. We also include hospitals that are

[31] See Cooper et al. (2019) for additional discussion regarding hospital/insurer contract types.
[32] We focus on the price in the following year to allow for a lagged response to rating information.

not rated in our analysis, resulting in ten groups. Due to the sparsity of observations within these narrow rating categories, we have insufficient power to estimate effects at all possible ratings. As such, we aggregate the possible ratings into four groups: low, middle, high, and unrated. The high rated group consists of hospitals with 4, 4.5, or 5 stars, those in the middle rated group have 3 or 3.5 stars, and hospitals in the low rated group have 2.5 stars or below. We also form an indicator for hospitals without a Yelp presence. Our delineation of the star rating groups follows from the natural cut-points observed in the data. Given the observed average rating of 2.9, a middle rated hospital falls in a group slightly above the average (i.e. 3 or 3.5 star rating), and a high rated hospital is at least a standard deviation above the average (i.e. 4, 4.5, or 5 stars).[33]

To construct our price variable, we create a measure of price for all non-Medicare inpatient discharges by taking the sum of inpatient charges reduced by the discount factor less Medicare payments, divided by the number of non-Medicare inpatient discharges (Dafny, 2009; Schmitt, 2018; Lin et al., 2020). We eliminate observations with price outliers at the 5th and 95th percentiles, and we deflate all values to 2012 dollars. Our final price measure reflects an average hospital-level negotiated payment between hospitals and commercial insurers from 2012 through 2017.[34] If we limit our study of hospital prices to hospitals in Florida—i.e., just to the hospitals that appear in our choice analysis—we are left with a 5% subsample of the observations available in the HCRIS data. We, therefore, take advantage

[33]The supplemental material provides further discussion of our rating group choices along with estimation results of our main specification modified to include more granular rating groups. We also provide results that use different thresholds for middle and high. Using these thresholds, the results are less precise but have commensurate point estimates.

[34]We are able to back out Medicare payments, but the data do not enable us to remove Medicaid payments. Specifics on the variables used to construct the price measure are detailed in the supplemental material. We also provide additional analysis regarding the robustness of our results to the presence of outliers in the supplemental material. We find that our qualitative results are not sensitive to the presence of price outliers.

of the full sample and estimate the price effect at the national level.[35]

Our outcome of interest is log price, i.e. $\ln(price_{t+1})$, which we simply refer to as "price."[36] The measure of time is calendar year, but the cost reports are compiled by hospital fiscal year, which means that for hospitals whose fiscal year differs from the calendar year, the price change is capturing less than a full year of "exposure" to that rating.[37] We examine the sensitivity of our results to these timing considerations in the supplemental material, in addition to falsification tests when considering prices at earlier time periods. The results of that analysis are consistent with our initial findings.

In addition to the data from Yelp and HCRIS, we incorporate county-level characteristics from the Area Health Resource Files, hospital quality information from CMS's Hospital Compare data, and hospital characteristics from the AHA Annual Survey data.[38] We limit our dataset to general acute care hospitals that have at least 30 beds for which we can construct a valid price estimate. Our final sample contains 15,854 hospital-year level observations, which are summarized in the supplemental material. On average, our hospital-year observations have 18 reviews, conditional on having a Yelp rating, with an average rating of 2.9. The hospitals in our sample are more likely to be private, non-profit, and members of a system, but are representative of an average, mid-to-large acute care hospital in the United States.

[35] When we conduct the analysis with only hospitals from Florida, we generally find qualitatively similar estimates to those presented in this section, but given that this restriction drops 95% of our observations, we find dramatically less precise estimates.

[36] We opt for log prices due to the wide range in the outcome variable. We find qualitatively similar results when estimating this equation in levels rather than logs.

[37] For example, consider a hospital whose fiscal year ends on June 30. The year-end rating for that hospital captures the rating as of December in year t, while the price measure for this hospital approximates the average commercial payment from July 1 in year t through June 30 in year $t+1$.

[38] The AHRF variables include population, unemployment and poverty rates, rate of uninsured, and median income. Quality measures include readmission and mortality rates for heart failure, pneumonia, and acute myocardial infarction.

5.2 Empirical Approach

Recall that our empirical analysis of hospital choice is procedure-specific and estimated using Florida claims data. Conversely, our estimated prices are at the hospital-level and are available for all hospitals submitting cost report information in the US. Our pricing data is therefore geographically broader than our choice data but limited with respect to procedure-specific information. As such, our data do not allow for the estimation of both demand and price effects collectively as part of a single bargaining framework. Instead, we estimate the effects of ratings on prices separately from our analysis of hospital choice. To do so, we again exploit the exogenous variation on the Yelp platform—wherein a continuous underlying score is rounded to the nearest half-star increment—as an instrument for the observed rating category in a two-stage least squares estimator.[39] We employ the following regression specification, where Equation 5a is the second stage and Equations 5b and 5c are the first stage:

$$\ln(Price_{i,t+1}) = \beta_1 \widehat{High_{it}} + \beta_2 \widehat{Mid_{it}} + \delta_1 None_{it} + \delta_2 TooFew_{it} + X_{it}\boldsymbol{\alpha} + \theta_{i,c,t} + \varepsilon_{it}, \quad (5a)$$

$$High_{it} = \lambda_1 RH_{it} + \lambda_2 RM_{it} + \zeta_1 None_{it} + \zeta_2 TooFew_{it} + X_{it}\boldsymbol{\gamma} + \theta_{i,c,t} + \sigma_{it}, \quad (5b)$$

$$Mid_{it} = \tau_1 RH_{it} + \tau_2 RM_{it} + \eta_1 None_{it} + \eta_2 TooFew_{it} + X_{it}\boldsymbol{\rho} + \theta_{i,c,t} + \mu_{it}. \quad (5c)$$

[39] While a regression discontinuity (RD) design is seemingly a natural starting point, several features of our application deem it inappropriate. For example, we observe relatively small sample sizes within different bandwidths, movement in and out of treatment over time, and a single running variable with multiple treatments. The sparsity of reviews near the rounding thresholds is particularly problematic as it suggests a violation of the continuity assumption. This contrasts with studies that use Yelp reviews in other settings, such as restaurants, where there are more businesses, more reviews, and the outcome of interest can be measured more frequently (Anderson and Magruder, 2012).

The variable $High_{it}$ is an indicator equal to 1 if the hospital has a year-end rating of 4 or above, and Mid_{it} is an indicator for hospitals with year-end ratings equal to 3 or 3.5.[40] We also include an indicator for hospitals without a Yelp presence ($None_{it}$), along with an indicator for hospitals with fewer than 3 ratings ($TooFew_{it}$). We define 3 ratings as the cutoff because a hospital must have at least 3 reviews to have the possibility of being rounded. Hospitals with fewer than three ratings do not have their aggregate rating included in the $High_{it}$ or Mid_{it} variables.[41] Additionally, X_{it} is a vector of hospital and county characteristics, and $\theta_{i,c,t}$ represents separate fixed effects for year, county, and hospital.[42] In the first stage Equations 5b and 5c, RH_{it} and RM_{it} are indicators for being rounded up into a high or a middle rating, respectively.

For a given rating category, hospitals located near the rounding threshold have similar underlying scores but different summary scores. Thus, we use an indicator for being rounded into a higher star rating as an instrument for the endogenous hospital rating. We impose a bandwidth of 0.15 around the 2.75 and 3.75 thresholds for the middle and high groups, respectively. This means that for a hospital's rating to be considered "rounded up" into high, the average rating would need to fall between 3.75 and 3.90. Similarly, hospitals rounded up into middle have an average rating between 2.75 and 2.90. Defining the instrument as such, approximately 10% of reviewed hospitals are rounded in each year.[43] Given the fact that a

[40] We do not include the continuous score in our primary specification, but we do include it in the supplemental material, where we add the continuous score with no change in our results.

[41] The appendix presents results where the minimum number of reviews ranges from 4 to 10. The results coincide with that of our main specification, noting that we find increasingly high magnitudes on the high category when the number of required reviews increases.

[42] Note that hospital mergers, acquisitions, and closures create a distinction between county and hospital fixed effects in this analysis. As such, we include both.

[43] The standard deviation of this value over the sample period is 0.012. The most rounding took place in 2013 with 11.5% of hospitals, and the least in 2017 with 8.25%.

patron on the site would not have information on whether or not a hospital was rounded up, it is reasonable to assume that the instrument only affects price through its effect on ratings, and thus plausibly satisfies the exclusion restriction.[44]

5.3 Results

Here we focus on our main results and reserve a detailed presentation of the first stage and reduced form analyses for the supplemental material. To summarize, the first stage results show a strong positive relationship between the rounding indicator, i.e., the instrument, and the respective rating group. The reduced form analysis shows a positive relationship between each instrument and price, which is significant in the majority of specifications.

Our primary IV results are presented in Table 8. The "High Rating" and "Middle Rating" coefficients reflect the estimated percentage point increase in price for a hospital in the high or middle group in comparison to the low rating group. We also show the coefficients for the no reviews and "too few reviews" variables. Our dependent variable across each of the four specifications is hospital price.

Across all specifications we find a price premium for hospitals that do not have a low rating. Recall that in our analysis, identification comes from those hospitals that were rounded up into a higher group. Thus we are identifying the local average treatment effect on the hospitals that move into a higher rating category (Mid_{it} or $High_{it}$) due to the rounding mechanism. Even in our most saturated specification (column 4), we see that relative to

[44]The supplemental material provides additional analyses that allay concerns that our results are particularly sensitive to the choices made in structuring our estimation. For example, we provide alternative results with a narrower bandwidth, different definitions of high and low rating groups, and more granular rating groups. We also examine the sensitivity of our results to potential violations of the exclusion restriction. Across these specifications we find little qualitative change in our results.

Table 8: Ratings and Hospital Prices

Price	(1)	(2)	(3)	(4)
High Rating	0.0705*	0.0819**	0.0550*	0.0710**
	(0.0395)	(0.0372)	(0.0316)	(0.0347)
Middle Rating	0.104***	0.0722**	0.0683***	0.0501
	(0.0402)	(0.0310)	(0.0256)	(0.0310)
No Reviews	-0.0139	0.000542	0.0385**	0.0406*
	(0.0248)	(0.0206)	(0.0196)	(0.0221)
Fewer than 3 Reviews	0.00851	0.00675	0.0289*	0.0265
	(0.0249)	(0.0197)	(0.0169)	(0.0197)
County Fixed Effects	No	Yes	Yes	Yes
Hospital Fixed Effects	No	No	Yes	Yes
Hospital Quality Measures	No	No	No	Yes
Observations	11850	11780	11693	8061
Kleinbergen-Paap LM Statistic	196.4	189.5	127.9	85.49
Kleinbergen-Paap F Statistic	507.0	350.1	144.7	97.91

NOTES: All specifications include a set of hospital and county level characteristics, along with year fixed effects. Additional fixed effects and controls are indicated in the respective columns. Robust standard errors clustered at the hospital level are in parentheses. The Kleinbergen-Paap L M and F statistics allow for non-i.i.d. errors. Stars indicate the following: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

low-rated hospitals, the average price for an inpatient stay is higher the following year at a high-rated hospital. Our point estimates, when rescaled based on the observed ratings among high versus low-rated hospitals, translate to around a 1.5% increase in prices for a half-star increase in ratings.

We also estimate an increase in price for middle-rated hospitals, although the point estimate for this effect is imprecise and not significant at conventional levels (p-value=0.106). We conclude from this analysis that, for sufficiently high ratings, hospitals are able to negotiate price increases in subsequent periods. To be clear, these price increases capture the effect of a higher reported rating, and not an increase in underlying quality, as our identification strategy isolates the impact of a change in *reported* quality on price.

Turning to the coefficient on the "no reviews" indicator, the results suggest a positive relationship between a hospital's lack of online reviews and price. These estimates indicate that on this platform, no information is better than negative information. This is consistent with other findings in the literature, which show that consumers may use ratings to avoid low quality in addition to actually seeking higher quality (Cabral and Hortacsu, 2010; Burkle and Keegan, 2015; Lu and Rui, 2018; Lantzy and Anderson, 2020).

These results hold up to a bevy of alternative specifications and robustness checks, which are detailed in the supplemental material. In short, we find qualitatively similar results even when we modify the minimum number of reviews required to be considered rated, use different rating groups and bandwidths, impose a balanced panel, and modify the covariates included. Our results are also commensurate with the existing literature. Even at the top end of our point estimates, the price effects from higher ratings are modest relative to other studies of hospital pricing. For example, Lin et al. (2020) finds a 3 – 5 percent increase

in hospital prices after vertical integration. Lewis and Pflum (2017) finds that hospitals acquired by out-of-market systems increase prices by around 17 percent, and that the prices of close competitors increase by approximately 8 percent. Other recent research on hospital mergers finds much larger price effects. For example, Dafny (2009) finds a one-time increase in price of 40 percent following the merger of nearby rivals, which is commensurate with the results found in various structural analyses of mergers (Capps et al., 2003; Gaynor and Vogt, 2003).

Reviewed in the context of the existing hospital pricing literature, our results are reasonable. We would *ex ante* anticipate a smaller price effect from online reviews as compared to, e.g., hospital mergers. Our central takeaway from the results, however, is that higher online reviews do appear to translate into higher hospital prices. Even at relatively modest magnitudes, this finding is important to help guide and understand potential effects of future transparency efforts from CMS and other agencies.

6 Discussion

This paper provides novel insights on the effects of online reviews on hospital choice. We find significant increases in willingness to travel for higher ratings for both labor and delivery and orthopedic surgery admissions. The magnitudes of these effects are commensurate with the existing literature and reflect the nature of the respective procedure. The results lend further support to other studies that show that aggregate measures of quality and measures motivated by the patient perspective of care drive hospital choice (Dranove and Sfekas, 2008; Romley and Goldman, 2011; Pope, 2009; Chandra et al., 2016).

These findings have important implications for policy efforts interested in improving information disclosure in health care markets. The results indicate that Yelp reviews may provide a more accessible way of understanding quality of care, provide new information, or some combination of these factors. Understanding the structure and substance of metrics that are relevant to health care decisions can help guide quality disclosure efforts and motivates research on other platforms, which will likely also be relevant to these decisions.

Consistent with a standard Nash bargaining framework, our analysis also suggests that higher online reviews translate into higher hospital prices. The magnitude of this increase is relatively small compared to changes in market structure (e.g., mergers), but economically meaningful nonetheless. One policy implication of this analysis is that efforts to make hospital quality information more accessible may have the unintended consequence of facilitating price increases even for hospitals of similar underlying quality.

Our findings also highlight important outstanding questions surrounding the incentive structures in hospital markets. Given that online reviews affect choice, and therefore increase demand, hospitals face incentives to invest in dimensions of care that will bolster their ratings and subsequently increase their market shares. Existing research shows that non-clinical features such as bedside manner and amenities drive these reviews; therefore, whether or not prioritizing these features of care is beneficial depends on the extent to which investments in these dimensions improve the efficiency of health care delivery. By further exploring this relationship, future research can advance our understanding of how various features of quality and its disclosure affect hospital markets which is pivotal to designing policy that fosters efficiency in health care.

References

Alker J, Hoadley J. 2013. Medicaid managed care in florida: Federal waiver approval and implementation. *Georgetown Health Policy Institute* .

American Hospital Association. 2016. Health organizations request delayed release of hospital star ratings on hospital compare website. https://www.aha.org/letter/2016-04-05-health-organizations-request-delayed-release-hospital-star-ratings-hospita Online; accessed 22 June 2020.

American Hospital Association. 2017. RE: Enhancements of the Overall Hospital Quality Star Rating, August 2017. https://www.aha.org/letter/2017-09-25-aha-cms-re-enhancements-overall-hospital-quality-star-rating-august-2017. Online; accessed 22 June 2020.

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

Austin PC. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine* **28**: 3083–3107.

Austin PC. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* **46**: 399–424.

Avdic D, Moscelli G, Pilny A, Sriubaite I. 2019. Subjective and objective quality and choice of hospital: Evidence from maternal care services in Germany. *Journal of Health Economics* **68**: 2–18. ISSN 0167-6296.
URL https://doi.org/10.1016/j.jhealeco.2019.102229

Bardach NS, Asteria-Peñaloza R, Boscardin WJ, Dudley RA. 2013. The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ Quality and Safety* **22**: 194–202.

Brot-Goldberg ZC, Chandra A, Handel BR, Kolstad JT. 2017. What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics. *The Quarterly Journal of Economics* **132**: 1261–1318.

Bundorf MK, Chun N, Goda GS, Kessler DP. 2009. Do markets respond to quality information? the case of fertility clinics. *Journal of health economics* **28**: 718–727.

Burkle CM, Keegan MT. 2015. Popularity of internet physician rating sites and their apparent influence on patients' choices of physicians. *BMC health services research* **15**: 1–7.

Cabral L, Hortacsu A. 2010. The dynamics of seller reputation: Evidence from ebay. *The Journal of Industrial Economics* **58**: 54–78.

Campbell L, Li Y. 2018. Are Facebook user ratings associated with hospital cost, quality and patient satisfaction? A cross-sectional analysis of hospitals in New York State. *BMJ Quality and Safety* **27**: 119–129. ISSN 20445415.

Capps C, Dranove D, Satterthwaite M. 2003. Competition and market power in option demand markets. *RAND Journal of Economics* : 737–763.

Cattaneo MD, Jansson M, Ma X. 2018. Manipulation testing based on density discontinuity. *The Stata Journal* **18**: 234–261.

Chandra A, Finkelstein A, Sacarny A, Syverson C. 2016. Health care exceptionalism? Performance and allocation in the US health care sector. *American Economic Review* **106**: 2110–2144. ISSN 00028282.

Chernew M, Cooper Z, Larsen-Hallock E, Morton FS. 2018. Are health care services shoppable? evidence from the consumption of lower-limb mri scans. Technical report, National Bureau of Economic Research.

Conley TG, Hansen CB, Rossi PE. 2012. Plausibly exogenous. *Review of Economics and Statistics* **94**: 260–272.

Cooper Z, Craig SV, Gaynor M, Van Reenen J. 2019. The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured. *The Quarterly Journal of Economics* **134**: 51–107.

Cutler DM. 2011. Where are the health care entrepreneurs? the failure of organizational innovation in health care. *Innovation Policy and the Economy* **11**: 1–28.

Cutler DM, Huckman RS, Landrum MB. 2004. The role of information in medical markets: an analysis of publicly reported outcomes in cardiac surgery. *American Economic Review* **94**: 342–346.

Dafny L. 2009. Estimation and Identification of Merger Effects: An Application to Hospital Mergers. *The Journal of Law and Economics* **52**: 523–550. ISSN 0022-2186.

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

Dellarocas C. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science* **49**: 1407–1424.

Desai S, Hatfield LA, Hicks AL, Sinaiko AD, Chernew ME, Cowling D, Gautam S, Wu Sj, Mehrotra A. 2017. Offering a price transparency tool did not reduce overall spending among california public employees and retirees. *Health affairs* **36**: 1401–1407.

Dranove D, Jin GZ. 2010. Quality disclosure and dertification: Theory and practice. *Journal of Economic Literature* **48**: 935–963.

Dranove D, Sfekas A. 2008. Start spreading the news: a structural estimate of the effects of new york hospital report cards. *Journal of health economics* **27**: 1201–1207.

Everson J, Hollingsworth JM, Adler-Milstein J. 2019. Comparing methods of grouping hospitals. *Health services research* **54**: 1090–1098.

Freue GVC, Ortiz-Molina H, Zamar RH. 2013. A natural robustification of the ordinary instrumental variables estimator. *Biometrics* **69**: 641–650.

Garthwaite C, Ody C, Starc A. 2020. Endogenous quality investments in the us hospital market. Technical report, National Bureau of Economic Research.

Gaynor M, Vogt WB. 2003. Competition among hospitals. *RAND Journal of Economics* **34**: 764–785.

Gutacker N, Siciliani L, Moscelli G, Gravelle H. 2016. Choice of hospital: Which type of quality matters? *Journal of health economics* **50**: 230–246.

Ho K, Lee RS. 2017. Insurer Competition in Health Care Markets. *Econometrica* **85**: 379–417.

Howard P, Feyman Y. 2017. Yelp for Health: Using the Wisdom of Crowds To Find High-Quality Hospitals. Technical Report April, Manhattan Institute.

Jin G, Sorensen AT. 2006. Information and consumer choice : The value of publicized health plan ratings. *Journal of Health Economics* **25**: 248–275.

Kuklina EV, Whiteman MK, Hillis SD, Jamieson DJ, Meikle SF, Posner SF, Marchbanks PA. 2008. An enhanced method for identifying obstetric deliveries: implications for estimating maternal morbidity. *Maternal and child health journal* **12**: 469–477.

Lantzy S, Anderson D. 2020. Can consumers use online reviews to avoid unsuitable doctors? evidence from ratemds. com and the federation of state medical boards. *Decision Sciences* **51**: 962–984.

Lewis MS, Pflum KE. 2017. Hospital systems and bargaining power: evidence from out-of-market acquisitions. *RAND Journal of Economics* **48**: 579–610.

Lin H, McCarthy IM, Richards M. 2020. Hospital Pricing following Integration with Physician Practices.

Lu SF, Rui H. 2018. Can We Trust Online Physician Ratings? Evidence from Cardiac Surgeons in Florida. *Management Science* **64**: 2557–2573.

McCarthy I, Huang SS. 2018. Vertical Alignment Between Hospitals and Physicians as a Bargaining Response to Commercial Insurance Markets. *Review of Industrial Organization* **53**: 7–29.

McCarthy I, Sanbower K, Aragón LS. 2020. Online reviews and hospital prices. Working paper.

Mehrotra A, Dean KM, Sinaiko AD, Sood N. 2017. Americans support price shopping for health care, but few actually seek out price information. *Health Affairs* **36**: 1392–1400.

Moscelli G, Siciliani L, Gutacker N, Gravelle H. 2016. Location, quality and choice of hospital: Evidence from england 2002–2013. *Regional Science and Urban Economics* **60**: 112–124.

Moscone F, Tosetti E, Vittadini G. 2012. Social interaction in patients' hospital choice: evidence from italy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **175**: 453–472.

Perez V, Freedman S. 2018. Do Crowdsourced Hospital Ratings Coincide with Hospital Compare Measures of Clinical and Nonclinical Quality? *Health Services Research* **56**: 4491–4506. ISSN 14756773.

Petrin A, Train K. 2010. A control function approach to endogeneity in consumer choice models. *Journal of marketing research* **47**: 3–13.

Pope DG. 2009. Reacting to rankings: evidence from "america's best hospitals". *Journal of health economics* **28**: 1154–1165.

Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, Asch DA, Ungar LH, Merchant RM. 2016. What can Yelp teach us about measuring hospital quality? *Health Affairs* **35**: 697–705. ISSN 1544-5208.

Romley JA, Goldman DP. 2011. How costly is hospital quality? A revealed-preference approach. *Journal of Industrial Economics* **59**: 578–608. ISSN 00221821.

Scanlon DP, Chernew M, McLaughlin C, Solon G. 2002. The impact of health plan report cards on managed care enrollment. *Journal of health economics* **21**: 19–41.

Schmitt M. 2018. Multimarket Contact in the Hospital Industry. *American Economic Journal: Economic Policy* **10**: 361–387.

Stuart EA. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**: 1.

Varkevisser M, van der Geest SA, Schut FT. 2012. Do patients choose hospitals with high quality ratings? empirical evidence from the market for angioplasty in the netherlands. *Journal of health economics* **31**: 371–378.

Zhang Z, Kim HJ, Lonjon G, Zhu Y, et al. 2019. Balance diagnostics after propensity score matching. *Annals of translational medicine* **7**.

Appendix A Yelp Data

We use the AHA Annual Survey database to match 2,935 hospitals to Yelp profiles with reviews.[45] For these hospitals, the name associated with the profile exactly matched the name listed in the AHA data. We then implemented the following process to ensure that the profiles are associated with the correct hospital. We refer to these profile-hospital matches as "exact" matches.

Appendix A.1 Data Cleaning

(a) We eliminated any observations without an address on the Yelp profile because we need to match the address in the Yelp profile to the AHA data to ensure the profile is describing the proper hospital. This leaves a total of 2,904 observations.

(b) We then reformatted the Yelp addresses to match the conventions used in the AHA data. For instance, 'E' for 'East', 'St' for 'Street', or '1st' for 'First', which is what AHA uses for street, direction, and number abbreviations.

(c) Next we parsed out the first part of the string from both addresses. Most often this is the street number, but it can also be a word (i.e., the street name) if the number is missing. We implement this process again for the second and third words of the addresses. This creates six new variables, namely the first, second, and third word or number for both addresses.

We use these new variables to compare the addresses and keep those that match. At each of the following steps, we reviewed the observations selected to ensure that the addresses are matched as intended.

(d) We kept 1,687 observations with exact address matches. This left 1,217 observations to be analyzed.

(e) We then kept 333 observations where the first three words of the addresses match, which handles cases where the address is the same but one has an additional directional term at the end of the address (i.e., SE for 'South East').

(f) We added 240 more observations where the first word matched along with a match between some combination of the second and third words. This allows us to keep observations where the street number matches, but one address states 'South Main Street' and the other is 'Main Street', for example.

(g) We manually reviewed 258 observations with only matching street numbers or those with an '&' or 'and' in the name. We inspected these manually because when the street number matches but the street does not, sometimes the hospital has its own street name that is connected to a larger street or highway. Additionally, the '&' or 'and' typically signifies a cross-street, where both addresses are likely describing the same hospital.

[45] The data used in this paper were also used in McCarthy et al. (2020).

(h) Lastly, we manually reviewed observations with different street numbers but the same subsequent address information to ensure that we did not include doctor's offices located in the same complex as the hospital.

(i) We dropped any observations where none of the first three words or numbers of the addresses matched. Any remaining observations were manually reviewed.

Appendix A.2 Manual Review of Observations

To manually review the remaining observations, we first inspected the addresses to see if anything slipped through the sorting process. This includes observations where, for example, the AHA address was 'Ridgeview Road' and on Yelp it was 'Ridge View Road.' In cases with the same street number but different street name, we used Google Maps to determine whether or not they were referring to the same location. We examined any remaining profiles using this process. Any addresses that did not refer to the same location were dropped.

Appendix A.3 Approximate Matches

Note that the process above referred to hospitals that were exact matches, meaning that the name in the AHA data and the name on Yelp matched precisely. However, the data also include hospital profiles that had approximate matches to the AHA data. An approximate match is a hospital name that matches the Yelp profile with the exception of one word. We used the process outlined above on these data, but, there were very few relevant observations. Many of them referred instead to veterinary hospitals, hospital cafeterias, and physician practices. The analysis, therefore, does not use the approximate match data, but we mention it here to provide additional clarity on the data collection process.

Appendix A.4 Evidence of Decision Makers Using Online Reviews

For online reviews to affect hospital choice, health care decision makers must actually use this information. This is an important underlying assumption in this analysis. While we cannot ask the patients directly if they used online reviews, we can analyze the text of the review comments to determine if reviewers mention using reviews to inform their decisions. We do this by first identifying all of the reviews that have the words "read" and/or "see" *and* any of the following words: review, rating, star, yelp, google.[46] We find that nine percent of reviews meet this criteria. This search finds reviews with comments like "Reading some of these reviews I was a little worried but I had an excellent experience." However, it also identifies reviews such as "If I could give this place no stars I would. This is the worse place I have ever been to. ... I have never seen anything like this in my entire life." This shows that the criteria here are relatively loose and may not limit the reviews to the sample of interest. We therefore apply a stricter set of criteria which requires the review to have a bigram (i.e. set of two words) from the following list: "read review", "read yelp", "read google", "see rating", "see review", "see yelp", "see google." Using this approach, We find

[46] I first preprocess the review text to impose all lower-case text.

that one percent of the reviews meet this criteria. We read a sample of 50 of these reviews and found that each explicitly mentions consulting online reviews.

Based on these findings, we argue that between one and ten percent of persons on the Yelp platform considered online reviews in selecting a hospital, but this is not to say that only 10% of potential patients use this information. These criteria miss comments such as: "Hope this helps, I know I felt I couldn't find a lot of reviews about it when I was looking," which indicates that this person consulted the reviews, but the verbiage slips through the search criteria. It is also possible that a patient consults online reviews prior to her hospital visit and then either does not mention it in her review or does not review the hospital at all. We cannot measure the extent to which that occurs, but this exercise lends confidence to the idea that online reviews are relevant to the decision making process.

Appendix B Florida Data

The Florida inpatient claims data contains the population of Florida inpatient stays over the sample period, i.e. 2012 through 2017. I limit the data to the respective procedure using the following processes.

Appendix B.1 Labor and Delivery

To identify labor admissions in each quarter, we first limit the data to all claims that include a diagnosis code for "outcome of delivery" (Kuklina et al., 2008). In the ICD-9 diagnosis codes this is `V27._`, where the digit in place of the underscore identifies the number of babies and whether or not they were live or stillborn. The analogous ICD-10 code is `Z37_`. This keeps all admissions that include an outcome of delivery code. Then, we use procedure codes to limit the data to admissions with a normal delivery or cesarean section. For the ICD-9 codes, these admissions consist of procedure codes `73.59, 74.0` and `74.1`, and for ICD-10, the codes are `10E0XZZ, 10D00Z0`, and `10D00Z1`.

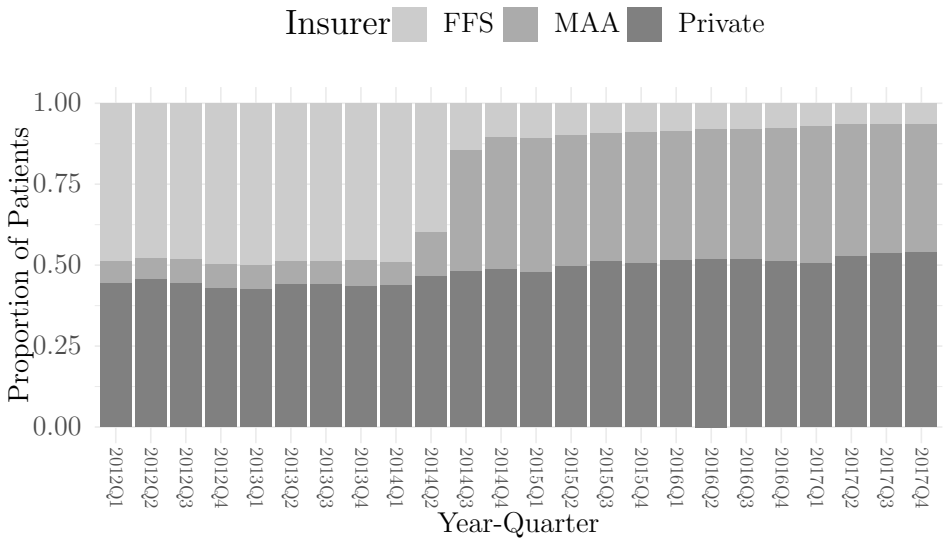
Among labor and delivery patients, we limit the data to admissions for normal delivery or cesarean section paid for by Medicaid, Medicaid HMO, or private insurance. We include each of these payer types because, for one, the data do not indicate that any hospitals only accept private insurance for labor and delivery. This suggests that hospitals are likely not turning away Medicaid patients for this type of admission. Additionally, in 2014, Florida launched its new Medicaid program (titled Managed Medical Assistance, i.e. MMA) where it moved the majority of its Medicaid beneficiaries to managed-care plans, as illustrated in Figure A2 (Alker and Hoadley, 2013). However, as Figure A2 as shows, the data still contain Medicaid fee-for-service (FFS) births in 2014 and after. This is because the FFS population comprises Medicaid recipients who are not included in MMA–either because they are not required to enroll or because they are members of an "Excluded" population. In the context of labor and delivery, recipients may be excluded from MMA because they are only eligible for family planning services or they are eligible for the Medically Needy program.[47] These institutional details indicate that while we should expect MMA to be the insurer for the majority of publicly funded births starting in 2014, there can still be births covered by Medicaid FFS after Florida overhauled its Medicaid program.

Appendix B.2 Orthopedic Surgery

The process to identify orthopedic surgery admissions is simpler. We limit the data to those observations for hip and knee replacement using the following diagnosis related group (DRG) codes: `469, 470, 461, 462, 466, 467, 468`. We limit these data to Medicare fee-for-service patients. Summary statistics and additional information on these data are detailed in the main text.

[47] Additional information on "Excluded," "Voluntary," and "Mandatory" populations is outlined here: https://ahca.myflorida.com/Medicaid/statewide_mc/pdf/mma/SMMC_MMA_Snapshot.pdf and https://www.medicaid.gov/sites/default/files/2019-12/fl-amrp-16.pdf

Figure A2: Composition of Insurance Coverage by Quarter



NOTES: The plot shows the proportion of patients covered by Medicaid Fee-for-Service (FFS), Medicaid Managed Medical Assistance (MMA), and private insurance by quarter. The data cover labor and delivery admissions for 2012 through 2017. Due to a change in the structure of Florida's Medicaid program in 2014, many patients who would otherwise be FFS patients shifted to MMA.

Appendix C Community Detection for Hospital Markets

Community detection (CD) relies on an adjacency matrix that indicates which zip codes go to common hospitals. This process begins by first creating a bipartite matrix relating zip codes and hospitals. The matrix is comprised of zeros and ones, where one indicates that people from that zip code went to the corresponding hospital. Without further restrictions, this means that even if a hospital only serves a small number of patients from a given zip code, that hospital and zip code would be connected. Instead, we impose a minimum share value of 0.15, meaning that at least 15% of a hospital's labor and delivery admissions must come from that zip code in order to be considered connected. This bipartite matrix is the basis for the unipartite adjacency matrix needed for CD. By multiplying the bipartite matrix by its transpose, We create the unipartite matrix, which is symmetric and indicates the number of hospitals that were selected by a sufficient portion of people in both zip codes. The community detection algorithms then use this unipartite matrix to identify the markets based on common hospitals between zip codes. Using the same unipartite matrix, we run multiple CD algorithms, but focus on one specific market definition for the main results.

Appendix D Support for Labor and Delivery Analysis

This section contains the robustness checks and alternative specifications for hospital choice in labor and delivery. We include modifications to the main specification and assess the sensitivity of our results to alternative market definitions. Taken together, these results lend support to the main results.

Appendix D.1 Discrete Choice Results without Instrument

The preferred specification uses an instrumental variable to produce consistent estimates of the effect of Yelp ratings on hospital choice. Hospital characteristics such as reputation, amenities, and other non-clinical aspects of care are likely to affect both Yelp star ratings and hospital choice, but in the absence of controls for these features, estimates without an instrument will suffer from omitted variable bias. For completeness, Table A9 presents the results corresponding to Equation 1, i.e., the specification that does not include the first stage residuals in the estimation. These percentile rank coefficients are approximately twice as large as those in the main specification, driving larger willingness to travel estimates. The results in Table A9 are biased upward due to the potential correlation between star ratings and other non-clinical, unobserved (to the researcher) features that affect choice, whereas the instrumental variable results explicitly capture the effect of the star ratings, devoid of these underlying quality elements.

Appendix D.2 Alternative Specifications

A common practice in analyses of hospital choice is to model utility as a function of differential distance, i.e., the distance between a patient's home and a given hospital, minus the distance to the closest hospital in the choice set. Table A10 presents these results, which replaces the raw distance variable used in the main specification with the differential distance measure. The results are unaffected by this modification, which coincides with the fact that the majority of these patients have a hospital in (or very close to) their home zip code.

Table A9: Discrete Choice Model Results: No Instrument

	(1)	(2)	(3)	(4)
Percentile Rank	0.536***	0.525***	0.530***	0.514***
	(0.018)	(0.018)	(0.018)	(0.018)
Not Rated	0.206***	0.204***	0.233***	0.210***
	(0.013)	(0.013)	(0.013)	(0.013)
Distance	-0.156***	-0.145***	-0.133***	-0.154***
	(0.001)	(0.010)	(0.010)	(0.011)
Distance^2		-0.001***	-0.002***	-0.001***
		(0.000)	(0.000)	(0.000)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
Willingness to Travel	1.42***	1.391***	1.429***	1.333***
	(0.043)	(0.043)	(0.044)	(0.043)

NOTES: The table presents the results corresponding to Equation 1, i.e., the specification without instrumenting for percentile rank. Each column includes all of the same covariates as Table 3, with the exception of the residuals used in the control function approach. All specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Standard errors on the willingness to travel measures are calculated using the Delta Method. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Discrete Choice Model Results: Differential Distance

	(1)	(2)	(3)	(4)
Percentile Rank	0.2809***	0.2663***	0.2812***	0.3061***
	(0.0706)	(0.0721)	(0.0730)	(0.0736)
Not Rated	0.0565	0.0528	0.0855*	0.0866*
	(0.0421)	(0.0430)	(0.0441)	(0.0445)
Differential Distance	-0.1558***	-0.1739***	-0.1597***	-0.1809***
	(0.0005)	(0.0088)	(0.0089)	(0.0091)
Differential Distance^2		-0.0003	-0.0008*	-0.0001
		(0.0005)	(0.0005)	(0.0005)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
Willingness to Travel	0.662***	0.611***	0.656***	0.680***
	(0.166)	(0.165)	(0.170)	(0.163)

NOTES: The table presents the results corresponding to Equation 5a, but replaces any distance variable with the differential distance, i.e., the distance to a given hospital minus the distance to the closest hospital in the patient's choice set. Each column includes all of the same covariates as Table 3. All specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Standard errors on the willingness to travel measures are calculated using the Delta Method. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

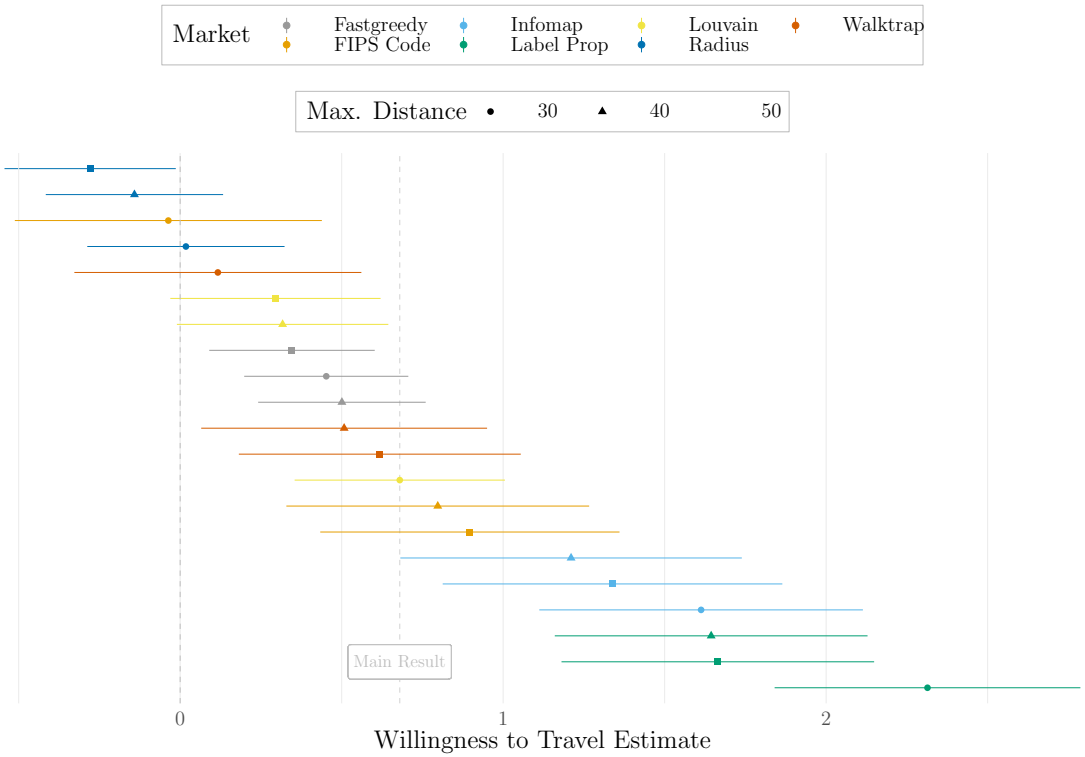
Appendix D.3 Sensitivity to Market Definition

Recall that the main specification uses community detection based markets with the additional restriction that any hospital in a patient's choice set must be within a 30 mile radius. I assess the robustness of the main results to the selected market definition using various community detection algorithms and FIPS codes, along with various radii around the patient's home zip code.

Figure A3 presents the willingness to travel estimates for the most saturated specification across various market definitions. The main result is approximately centered among the alternative results. The "Fast Greedy", "Info Map", "Label Prop", "Louvain", and "Walk Trap" markets are all based on the corresponding community detection algorithms, and overall, these results with these market definitions lend further support for to the conclusions of the main results. Note that for most of the community detection methods, the estimates are larger when I layer in the 30-mile restriction compared to the same market with the 40 or 50 mile restriction. This appears to be driven by the fact that, on average, as I expand the radius, there are more markets where only a small percent (less than 20%) of the hospitals in the market are rated. In contrast to the community detection markets, the radius based markets find null effects. This is not surprising given that this broad market definition may not necessarily reflect the set of hospitals from which expectant mothers are likely to choose. Overall, the estimates in Figure A3 demonstrate that the results are robust to various market definitions, with the arguably reasonable caveat that the markets must reflect patient flows.[48]

[48]Note that I do not estimate the model using HSA markets because those are based on Medicare patient flows and represent a fundamentally different patient population than the patients seeking care for labor and delivery. Similarly, HRR Code market definitions are based on tertiary care, which is likely not reflective of the referral patterns for labor and delivery. These boundaries can also cross state lines, but my admissions are limited to patients living in and admitted to hospitals in Florida. For these reasons, I concentrate on the hospital market definitions from the community detection methods.

Figure A3: WTT Estimates across Market Definitions



NOTES: The main results use the market definitions produced by the Louvain community detection method. The "Radius" markets include all of the hospitals within the respective mile radius around the patient's home zip code. FIPS Code markets are based on county FIPS codes with an added distance radius restriction as indicated. All other market definitions come from community detection methods.

Appendix E Support for Orthopedic Surgery Analysis

This section contains the robustness checks and alternative specifications that we conduct for the orthopedic surgery analysis. We present results with alternative market definitions and modifications of the main specification. The results coincide with the main results, bolstering the overarching conclusion.

Appendix E.1 Discrete Choice Results without Instrument

Recall that the preferred specification relies on an instrumental variable to deal with the potential endogeneity in the relationship between hospital Yelp ratings and hospital choice. Table A11, however, presents the results corresponding to Equation 1, i.e., the specification that does not include the first stage residuals in the estimation. The coefficients on the percentile rank are smaller than that of the main specification, resulting in smaller willingness to travel estimates.

Table A11: Discrete Choice Model Results: No Instrument

	(1)	(2)	(3)	(4)
Percentile Rank	0.2749***	0.3082***	0.2757***	0.2788***
	(0.0233)	(0.0236)	(0.0236)	(0.0236)
Not Rated	0.0733***	0.0968***	0.0766***	0.0773***
	(0.0170)	(0.0172)	(0.0172)	(0.0172)
Distance	-0.1172***	0.0080	-0.0026	-0.0029
	(0.0005)	(0.0119)	(0.0121)	(0.0121)
Distance^2		-0.0007***	-0.0006***	-0.0006***
		(0.0002)	(0.0002)	(0.0002)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
Willingness to Travel	0.860***	0.824	0.735	0.743
	(0.073)	(0.063)	(0.063)	(0.063)

NOTES: The table presents the results corresponding to Equation 1, i.e., the specification without instrumenting for percentile rank. Each column includes all of the same covariates as Table 5, with the exception of the residuals. All specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Standard errors on the willingness to travel measures are calculated using the Delta Method. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix E.2 Alternative Specifications

One potential concern about the main specification is that online reviews at the end of quarter $t-1$ may not be useful to patients going in for surgery early in quarter t. We

therefore conduct a supplemental analysis where hospital choice at time t is based on the percentile rank of a hospital's Yelp rating in $t-2$. The results are presented in Table A12 and find effects that are largely similar to the main results, particularly in the most saturated specification.

Table A12: Discrete Choice Model Results: Percentile Rank in $t-2$

	(1)	(2)	(3)	(4)
Percentile Rank	1.9622***	1.9995***	1.9612***	1.6851***
	(0.0996)	(0.1004)	(0.1005)	(0.1008)
Not Rated	1.0213***	1.0445***	1.0214***	0.8623***
	(0.0571)	(0.0576)	(0.0577)	(0.0578)
Distance	-0.1172***	0.0050	-0.0040	-0.0028
	(0.0005)	(0.0119)	(0.0121)	(0.0121)
Distance^2		-0.0006***	-0.0006***	-0.0006***
		(0.0002)	(0.0002)	(0.0002)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
Willingness to Travel	6.126***	5.313***	5.197***	4.463***
	(0.312)	(0.267)	(0.266)	(0.267)

NOTES: The table presents the results corresponding to Equation 5a, i.e., but replaces percentile rank in $t-1$ with percentile rank in $t-2$. Analogously, the first stage includes an indicator for being rounded up in $t-2$ instead of $t-1$. Each column includes all of the same covariates as Table 5. All specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Standard errors on the willingness to travel measures are calculated using the Delta Method. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

While the main results include the centroid distance between a patient's home and a given hospital, another possible way to measure this variable is the differential distance between a given hospital and the distance to the closest hospital available to the patient. Table A13 presents these estimates where we replace the raw distance variable used in the main specification with the differential distance measure. The results are similar to the main specification, particularly when controlling for other hospital characteristics and clinical quality.

Table A13: Discrete Choice Model Results: Differential Distance

	(1)	(2)	(3)	(4)
Percentile Rank	2.0888***	2.1517***	2.1097***	1.7063***
	(0.1116)	(0.1123)	(0.1125)	(0.1120)
Not Rated	1.1448***	1.1873***	1.1619***	0.9203***
	(0.0666)	(0.0670)	(0.0671)	(0.0668)
Differential Distance	-0.1171***	0.0228**	0.0125	0.0139
	(0.0005)	(0.0105)	(0.0109)	(0.0109)
Differential Distance^2		-0.0011***	-0.0010***	-0.0010***
		(0.0002)	(0.0002)	(0.0002)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
Willingness to Travel	6.543***	5.735***	5.604***	4.542***
	(0.351)	(0.300)	(0.300)	(0.298)

NOTES: The table presents the results corresponding to Equation 5a, but replaces any distance variable with the differential distance, i.e., the distance to a given hospital minus the distance to the closest hospital in the patient's choice set. Each column includes all of the same covariates as Table 5. All specifications include a not rated indicator, hospital characteristics, hospital quality, and distance. Interactions with individual characteristics are layered in as indicated. Standard errors on the willingness to travel measures are calculated using the Delta Method. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix E.3 Sensitivity to Market Definition

The main specification for orthopedic surgery uses community detection based markets with the additional restriction that any hospital in a patient's choice set must be within a 100 mile radius. Figure A4 presents the willingness to travel estimate for the main results relative to various additional estimates based on market definition using other community detection algorithms, FIPS codes, HSAs, and various radii around the patient's home zip code.

The main result is at the higher end of the alternative estimates, but among other comparable markets, the estimates are reasonable. The radius and Fast Greedy markets are the most comparable to the sample sizes for the main specification, whereas the FIPS code, Infomap, and Walktrap markets, each result in sample sizes that are about 50% smaller than the sample used for the main results. Each of these approaches produces much more granular markets than are applicable to this setting, and upon layering in additional limitations to ensure that there are sufficient hospitals on Yelp in a given market, we are left with a sample that is likely not a representative subset of admissions. Similarly, the HSA market estimates have much smaller sample sizes, and when compounded with the additional restrictions necessary to analyze the effect of star ratings on choice, the sample is limited to no more than 14,000 admissions.[49] This consists of only four markets, namely Jacksonville, Tampa,

[49]Note that the HRR Code market definitions are based on tertiary care, which is likely not reflective of the referral patterns for secondary care, such as orthopedic surgery. Additionally, these boundaries can cross state lines, but my admissions are limited to patients living in and admitted to hospitals in Florida. Of the

St. Petersburg, and West Palm Beach. Results with this market definition, therefore, are not comparable with the main results, and instead provide an estimate of how these star ratings affect a subset of urban markets.

Figure A4: WTT Estimates across Market Definitions

NOTES: The main results uses the Louvain market definition. The "Mile Radius" markets include all of the hospitals within the respective mile radius around the patient's home zip code.

Appendix F Support for Price Analysis

The following subsections provide the supporting material for our analysis of the relationship between online reviews and hospital prices.

Appendix F.1 Summary Statistics

Table A14 presents the summary statistics for the data used for the price analysis.

existing market definitions, HSA codes are theoretically better-suited for this analysis.

Table A14: Summary Statistics

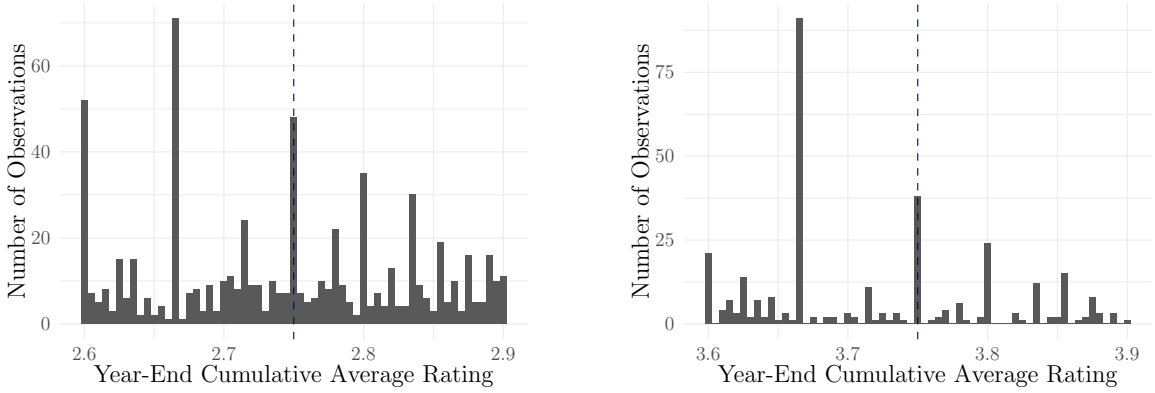
	Mean	St. Dev.	5th %tile	95th %tile
Price	9,340	3,970	3,794	16,908
Number of Reviews	18	33	1	78
Year-End Rating	2.9	1.1	1	5
Total Beds	234	205	45	627
Government	.13	.34	0	1
Non-Profit	.65	.48	0	1
System	.71	.45	0	1
Total Physicians	32	92	0	132
Total Nurses	441	480	68	1,335
Total Discharges	10,275	10,008	1,192	29,370
Total Medicaid Discharges	1,276	1,908	45	4,615
Cost per Discharge	24,760	12,093	12,896	43,949
Major Teaching Hospital	.071	.26	0	1
Any Teaching Hospital	.48	.5	0	1
Population	850,936	1721390	25,311	4242997
Unemployment	6.3	2.3	3.4	11
Poverty Rate	16	5.5	7.3	26
Uninsured	14	5.8	5.5	24
Median Income	53,018	14,103	35,165	81,992
Case Mix Index	1.5	.25	1.1	1.9
30-Day Mortality (Heart Failure)	12	1.6	9.4	15
30-Day Readmission Rate (Heart Failure)	23	2.1	20	26
30-Day Mortality (Pneumonia)	13	2.9	9.5	19
30-Day Readmission Rate (Pneumonia)	18	1.6	15	20
30-Day Mortality (AMI)	15	1.5	12	17
30-Day Readmission Rate (AMI)	18	1.6	16	21

NOTES: The first panel presents the deflated hospital price measure, followed by Yelp review data, and then hospital and county characteristics, and hospital quality metrics. Variables that range from zero to one are indicators. "Major Teaching Hospital" indicates hospitals that are members of the Council of Teaching Hospital of the Association of American Medical Colleges. "Any Teaching Hospital" indicates hospitals that satisfy a broader definition of teaching hospital (i.e. residency training approval, medical school affiliation, etc.).

Appendix F.2 Instrument Construction

Figures A5a and A5b show the distribution of the average rating around the middle and high thresholds, respectively. Figure A5a contains some spikes but appears to be more uniformly distributed than Figure A5b. The distribution in Figure A5b is particularly dense at 3.67 and 3.75, which is likely because a hospital can have an aggregate rating of 3.67 with just 3 total reviews and 3.75 with just 4 reviews. The uniformity of the distribution shown in Figure A5a is plausibly driven by the fact that hospitals in the bandwidth around the middle threshold have nearly twice as many reviews as the hospitals situated around the high threshold.

Figure A5: Distribution of the Average Rating at Each Threshold



(a) Middle Threshold

(b) High Threshold

NOTES: The figures are limited to the observations within the 0.15 bandwidth around the respective threshold. The threshold is indicated by the vertical dotted line.

Appendix F.2.1 Manipulation of the Rating

Given the potential for higher ratings to affect hospital choice and ultimately increase prices, hospitals face incentives to manipulate their reviews to improve their ratings. If hospitals behave in such a way, this would invalidate our estimation strategy because the underlying assumption that rounding is exogenous would be violated. In this subsection, we examine the details of the platform and further analyze our data to address this concern.

We first consider the rules and restrictions that Yelp has in place to prevent businesses from manipulating the reviews on their profiles. One way that hospitals may attempt to affect their online presence is by deleting reviews; however, Yelp does not allow businesses to remove reviews from their profiles. If a business believes that a review violates Yelp's content guidelines, they (or any other user on the site) may report the review.[50] If it is determined that the review is in fact a violation of the guidelines, then it will be removed. This feature of the platform suggests that it is unlikely that hospitals are able to precisely manipulate their ratings simply by eliminating negative reviews. Another possibility would be for hospitals to counteract negative reviews by either soliciting flattering reviews from patients or posting fraudulent positive reviews; however, Yelp is adamant that business owners should not solicit reviews from their customers due to the obvious conflict of interest.[51]

Further, one of Yelp's most boasted characteristics is its proprietary recommendation software, which systematically applies a set of quality standards to reviews and does not allow the business or Yelp employees to override the output of the software.[52] The software takes into account a variety of aspects about both the reviewer and the review content when

[50] See https://www.yelp.com/guidelines for more details.

[51] Yelp's policy on soliciting reviews is outlined here: https://www.yelp-support.com/article/Don-t-Ask-for-Reviews.

[52] Information on this policy can be found here: https://www.yelp-support.com/article/Does-Yelp-allow-employees-to-manually-override-the-recommendation-software.

determining whether or not to recommend a review.[53] These institutional details suggest that while business owners may attempt to manipulate their ratings, there are numerous policies and limitations that make doing so effectively rather complicated.

Nonetheless, in order to ensure that this does not occur in our data, we investigate the hospitals that fall around the threshold. If hospitals cannot delete reviews, their only option is to attempt to get positive reviews past the recommendation software. To assess this possibility, we analyzed the ratings of hospitals that fell within the 0.15 bandwidth at some point during our study period. Of the hospitals in the bandwidth around the high rated threshold (3.75), nearly 60 percent of the hospitals were rounded down, i.e., had a cumulative average of less than 3.75. Analogously, for the hospitals around the middle rated threshold (2.75), 47 percent of the hospitals had a cumulative average less than 2.75, meaning that they were rounded down. If hospitals were attempting to manipulate their aggregate rating, we would expect to see a clear majority of the hospitals in the bandwidth above the threshold, but that is not the case.

Further, if hospitals are exhibiting this behavior in a way that would invalidate our results, then we would expect to see high reviews submitted to bolster a low average rating. For example, if a hospital received n reviews in a given year, and its cumulative average with $n-1$ reviews was low, we would expect that hospital's n-th review to exceed the existing cumulative average. We examine this possibility in Figure A6, which plots the difference between the final rating a hospital received in a given year and the cumulative average up until that point. The mean values for high and middle rated hospitals are shown by the circles and triangles, respectively. The interval lines represent one standard deviation. If hospitals were posting positive reviews to counteract a lower aggregate rating, we would expect to see these point estimates consistently fall above the dashed line at zero. However, that does not appear to be the case around either of the rating thresholds.

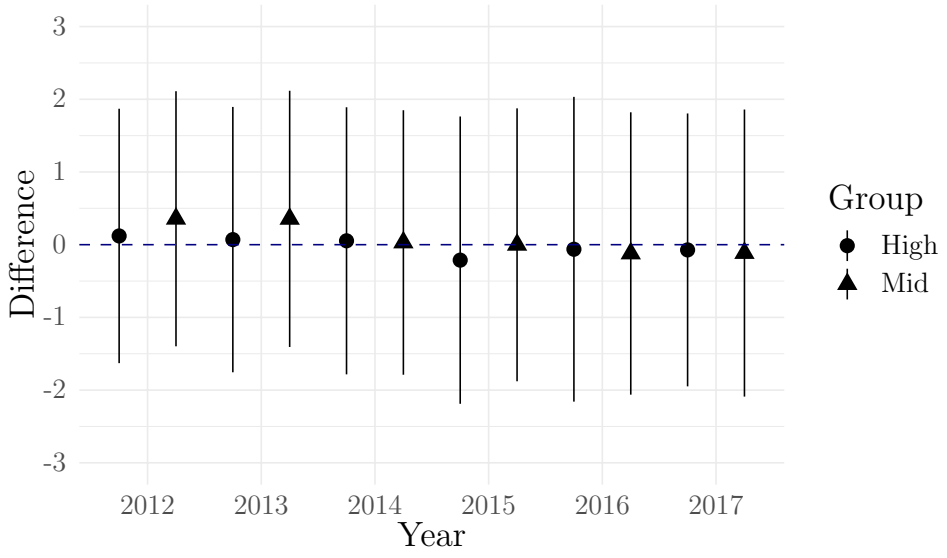
Lastly, there does not seem to be any apparent clustering to the right of either threshold, which would possibly be evidence of manipulation. To explore this possibility more formally, we present density tests around each threshold using the methodology presented in Cattaneo et al. (2018). The results of the density tests are reflected in Figures A7a and A7b. Both show statistical evidence of sorting; however, this result seems to be driven by the mass of hospitals that have an average rating exactly at the threshold. Recall that with as few as four reviews, the average rating can be exactly equal to the threshold value. For instance, 58% (68%) of the observations with the average rating equal to 2.75 (3.75) have exactly four reviews. Apparent jumps in the density of the ratings may therefore be a mechanical byproduct of the rounding. Indeed, when we consider the same density test but require a minimum of five reviews to be considered a "rated" hospital, in which case we find no statistical evidence of sorting at either threshold.[54]

The data and institutional details of the Yelp platform therefore suggest that hospitals cannot precisely manipulate their ratings. This supports the underlying assumption that being rounded into a higher rating is exogenous and is not affected by unobserved hospital

[53] This source outlines Yelp's practices regarding review recommendation: https://www.yelp-support.com/article/Why-would-a-review-not-be-recommended.

[54] We present the results of our analysis that correspond to this restriction, along with higher review count requirements. The point estimates are higher and remain significant, indicating that selective sorting is not inflating our results.

Figure A6: Difference between Last Rating and Prior Cumulative Average for Rounded Hospitals



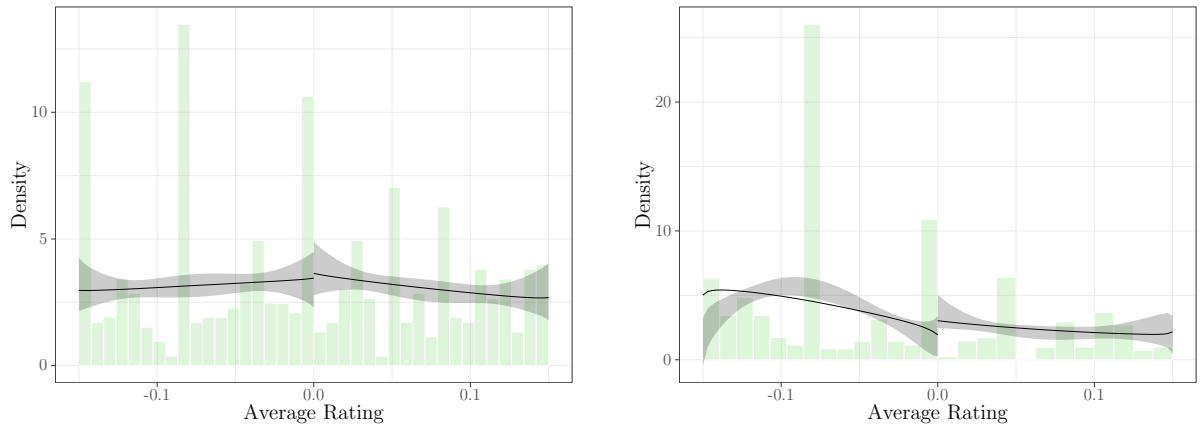
NOTES: The points represent the mean value for the difference, and the interval lines represent one standard deviation.

characteristics that also affect prices.

Appendix F.3 Covariate Balance

Lastly, to test the assumption that hospitals on either side of each threshold are comparable with the exception of their rounding status, we present the covariate balance for the observations within the bandwidth around the middle and high thresholds (Austin, 2009; Stuart, 2010; Austin, 2011; Zhang et al., 2019). Figure A8 shows the balance for hospitals rounded into high versus middle (circles) and hospitals rounded up to middle versus low (triangles). The majority of the covariates show no discernible difference, and for the two covariates that fall outside of the 0.10 bandwidth, the differences are less than 0.15. This further supports the assertion that observations on either side of the threshold are comparable with the exception of their rounding status. In the context of our estimation, this ameliorates the concern that our results could be driven by other hospital characteristics that happen to be more prevalent in hospitals that were rounded up into a higher rating category.

Figure A7: Manipulation Tests at the Middle and High Thresholds

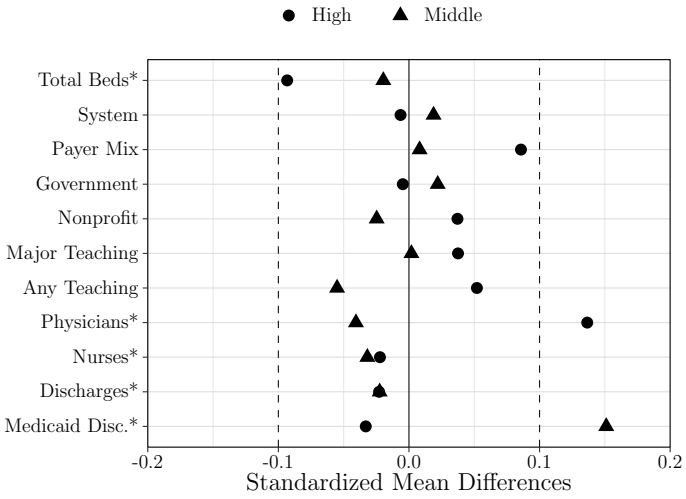


(a) Middle Threshold

(b) High Threshold

NOTES: The graph depicts the density tests presented in Cattaneo et al. (2018). The x-axis shows the average rating for a hospital at year-end. The density estimates are on the y-axis. The light green bars show the histogram of the average ratings. The bandwidth is 0.15. At the middle threshold there is evidence of sorting (p-value is 0.0418). We reach the same conclusion for the hospitals around the high threshold (p-value is < 0.0000). Note that at the high threshold, the point estimates are not contained in the shaded confidence interval. Upon further inspection, this appears to be driven by the density of hospitals that have an average of 3.67; when we require hospitals to have more reviews the point estimates are contained in the confidence intervals. Lastly, note that the confidence intervals are not always symmetric around the point estimates. Cattaneo et al. (2018) states that their test uses robust bias-corrected methods which causes the asymmetric confidence intervals.

Figure A8: Covariate Balance Plot of Hospitals around High and Middle



NOTES: The standardized mean differences between hospitals above the high threshold (3.75) and those below it, are shown with circles. The standardized mean differences between hospitals that fall above the middle threshold (2.75) and below it are represented by triangles. In either group, the observations under consideration are those within the 0.15 bandwidth around the threshold. The covariates analyzed here are defined in the discussion of Table A14. The symbol * denotes variables that are shown per capita.

Appendix F.4 First Stage and Reduced Form Results

Tables A15 and A16 present our first stage and reduced form results, respectively. Each table presents the results of four specifications. The baseline specification includes a set of covariates that control for hospital and county level characteristics, along with year fixed effects.[55] County and hospital fixed effects along with hospital quality measures are then added across the specifications as indicated in the tables.

To investigate the validity of the instruments, we first discuss the first stage results shown in Table A15, which are based on the first stage equations presented in the main text. The first panel in Table A15 summarizes the relationship between the high rating and being rounded into high and middle, followed by the second panel which shows the relationship between the middle rating group and the two rounding instruments. As one would expect, there is a strong positive relationship between being rounded into the high group and a high rating, and a strong negative relationship with the high rating and being rounded into middle. The analogous relationship holds when we consider the middle rating group as the outcome. It is clear that the expected relationship between the rounding instruments and the corresponding endogenous variables is quite strong.

To understand the connection between our instruments and our outcome of interest, we

[55]Note that we include this specification because we lose variation in ratings over time once we condition on a hospital fixed effect. Further, the instrument should account for any endogeneity, meaning that a hospital fixed effect is not necessary for identification.

Table A15: First Stage Results

	(1)	(2)	(3)	(4)
Panel (A)				
High Rating				
Rounded into High	0.875***	0.831***	0.755***	0.748***
	(0.0107)	(0.0202)	(0.0344)	(0.0377)
Rounded into Middle	-0.112***	-0.0917***	-0.0641***	-0.0808***
	(0.00957)	(0.0102)	(0.0119)	(0.0161)
Panel (B)				
Middle Rating				
Rounded into High	-0.384***	-0.399***	-0.584***	-0.588***
	(0.0152)	(0.0209)	(0.0370)	(0.0402)
Rounded into Middle	0.628***	0.604***	0.522***	0.536***
	(0.0146)	(0.0192)	(0.0286)	(0.0344)
County Fixed Effects	No	Yes	Yes	Yes
Hospital Fixed Effects	No	No	Yes	Yes
Hospital Quality Measures	No	No	No	Yes

NOTES: Panel (A) shows the regression where the outcome of interest is an indicator for having a high rating, and the independent variables of interest are the indicators for being rounded into middle and high. Panel (B) shows a different regression where the outcome of interest instead is an indicator for having a middle rating, but the independent variables are unchanged. All specifications include a set of hospital and county level characteristics, along with year fixed effects. Additional fixed effects and controls are indicated in the respective columns and apply to both panels. Robust standard errors clustered at the hospital level are in parentheses. Stars indicate the following: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A16: Reduced Form Results of Rounding Instrument

Price	(1)	(2)	(3)	(4)
Rounded into High	0.0219	0.0395	0.00144	0.0230
	(0.0280)	(0.0258)	(0.0158)	(0.0154)
Rounded into Middle	0.0575**	0.0363**	0.0324***	0.0214
	(0.0230)	(0.0173)	(0.0123)	(0.0150)
County Fixed Effects	No	Yes	Yes	Yes
Hospital Fixed Effects	No	No	Yes	Yes
Hospital Quality Measures	No	No	No	Yes
F-test of Coefficients (p-value)	0.0353	0.0394	0.0318	0.128

NOTES: All specifications include a set of hospital and county level characteristics, along with year fixed effects. Additional fixed effects and controls are indicated in the respective columns. Robust standard errors clustered at the hospital level are in parentheses. The F-test results show the p-values for the joint significance of the coefficients on the two variables shown in the table. Stars indicate the following: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

present the reduced form results—i.e., the regression of price on the instruments—in Table A16. The four specifications shown in the table are analogous to those discussed above: each includes hospital and county characteristics and year fixed effects, with additional controls indicated in the respective column. From Table A16, we see that there is a weak but positive relationship between the "rounded into high" instrument and price. Additionally, across all specifications, there is a positive relationship between the "rounded into middle" instrument and price, which is highly significant until the most saturated specification. The F-test of the coefficients shows that the instruments are jointly significant at conventional levels for the first three specifications.

Appendix F.5 Alternative Specifications

We expand on our main results—namely column (4) of main results table—with a series of alternative specifications and sensitivity analyses, including different bandwidths to define the instrument, increases in the minimum number of ratings required to be considered a "rated" hospital, and different threshold values for defining middle and high rated hospitals. Figure A9 presents the coefficient estimates on the high and middle rating groups, along with the "no reviews" variable for these specifications. They are shown in tandem with our main results, which are presented in blue and are indicated in the row "Main." The specification corresponding to a given point estimate is indicated with a black circle in the "Specifications" panel, and the relevant minimum number of reviews is similarly identified in the "Min. Reviews" panel below. The following subsections detail each of the alternative

specifications presented in Figure A9.

Appendix F.5.1 Minimum Number of Reviews

Our baseline results require a hospital to have a minimum of 3 reviews to be included in a rating group, but we also consider alternative minimum numbers of reviews. These are indicated in the "Min. Reviews" panel in Figure A9, where the minimum value ranges from 4 to 10. We also include one specification that sets the minimum number to zero, meaning that any rated hospital is included in the low, middle, and high groups, thereby eliminating the "too few" designation. These results coincide with the main specification.

We include each of these estimates to show how the informational value of the rating is affected by the number of ratings that comprise it. For example, a review of 4.5 stars based on 10 reviews may offer a stronger signal than that of a 4.5 rating based on just 3 reviews. The results shown here indicate that this may be the case. The point estimates on the high and middle groups exceed that of the main specification when the minimum number of reviews is 6 or above. While the confidence intervals on these are relatively wide, they indicate that the returns to a higher rating may be heightened by a larger number of reviews, particularly for the high rated group.

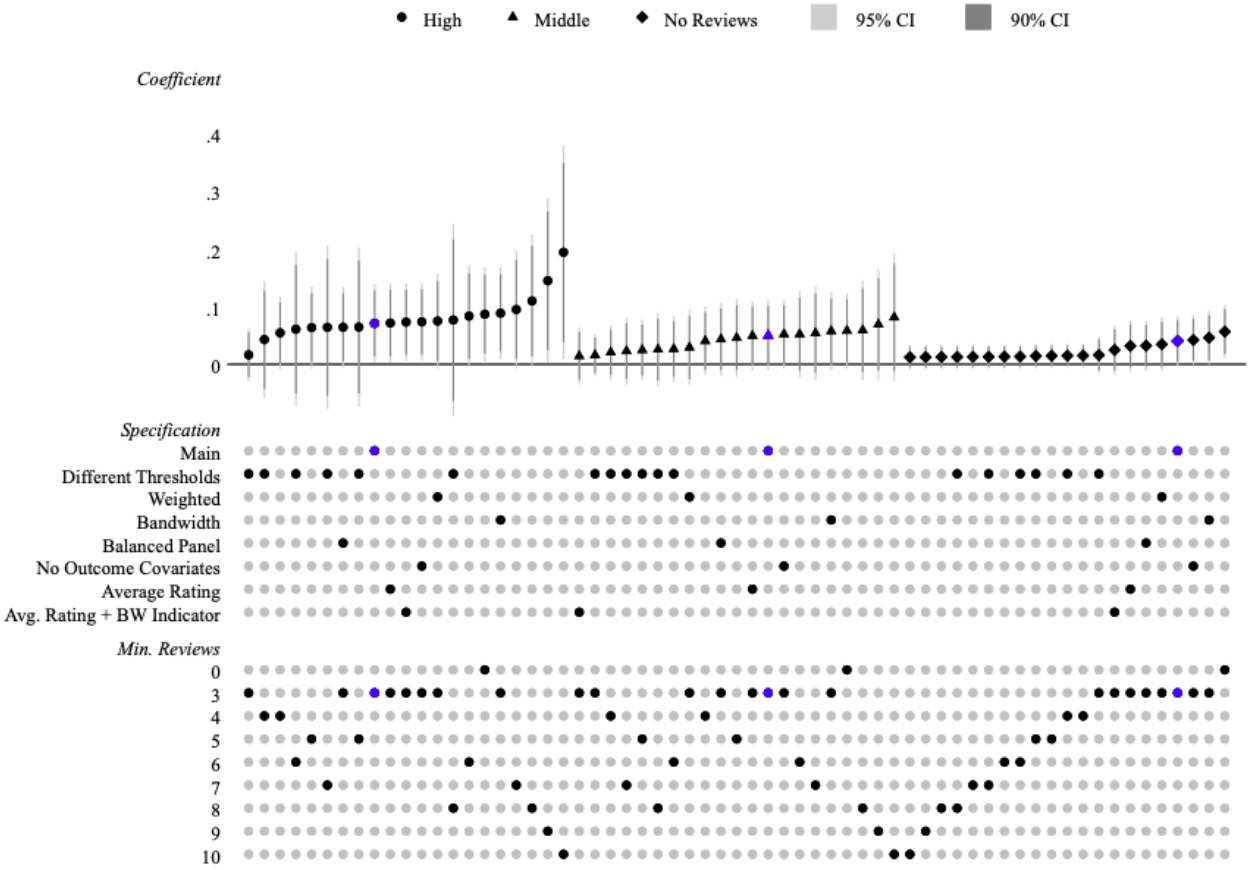
We also present results where we weight the estimation by the number of reviews (the "Weighted" specification), again with a minimum of 3 reviews to be considered "rated." As with the prior specification, the goal here is to accommodate the idea that the number of reviews may be informative in addition to the rating itself. While the coefficient on middle rating is smaller and less precise than the main result, the coefficient on high rating is quite similar to that of the main specification.

Appendix F.5.2 Different Rating Groups and Bandwidths

The "Different Thresholds" specification changes the definition of a high rating to hospitals that have 4.5 or 5 stars (instead of 4, 4.5, or 5 stars), and the middle rating to hospitals that have 3.5 or 4 stars (instead of 3 or 3.5 stars), with low rated hospitals having 3 stars or fewer. The goal is to test the sensitivity of our results to our definitions of the rating groups. For fewer numbers of minimum reviews, these estimates are lower than the main results, and estimates are larger with an increased number of minimum reviews. This is again consistent with the idea that the number of reviews may act as a proxy for the informational value of the observed rating. We estimate another set of results with more granular rating groups than in our preferred specification. While the narrowly defined rating groups offer less precise estimates, they align with the conclusions of our main results.[56] Lastly, we include the "Bandwidth" specification, which changes the 0.15 bandwidth used throughout the paper to 0.1. The results are robust to this change, finding slightly higher, statistically significant point estimates for each coefficient of interest.

[56]These results are available upon request.

Figure A9: Alternative Specifications



NOTES: The figure shows the coefficients for our variables of interest for our main specification (indicator for "Main") in conjunction with the results for various alternative specifications. The "Specification" panel signifies which approach is used, and the "Min. Reviews" panel corresponds to the number of reviews required in that specification for a hospital to be considered rated. "Different Thresholds" changes the definition of a high rating to 4.5 or 5 stars, a middle rating to 3.5 or 4 stars, and a low rating to 3 stars and below. "Weighted" weights the regression by the number of reviews that hospital received. The "Bandwidth" specification changes the bandwidth from 0.15 to 0.10. "Balanced Panel" imposes a balanced panel, eliminating the hospitals that do not appear in the data entire sample period. "No Outcome Covariates" drops any covariates that may be outcomes for ratings such as staffing and discharge values. The final two specification include the average rating as a covariate and the final specification also includes an indicator for whether or not the hospital is in the 0.15 bandwidth.

Appendix F.5.3 Balanced Panel

Next we turn to the "Balanced Panel" results. In our main results, we do not impose a balanced panel, and as such, we make this imposition here to ensure our results are not particularly sensitive to that choice. Hospitals may not appear in all periods due to mergers, acquisitions, and closures, or because of outlier prices in some years. We see that in the case of a balanced panel, the coefficients are nearly identical to those of the main results.

Appendix F.5.4 Average Rating

The "Average Rating" specification simply modifies our main results by including the average rating as a covariate. Recall that the ratings that appear on Yelp are discrete signals of quality, based on the continuous, underlying average rating. Further, our analysis estimates the change in price as a result of an improved quality *signal*, conditioning on underlying quality—not a change in quality itself. Thus we include the average rating as a covariate here to capture any additional underlying quality that is not controlled for in the existing covariates. As shown in Figure A9, our results are unchanged by including this covariate. We further this specification by including an indicator to control for hospitals that fall within the bandwidth (titled "Avg. Rating + BW Indicator"). Here, the results at the high group are largely unchanged, and for the middle rating group, the effect remains positive but is smaller and statistically insignificant.

Appendix F.5.5 Included Covariates

Our preferred specification captures several observable hospital characteristics that may directly affect prices; however, some of these variables may also be affected by patient demand (and thus potentially affected by quality ratings). Therefore, in the "No Outcome Covariates" section of Figure A9, we drop covariates that may themselves be outcomes of the ratings. Such variables include the number of physicians and nurses, along with Medicare discharges and total discharges. The results are quite similar in light of the omission of these variables.

Appendix F.6 Sensitivity Analyses and Falsification

We consider the sensitivity of our estimates presented in the main text to several potential concerns. First, we note that we are able to precisely reject underidentification and reject the null hypothesis of weak instruments.[57] Second, following the work on "plausible exogeneity" in Conley et al. (2012), we show that our results are robust to mild violations of the exclusion restriction. We also show that our estimates are robust to the presence of outliers, which are known to be a potentially severe problem in IV estimates (Freue et al., 2013). Each of these results are available upon request.

Several falsification tests also lend confidence to our results. These results are available upon request. To summarize, we consider future clinical quality measures and lagged prices

[57] The Kleibergen-Paap LM statistic shown in main result is 85.49 and has a p-value of < 0.000. In the presence of cluster-robust standard errors, the test for weak identification is the Kleibergen-Paap rk Wald F statistic, the value of which is is 97.91.

as outcomes and find no evidence of a relationship between those variables and our rating categories. This tends to support our estimates as revealing a true underlying effect of higher ratings rather than a simple correlation between online reviews, clinical quality, and prices.