Distribution Agreement

Kaylyn R. Sanbower | Date

Essays on the Effects of Information Disclosure in Hospital Markets

By

Kaylyn R. Sanbower

Doctor of Philosophy

Economics

Ian McCarthy, Ph.D.
Advisor

Sara Markowitz, Ph.D.
Committee Member

Stephen O'Connell, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Essays on the Effects of Information Disclosure in Hospital Markets

By

Kaylyn R. Sanbower

M.S., Emory University, 2020
B.S., Seton Hall University, 2015

Advisor: Ian McCarthy, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics
2022

Abstract

Essays on the Effects of Information Disclosure in Hospital Markets

By Kaylyn R. Sanbower

Hospital quality is complex and multidimensional, making it difficult to quantify and communicate. These challenges drive information problems that may affect the hospitals that patients choose, hospital prices, and hospital investment decisions. This creates the possibility for novel measures of hospital quality to affect hospital markets and the efficiency of health care delivery. Online reviews are one such measure; they provide an aggregate, accessible measure of hospital quality, which may affect these markets. Using hospital reviews from the online rating platform, Yelp, and numerous hospital data sources, the papers in this dissertation employ causal inference techniques to provide new evidence about the effects of online reviews on hospital markets and how hospitals might work to improve patient experience of care.

The first analysis shows that patients are willing to travel further to receive care from a hospital with a higher star rating. Reflective of this underlying mechanism, the second study shows hospitals can charge higher prices following a higher aggregate star rating. These findings highlight the incentives that hospitals face to prioritize patient experience and motivate the final chapter, which shows that hospital star ratings increase in response to windfall profits. Using the review text, additional analysis shows that this is driven, at least in part, by the presence of amenities, which is suggestive of hospital investment behavior. These studies further our understanding of information disclosure in hospital markets, informing policy and providing a foundation for further research.

Essays on the Effects of Information Disclosure in Hospital Markets

By

Kaylyn R. Sanbower

M.S., Emory University, 2020
B.S., Seton Hall University, 2015

Advisor: Ian McCarthy, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics
2022

Acknowledgments

It is hard to comprehend that around this time six years ago, I was gearing up for the next chapter of my life, blissfully unaware of the daunting uphill battle that awaited me in Atlanta. Graduate school has taken me on a journey through frustration, tears, and omnipresent self doubt, to breakthroughs, understanding, and—dare I say—confidence, which would not have been possible without an incredible support system.

I will forever be indebted to my committee—Ian, Sara, and Stephen—for helping me develop the research skills to frame and investigate important questions and for supporting me as I take the next step in my career. Ian—it is by no means hyperbolic to say that you are an incredible advisor, mentor, cheerleader, and friend. It has been a privilege to be your apprentice. I have learned so much from you, but most importantly, you've been a testament to the fact that your contribution is not limited to ideas and publications, but the role your character plays in improving the profession.

Sara, thank you for always challenging me to think about and express my ideas more clearly. You have made me a better economist and writer, for which my both audience and I are beyond grateful. Stephen, you have always given me such useful feedback and pragmatic advise. You helped me become a better presenter, feel confident in my career path, and take the job market in stride, and for that, I could not be more grateful. The three of you have had an incredible impact on my experience at Emory and budding career; I am quite fortunate to have had this dream team of a committee.

I am also incredibly grateful to the faculty and staff of the Economics Department. You have prioritized the growth and development of our graduate program, and my experience would not have been the same without you. In particular, I would like to thank David Jacho, Liz Eichinger, and Renee Sevy-Hasterok. David, you are intense but absolutely steadfast in your commitment to the program, and my experience was better for it. Liz, since I arrived at Emory you've helped me branch out and experience Atlanta outside of my role as a graduate student. Thank you for the wine, the laughs, and of course, the friendship. Renee, you have been an absolute gift to the program. You are kind, supportive, and extraordinarily organized. The latter feels like a lame compliment, but I'm sure you know that I say this with the utmost appreciation.

I would be remiss to not thank my wonderful cohort-mates and fellow graduate students. Specifically, Carla, Santi, Diego, Drew, Hanna, Katie, Pablo, and Noah—you have become some of my dearest friends, and I can't wait to visit you wherever building your lives and careers may lead. It has been a privilege to be surrounded by so many smart, ambitious people. Seeing all of you continually rise to the challenge has forced me to hold myself to a higher standard. Having struggled, failed, learned and succeeded together, I will forever feel a unique kinship with you.

There are two graduate students for whom this sentiment is overwhelmingly true: Cheng and Juan. Cheng, you are a ray of sunshine and my chosen family. You are brilliant, kind, and hilarious, and your friendship has brought a richness to the last five years that I could never have imagined. Juan, I admire your curiosity and dedication and trying to keep pace with you has made me a better researcher. (You'll likely refer to this as a peer effect, but in

this one context can we please just call it "love"?) You believed in me even when my own confidence was in short supply. I could not have imagined a better partner, now and forever.

I might never have mustered up the courage to pursue this degree without the support of my friends and family. Moreover, I might not have considered Emory it were not for my friend and mentor, Katherine Fidler. You supported me through the decision to leave the comfort of my prior career to find something more challenging, more rewarding, and were there for me in the tumultuous journey that followed, and for that, I am truly grateful. Sam–you have been my rock through two of the most challenging periods my life: graduate school and, while it will sound like a joke to anyone else reading this, college softball. You—my one-woman hype squad—were the best antidote to imposter syndrome; I am so astonishingly lucky to have met you and grown with you. Thank you for always being there for me. Soledad and Resfa—the first time I landed in Colombia, you welcomed me with open arms, and now, you've welcomed me into your family. Thank you for your ongoing love and support.

Finally, to my Mom and Dad—I dedicate this dissertation, the culmination of doctoral studies, to you. You taught me to believe that I can do anything I put my mind to and provided me with the unwavering support to do so. Throughout this journey you were there to celebrate my wins and comfort me through my losses; there would be no acknowledgements to write if not for you. And Dad, it likely goes without saying, but thank you for always reminding me to "be excellent." It came in handy a time or two.

I feel unbelievably fortunate to have had amazing family, friends, and mentors by my side over the past five years. While these acknowledgements cannot begin to express the extent of my gratitude, I hope they make clear the immeasurable role your love and support played in making this work and this degree possible.

Contents

List of Tables

List of Figures

Chapter 1

Online Reviews and Hospital Choice

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. Using the universe of hospital Yelp reviews and inpatient claims data for elective procedures in Florida from 2012 through 2017, this analysis exploits exogenous variation in online hospital ratings over time to identify the effect of this information on hospital choice. The analysis finds 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. These results indicate that novel, accessible sources of information have the potential to affect health care decisions.

1.1 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 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 information, but they do not encompass the breadth of features of care relevant to hospital choice (Dafny and Dranove, 2008; Dranove and Sfekas, 2008; Zhe 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, metrics that provide new and relevant information or present existing information in a more accessible way may have notable effects on hospital choice (Dranove and Jin, 2010; Cutler, 2011; Garthwaite et al., 2020). Online review platforms provide a novel, accessible means by which to gather information about the experience of care at a given hospital and may therefore be relevant to the hospital decision making process (Ranard et al., 2016).

This study examines the relationship between online reviews and hospital choice using hospital Yelp reviews and inpatient claims data from 2012 through 2017. During that time, Yelp was a particularly popular site for hospital reviews. Hospital profiles on the platform grew tenfold between 2010 and 2018, with nearly 50% of all general acute care hospitals represented 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. I use this rounding to construct an instrument for the percentile rank of a hospital's star rating in its respective market. I capture hospital choice using Florida inpatient claims data for planned or elective procedures—namely labor and delivery and orthopedic surgery—because these patients are better able to shop for a hospital than those with urgent or emergency admissions. I 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.

Overall, online reviews matter for hospital choice. The results show 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. 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. The results for both procedures are robust to alternative specifications and different market definitions. Falsification analyses that estimate the model in the context emergency department admission 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.

In identifying this causal relationship, my analysis contributes to the growing literature on the effect of information disclosure in health care. I 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; Zhe 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; Zhe Jin and Sorensen, 2006; Dafny and Dranove, 2008; Bundorf et al., 2009). 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, clinically driven measures (Dranove and Sfekas, 2008; Romley and Goldman, 2011; Scanlon et al., 2002; Pope, 2009; Chandra et al., 2016). My analysis both lends support to those existing findings and provides new evidence that identifies how a novel source of information—online reviews—is relevant to care decisions.

While economic theory predicts that decision makers will respond to a given type of information, whether or not they do must be investigated empirically. 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. My analysis informs this open question, demonstrating that online reviews affect hospital choice.

1.2 Information Disclosure in Hospital Care

Various efforts seek to divulge information on hospital quality. Existing 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.

Information disclosure 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. Dranove and Sfekas (2008) finds that when report card scores do not conform with existing beliefs patients select different hospitals, and Pope (2009) shows that the U.S. News and World Report ratings affect hospital choice.[1]

While these metrics provide valuable information, they also have notable shortcomings that can affect both decision making in hospital markets and our understanding of hospital quality more generally. For example, CMS has gathered quality of care metrics for some time, but their disclosure was fragmented due to legislative issues and other difficulties. CMS sought to improve hospital quality disclosure through its website,"Hospital Compare," which made these data available and aggregated them into 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).[2] The inconsistent availability of this information limits its potential to inform hospital choices.

Further, existing metrics may not address all of the relevant features of hospital care because

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

[2] In November 2020, CMS announced that they would aggregate each of their independent "Compare" platforms into a single quality compare site title "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.

they are limited to clinical quality measures and structured survey instruments. 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.[3] Additionally, these metrics report lagged quality information, which creates a notable barrier for decisions-makers who want to include timely information into their choice criteria. In summary, 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. Existing research finds that online 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). 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). Even if hospital reviews do not provide new information, they could affect choice

[3] 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.

if they makes 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. While this discussion demonstrates that online reviews possess the necessary characteristics to affect hospital choice, whether or not they do is an empirical question that this study informs.

1.3 Data

This paper analyzes the effect of online reviews on hospital choices using 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 I limit to elective admissions for specific medical needs (namely, labor and delivery, and orthopedic surgery). The study incorporates data from the American Hospital Association (AHA) Annual Survey, which is the most comprehensive source of data on hospital characteristics. Lastly, the analysis includes 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 the analyses.

1.3.1 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.[4] The platform continued to gain popularity through the 2010s, and while Google is now the most commonly used review site, Yelp appears to be the most prominent source of online review information during the study period, which begins in 2012 and ends in 2017.[5]

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. Then, users may 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 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

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

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

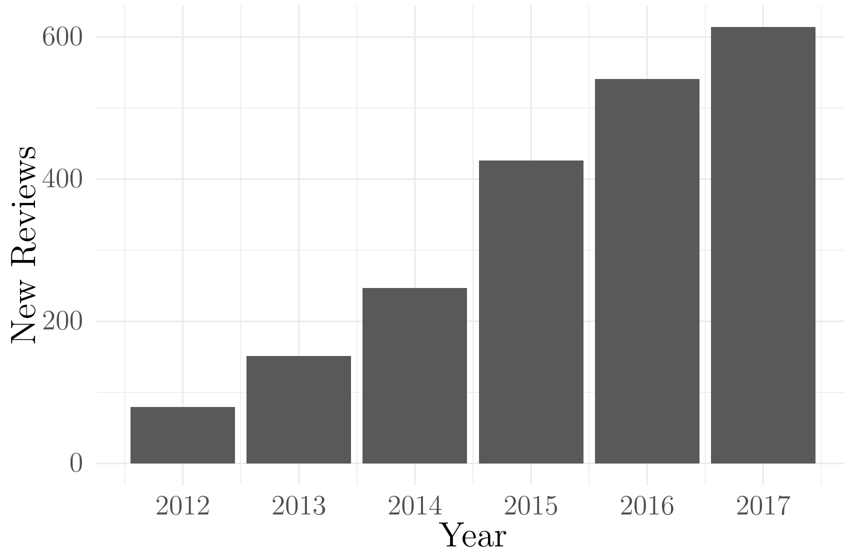
using web scraping methods to compile a dataset that consists of the date, the star rating, the review narrative, and the user ID for all of the hospital reviews on Yelp through year-end 2018.[6]

Using the star ratings from each of these reviews, I 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. I 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.

I 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 (McCarthy et al., 2020). 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 1.1 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 provides evidence that people use it to share their experiences.

[6] 1.9.1 provides details on the data collection and cleaning process.

Figure 1.1: New Hospital Reviews on Yelp by Year



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

1.3.2 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 time of admission. Lastly, in lieu of admission and discharge dates, the agency discloses quarter of admission.[7]

I limit the data to elective admissions at general acute care hospitals. This eliminates urgent or emergency room admissions, in order to isolate patients whose hospital choices were not

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

impaired by the circumstances of their admission. I 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, I eliminate observations with missing or foreign zip codes, along with any patients who are not from Florida. I drop any admissions to hospitals that are not included in the AHA data.

To analyze the effect of online reviews on hospital choice, I 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.[8] 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 patients—namely 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.

[8] 1.9.3 details the ICD-9 and ICD-10 codes used to identify the admissions for the respective procedures.

Labor and Delivery Data

I limit the labor and delivery data to admissions for patients with ages between the 5th to 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. I 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. 1.9.3 provides details on the public insurance options in Florida. Lastly, I 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 1.1 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 a third identify as Latina. The second panel in Table 1.1 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 I can calculate a hospital's observed rating at any point in time, I 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 1.1: 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 reviews data are measured at the quartlerly level to correspond with the unit of time available in the inpatient claims data.

Orthopedic Surgery Data

For orthopedic surgery admissions, I limit the data to Medicare fee-for-service (FFS) claims for knee or hip replacement among beneficiaries aged 65 and above.[9] By limiting the data to FFS admissions, I can focus on patients whose choices are not restricted by specific insurance networks. [10] I then omit any admissions in the top 5th percentile of ages, which results in an age range of 65 to 85. Lastly, I limit the analysis to hospitals that average at least one orthopedic surgery admission per quarter, to eliminate any hospitals that are not viable choices.[11] The resulting sample comprises 128,862 admissions at 132 hospitals.

Table 1.2 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 1.1.

1.3.3 Hospital Markets

To analyze hospital choice, I also need to determine the relevant market for each respective patient. In both antitrust litigation and economic research, hospital market definitions are a

[9] Information on the relevant diagnosis related group (DRG) codes is in 1.9.3.

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

[11] This differs from the minimum requirement that I 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.

Table 1.2: 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 reviews data are measured at the quartlerly level to correspond with the unit of time available in the inpatient claims data.

point of contention given the inextricable link between these definitions and the conclusions of the analyses in which they are employed. Gaynor et al. (2013) notes that market definitions are often the determining factor in antitrust cases. In economic research, hospital market definitions play a critical role in analyzing the various competitive forces in these markets. 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 also have notable shortcomings. HRRs, for instance, are based on referral patterns for tertiary surgery and have not been updated since 1993. They are, therefore, unlikely to accurately capture hospital markets for patients seeking other types of care in recent years (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. However, depending on the analysis, they are still limited to what is likely not a representative sample of patients. Lastly, commonly used geographic boundaries may impose limits that may not be characteristic of a patient's true choice set.

Community detection (CD) algorithms provide a novel way for researchers to define hospital markets that addresses the shortcomings of existing market definitions (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.[12] 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 indicates that researchers

[12] 1.9.4 describes the community detection method in greater detail.

can define markets in ways that may better reflect the competitive landscape and observed choices of the hospitals in their analyses.

I employ these methods to define procedure-specific hospital markets using zip code level patient flows.[13] Separately for each procedure, I aggregate the claims data to determine the number of patients from a given zip code admitted to each hospital. I then use these aggregate values to implement the CD method and determine the relevant hospital markets.[14]

Regardless of procedure, there are additional features of the data and my 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, I use the broadest market definition from the community detection methods, and then I layer in additional restrictions. I 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, I 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.[15] Ultimately, these

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

[14]1.9.4 details the CD methodology in the context of this analysis.

[15]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.

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, I limit the patient's choice set to hospitals whose centroid distance from the patient's home zip code falls within a certain radius. Sections 1.5 and 1.6 detail their respective choice sets used for estimation, which include these restrictions and procedure-specific caveats.[16]

1.4 Empirical Approach

To investigate the effect of online reviews on hospital choice, I model 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.[17] Therefore, I estimate the following model where patient i's utility from receiving care at hospital j at time t is defined as:

$$\begin{aligned} u_{ijt} &= v_{ijt} + \epsilon_{ijt} \\ &= \beta_1 P_{j,t-1} + \beta_2 NR_{j,t-1} + D'_{ij}\alpha_d + H'_{jt}\alpha_h + Q'_{jt}\alpha_q + X'_{ijt}\alpha_x + \varepsilon_{ijt}. \end{aligned} \tag{1.1}$$

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

[17] 1.9.2 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.

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., $t-1$.[18] Specifically, $P_{j,t-1}$ is the percentile rank of a hospital's star rating among the hospitals in its market. By using the percentile rank instead of raw star ratings, I capture a hospital's quality information relative to other hospitals on the platform.[19] 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. 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

[18] 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 this information. The supplemental material includes results with a two quarter lag, but the results are unchanged, which likely reflects the limited flow of new reviews per quarter (1.3 on average).

[19] 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 and an indicator to designate that they are not rated.

for both procedures are age, and race and ethnicity indicators including Black and Latino. For labor and delivery, I also include an indicator for public insurance because the data have a mix of private and publicly insured patients. For orthopedics, I 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 these data, there is no outside option. I estimate the underlying utility parameters using maximum likelihood.

I am 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.[20] Moreover, I am 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, I use a control function approach to 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 I use to instrument for the percentile rank of a hospital's star rating. Specifically, I 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

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

midpoint.

Using this instrument, I 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 (Petrin and Train, 2010). To implement, I first estimate the following equation,

$$P_{j,t-1} = \gamma R_{j,t-1} + \zeta NR_{j,t-1} + D'_{ij}\psi_d + H'_{jt}\psi_h + X'_{ijt}\psi_x + \epsilon_{ijt}, \tag{1.2}$$

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 1.1. I then include the residuals, $\widehat{\epsilon}_{ijt}$, in the following equation to recover a consistent estimate of β_1. The second stage, therefore, is

$$u_{ijt} = \beta_1 P_{j,t-1} + \beta_2 NR_{j,t-1} + D'_{ij}\alpha_d + H'_{jt}\alpha_h + X'_{ijt}\alpha_x + \widetilde{\varepsilon}_{ijt}, \tag{1.3}$$

where $\varepsilon_{ijt} = \eta\epsilon_{ijt} + \widetilde{\varepsilon}_{ijt}$, and $\widehat{\epsilon}_{ijt}$ is an estimate for ϵ_{ijt}. I use this specification for the main results. Additionally, to provide a more readily interpretable result, I 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

$$WTT = -\frac{\partial U_{ij}}{\partial P_{ij}} \Bigg/ \frac{\partial U_{ij}}{\partial D_j} \times SD(P). \tag{1.4}$$

For Column (1) of Tables 1.4 and 1.7, 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

$$WTT = \frac{-\beta_1}{\alpha_d + 2\alpha_{d^2} D + A + B} \times SD(P), \tag{1.5}$$

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 to the terms of $\frac{\partial U_{ij}}{\partial D_j}$ that correspond to the interactions between distance squared and patient characteristics.[21] I calculate standard errors using the delta method. The following sections detail the choice sets, model results, and willingness to travel estimates by procedure.

1.5 Choice in Labor and Delivery

I 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 con-

[21] 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 $B = 2\psi_{d^2 \times a} D x^a + 2\psi_{d^2 \times b} D x^b + 2\psi_{d^2 \times l} D x^l + 2\psi_{d^2 \times pm} D x^{pm}$. The subscripts on the ψ terms signify to which interaction term the coefficient correspond. 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 payer 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.

tained more than one hospital.[22] 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, I drop hospitals from a patient's choice set if they are over 30 miles away. Lastly, as detailed in Section 1.3, I 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 three stars, with an average of 17 reviews (median 13).

Table 1.3 presents the estimation results for the first stage as shown in Equation 1.2. 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 beds, indicators for teaching hospital status, system membership, total nurse and physician counts, 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 and distance squared and its interactions with individual characteristics.

[22]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.

Table 1.3: First Stage Regression Results

	(1)	(2)	(3)	(4)
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)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
F-Statistic	314934	159167	112143	101722
R^2	0.7470	0.7491	0.7545	0.7550

NOTES: The dependent variable across all specification is percentile rank. 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

I then implement the control function approach by including these residuals in Equation 1.3 (Petrin and Train, 2010). Table 1.4 presents these results, where each column includes the residuals from the corresponding column in Table 1.3. These 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 interactions terms between distance and individual characteristics inhibits the ability to interpret this directly. Instead, the bottom row of Table 1.4 presents the WTT results, where, in the most saturated specification, I find that patients are willing to travel an additional 0.68 miles to receive care from a hospital with 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

Table 1.4: Discrete Choice Model Coefficient Estimates

	(1)	(2)	(3)	(4)
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)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
Willingness to Travel	0.662***	0.611***	0.655***	0.680***
	(0.166)	(0.167)	(0.172)	(0.166)

NOTES: The table presents the results corresponding to equation 1.3. Each column includes as a covariate the residuals from the corresponding column in Table 1.3. 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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 1.4 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 1.5 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.

These results are robust to a variety of modifications and alternative specifications, all of which are detailed in 1.9.5. I estimate the model where I 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 my result to the chosen market definition, I 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.

Table 1.5: Discrete Choice Model Coefficient Estimates: Ratings in $t-2$ and $t-3$

	(1)	(2)	(3)	(4)
Panel A: 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 B: 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 1.3, 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 1.4. 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

1.6 Choices in Orthopedic Surgery

I identify markets for orthopedic surgery using patient flows to the 132 hospital 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, I 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, I 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 Table 1.6 and indicate that the instrument for being rounded into a higher rating is positively and significantly related to percentile rank. Each column also includes hospital characteristics (total beds, indicators for teaching hospital status, system membership, total nurse and physician counts, payer mix, and case mix index), readmission rates for hip and knee replacement and for the hospital overall, and the distance between the patient's home zip code and the hospital. I include interactions between these terms and individual characteristics (age, sex, and indicators for Black, and Hispanic or Latino) as indicated in the table. In columns (2) through (4) where interactions with "distance variables" are included, the specification consists of distance and distance squared and its interactions with individual characteristics.

Table 1.6: First Stage Regression Results

	(1)	(2)	(3)	(4)
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)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
F-Statistic	187027	97087	67312	55883
R^2	0.7414	0.7416	0.7419	0.7420

NOTES: The dependent variable across all specification is percentile rank. 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.7: Discrete Choice Model Coefficient Estimates

	(1)	(2)	(3)	(4)
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)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
Willingness to Travel	5.28***	4.534***	4.436***	4.41***
	(0.349)	(0.298)	(0.297)	(0.298)

NOTES: The table presents the results corresponding to equation 1.3. Each column includes as a covariate the residuals from the corresponding column in Table 1.6. 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

To my 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 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 are notably smaller than what I find in my 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. My estimates 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 1.5, wherein I would expect a larger WTT estimate for orthopedic surgery relative to labor and delivery.

My main results hold up to various alternative specifications, all of which are detailed in 1.9.5. I estimate the model where I replace distance with differential distance and the estimates are nearly identical. I also estimate hospital choice at time t based on ratings in $t-2$, whereas the main specification uses star ratings in $t-1$. These results are quite similar to the main results. Lastly, I assess the sensitivity of my 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.

1.7 Falsification Analysis

To assess the credibility of the results in Sections 1.5 and 1.6, I estimate the same model, but among patients for whom online reviews should not affect their chosen hospital. I therefore 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, I limit the sample using the same criteria outlined in Section 1.3.[23]

I 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

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

19% have private insurance. This results in 7,321,225 emergency admissions across 139 hospitals. I use all of these admissions to construct hospital markets, but due to computational limitations, I use a random sample of 250,000 admissions to estimate the discrete choice model. Table 1.8 summarizes the entire sample. The second and third panel in Table 1.8 show that the hospital characteristics and their Yelp presence are similar to the labor and delivery and orthopedic surgery contexts.

Table 1.8: 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. I use a random sample of 250,000 observations to estimate the model. The hospital characteristics are measured annually. The Yelp reviews data are measured at the quartlerly level to correspond with the unit of time available in the inpatient claims data.

Table 1.9: First Stage Regression Results

	(1)	(2)	(3)	(4)
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)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
F	150523	80995	57694	45189
R^2	0.6762	0.6771	0.6809	0.6848

NOTES: The dependent variable across all specification is percentile rank. 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

I then use these data to define hospital markets. Analogous to the approach for labor and delivery and orthopedic surgery, I start with the most broad market definition from the community detection algorithms and then layer in additional restrictions. This means I 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. I 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, I estimate the model starting with the first stage as outlined in Equation 1.2. Table 1.9 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 Tables 1.3 and 1.6. 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.

Table 1.10: Discrete Choice Model Coefficient Estimates

	(1)	(2)	(3)	(4)
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)
Interactions with Individual Characteristics				
× Distance Variables		X	X	X
× Hospital Characteristics			X	X
× Clinical Quality				X
Willingness to Travel	-0.001	0.092	0.042	-0.010
	(0.133)	(0.112)	(0.112)	(0.112)

NOTES: The table presents the results corresponding to equation 1.3. Each column includes as a covariate the residuals from the corresponding column in Table 1.6. 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 1.10 presents the main results, where each column includes the residuals from the corresponding column in Table 1.9. Across each column, the coefficient on percentile rank is small and insignificant. Following Sections 1.5 and 1.6, I 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.

1.8 Discussion

This paper provides novel insights on the effects of online reviews on hospital choice. I 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.

This analysis also highlights 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.

1.9 Appendix

1.9.1 Yelp Data

I use the AHA Annual Survey database to match 2,935 hospitals to Yelp profiles with reviews.[24] For these hospitals, name associated with the profile exactly matched the name listed in the AHA data. I then implemented the following process to ensure that the profiles are associated with the correct hospital. I refer to these profile-hospital matches as "exact" matches.

1.9.2 Data Cleaning

1. I eliminated any observations without an address on the Yelp profile because I 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.

2. I 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.

3. Next I parsed out the first part of the string from both addresses. Most often this is the

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

street number, but it can also be a word (i.e. the street name) if the number is missing. I 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.

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

4. I kept 1,687 observations with exact address matches. This left 1,217 observations to be analyzed.

5. I 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').

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

7. I manually reviewed 258 observations with only matching street numbers or those with an '&' or 'and' in the name. I 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.

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

9. I dropped any observations where none of the first three words or numbers of the addresses matched. Any remaining observations were manually reviewed.

Manual Review of Observations

To manually review the remaining observations, I 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, I used Google Maps to determine whether or not they were referring to the same location. I examined any remaining profiles using this process. Any addresses that did not refer to the same location were dropped.

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. I 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 I mention it here to provide more additional clarity on the data collection process.

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 I cannot ask the patients directly if they used online reviews, I can analyze the text of the review comments to determine if reviewers mention using reviews to inform their decisions. I 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.[25] I 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. I 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, I find that one percent of the reviews meet this criteria. I read a sample of 50 of these reviews and found that each explicitly mentions consulting online reviews.

Based on these findings, I 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

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

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. I 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.

1.9.3 Florida Data

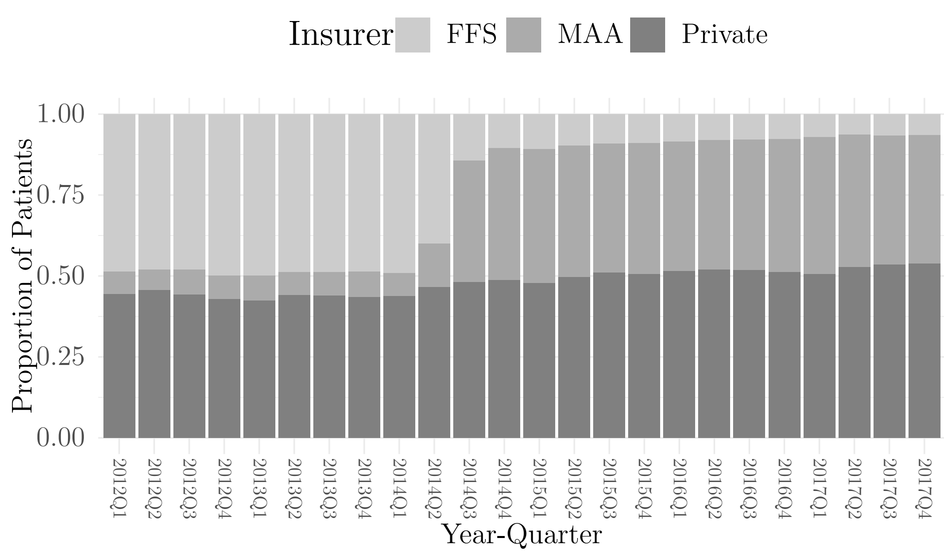
The Florida inpatient claims data contains the population of Florida inpatient stays over the sample period, i.e. 2012 through 2017. I limit the data to the respective procedure using the following processes.

Labor and Delivery

To identify labor admissions in each quarter, I 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, I 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, I limit the data to admissions for normal delivery or cesarean section paid for by Medicaid, Medicaid HMO, or private insurance. I 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 1.2 (Alker and Hoadley, 2013). However, as Figure 1.2 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.[26] 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.

Figure 1.2: 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.

[26]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

Orthopedic Surgery

The process to identify orthopedic surgery admissions is simpler. I 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`. I limit these data to Medicare fee-for-service patients. Summary statistics and additional information on these data are detailed in the main text.

1.9.4 Community Detection for Hospital Markets

Community detection 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, I 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 community detection. By multiplying the bipartite matrix by its transpose, I 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 (CD) algorithms then use this unipartite matrix to identify the markets based on common hospitals between zip codes. Using the same unipartite matrix, I run multiple CD algorithms, but focus on one specific market definition for the main results.

1.9.5 Labor and Delivery Sensitivity Analyses

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

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 1.11 presents the results corresponding to equation 1.1, 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 1.11 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.

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

Table 1.11: 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 1.1, i.e. the specification without instrumenting for percentile rank. Each column includes all of the same covariates as Table 1.4, 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

distance to the closest hospital in the choice set. Table 1.12 presents these results, which replaces of 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 1.12: 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 1.3, 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 1.4. 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

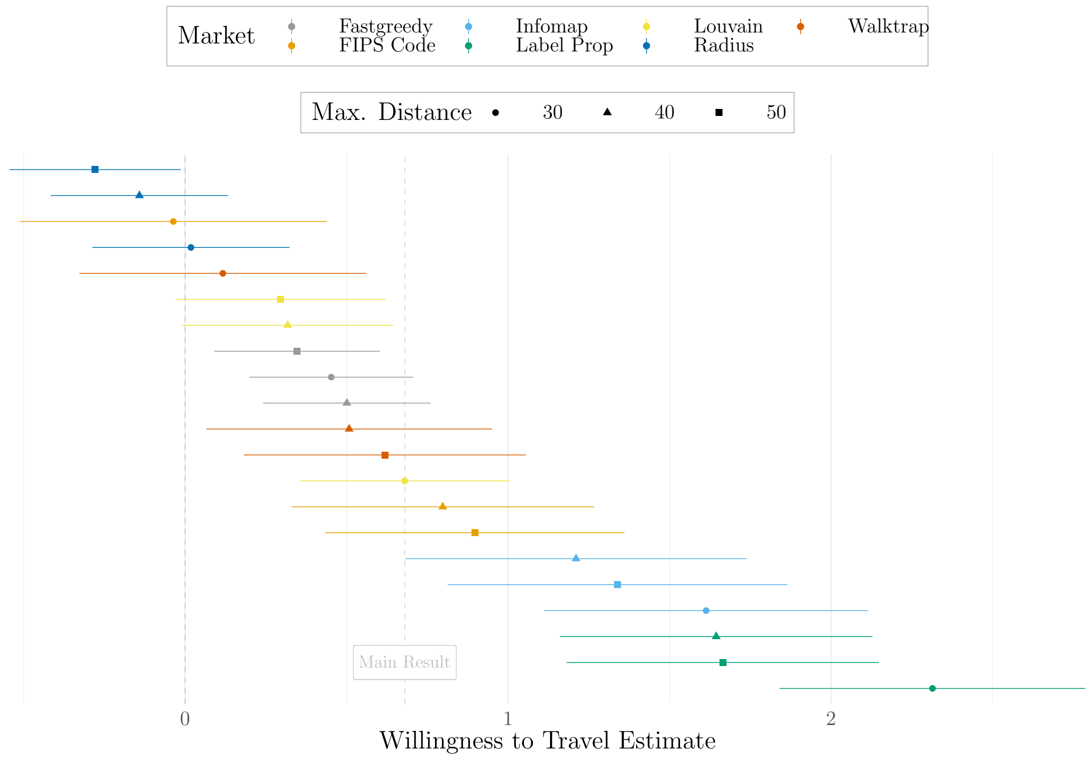
Sensitivity to Market Definition

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 1.3 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 market definition is relatively arbitrary and may not necessarily reflect the set of hospitals from which one is likely to choose. Overall, the estimates in Figure 1.3 demonstrate that the results are robust to various market definitions, with the arguably reasonable caveat that the markets must reflect patient flows.[27]

[27]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 1.3: 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.

1.9.6 Orthopedic Surgery Sensitivity Analyses

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

Table 1.13: 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 1.1, i.e. the specification without instrumenting for percentile rank. Each column includes all of the same covariates as Table 1.7, 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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 1.13, however, presents the results corresponding to equation 1.1, i.e. the specification that does not include the first stage residuals in the estimation. The coefficients on the percentile rank rank are smaller than that of the main specification, resulting in smaller willingness to travel estimates.

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. I 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 1.14 and find effects that are largely similar to the main results, particularly in the most saturated specification.

Table 1.14: 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 1.3, 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 1.7. 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 1.15 presents these estimates where I 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 1.15: 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 1.3, 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 1.7. 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

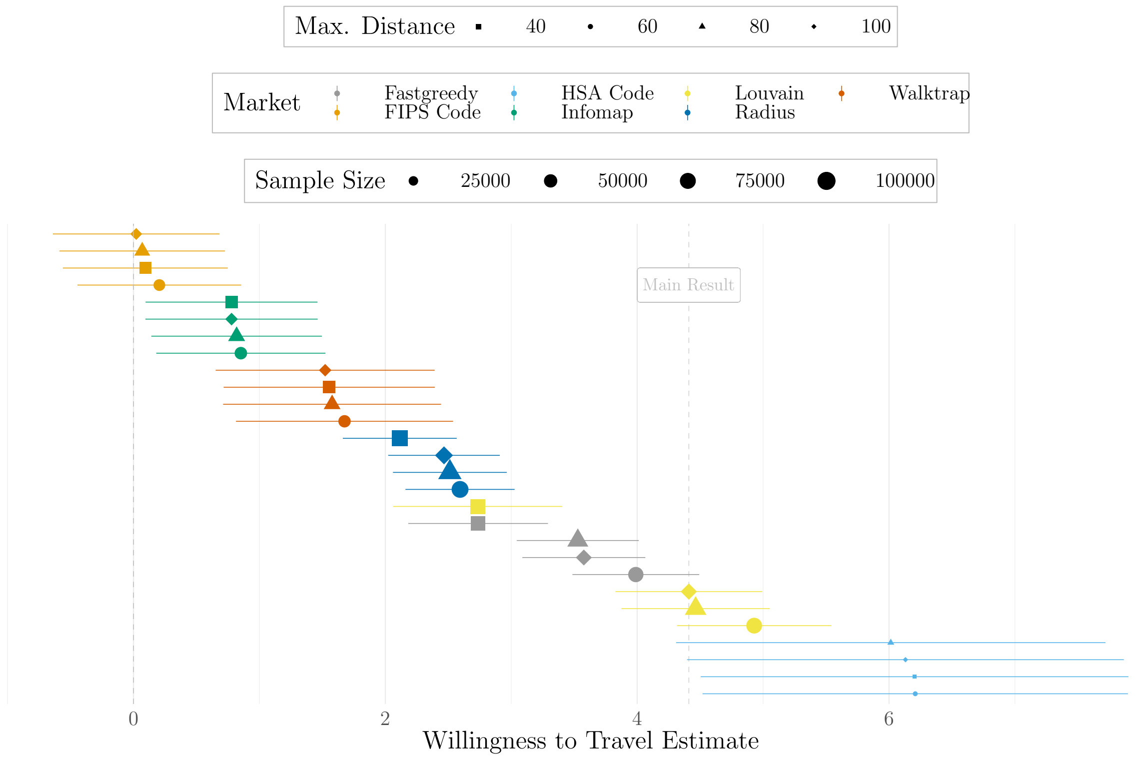
Sensitivity to Market Definition

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 1.4 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, I am left with a sample 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.[28] This consists of only four markets, namely Jacksonville, Tampa, 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.

[28]Note that the HRR Code market definitions are based on tertiary care, which is likely not reflective of the referral patterns use 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 existing market definitions, HSA codes are theoretically better-suited for this analysis.

Figure 1.4: 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.

Chapter 2

Online Reviews and Hospital Prices

User-generated review forums provide a novel source of quality information on businesses across many industries. In the context of hospital care, online reviews may be particularly valuable as an accessible source of quality information on an otherwise complicated product for which quality is multifaceted and difficult to measure. If reviews are informative to insurers or health care decision makers, then we would expect a corresponding effect on price. Using hospital Yelp reviews and Healthcare Cost Report Information System data, we employ an instrumental variable strategy that exploits plausibly exogenous variation in the Yelp algorithm to identify the effect of higher ratings on hospital prices. Our analysis consistently shows that higher user ratings tend to increase hospital prices, albeit with relatively modest magnitudes.

2.1 Introduction

Information problems are characteristic of many markets, especially those with complicated products. In some cases, brands can directly signal product quality to potential customers; in others, certification or mandatory disclosure from regulatory agencies provides third party information on quality (Dranove and Jin, 2010). More recently, online user reviews have emerged as a popular and accessible source of information, and consumers can now easily incorporate an *ex-post* quality assessment from a mass of consumers into their purchasing decisions.[1]

Given the increased availability of quality information via online reviews, a natural question is whether these reviews affect other product characteristics, particularly prices. Such a relationship is reflected, for example, in the case of certified goods, which have higher market prices compared to those that are not certified (Wimmer and Chezum, 2003; Dewan and Hsu, 2004). Higher user ratings have also been found to affect prices in a variety of other settings (Chevalier and Mayzlin, 2006; Resnick et al., 2006; Cabral and Hortacsu, 2010; Luca, 2016).

In this paper, we study the effect of online reviews on prices in the hospital market—a market characterized both by high prices and significant barriers to quality information. We have in mind a theoretical structure in which higher online reviews first affect demand for care, whether it be through the patient themselves or some other decision maker. Such a demand effect has been shown in the case of physician choice (Chen, 2018), as well as in the case of expert opinions (Reinstein and Snyder, 2005). A similar relationship has been

[1]Based on a 2016 survey from the Pew Research Center, 83 percent of U.S. adults reported that they at least sometimes consult online reviews before a first-time purchase, and 40 percent say that they always or almost always use online reviews in their decision-making processes. See https://www.pewresearch.org/internet/2016/12/19/online-reviews/ for further details.

found with regard to hospital demand and overall hospital quality measures (Chandra et al., 2016). In a bilateral negotiation between insurers and hospitals, it follows that any influence of online reviews on demand should also influence hospital pricing. Further, online reviews may provide additional information directly to insurers, which may also translate into higher negotiated prices. We discuss our theoretical motivation in more detail in Section 2.2.

Our empirical analysis is based on the universe of hospitals on the web-based rating platform, Yelp. This is a compelling empirical setting as the number of hospitals on this platform and the quantity of reviews per institution grew substantially over the past decade. By the end of our sample period (2012 – 2017), around 50 percent of general acute care hospitals were represented on Yelp, which facilitates the analysis of the price response for a wide range of hospitals. Using the data from each hospital's profile, we are able to construct the aggregate rating that a visitor to the site would have seen at a given point in time, and we pair these ratings with estimated hospital prices from the Healthcare Cost Report Information System (HCRIS).[2]

The Yelp platform provides a convincing source of exogenous variation from which we can identify the effect of ratings on prices. Specifically, the star rating presented on Yelp is an average of the prior user reviews, rounded to the nearest half-star. We exploit this rounding to construct an instrument for a hospital's reported star rating in each year. Using this instrumental variable strategy, we find significant increases in price as a result of having a sufficiently high rating on this platform. In our most saturated specification, we find that

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

in comparison to hospitals with fewer than 3 stars, more favorably rated hospitals negotiate higher prices in the following year. We also find a positive relationship between having no rating and price, indicating that there may exist a premium for the absence of negative information on this platform. The magnitudes of our estimates are relatively small across the majority of our specifications but reasonable given our empirical setting and in context of the existing literature. Our findings are robust to a variety of empirical concerns, including violations of the exclusion restriction and the presence of price outliers. Our results are also robust to a bevy of alternative specifications, including hospital fixed effects and controls for other measures of hospital quality from Hospital Compare. Finally, we consider the potential demand response to quality information in Section 2.6; although our data in this area are more limited, our results suggest that patients are responsive to higher quality ratings in their hospital choice, supporting a demand response as a potential underlying mechanism.

Our analysis contributes directly to the literature on information transparency in health care.[3] Much of this literature investigates quality disclosure in the context of report cards, noting that providers or insurers with higher reported quality experience increases in market share and that report cards augment consumers' existing knowledge (Dafny and Dranove, 2008; Zhe Jin and Sorensen, 2006). Conversely, Dranove and Sfekas (2008) find little change in market shares based on New York's cardiovascular surgery report cards, which may have been due to difficulty in understanding the information or a lack of novel information to patients.

Despite the natural link between demand effects and negotiated prices, there are no studies

[3] 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.

of which we are aware that directly examine the effect of quality information on prices for health care services. One likely explanation for this gap in the literature is the inherent difficulty and complexity in defining and measuring hospital quality (Dranove and Jin, 2010; Romley and Goldman, 2011). While there have been several regulatory attempts to improve hospital quality disclosure, many of these attempts have failed to gain substantial traction due to legislative delays and other practical barriers.[4] As a result, patients and other health care decision makers continue to face significant barriers when attempting to incorporate timely quality data into their hospital decision-making criteria, and the fragmented and delayed nature of existing hospital quality information would tend to depress any estimated effects of such measures on patient demand or price (Ranard et al., 2016).

Online user review platforms help to fill this informational gap in at least three ways. First, the information is more easily accessible. This is evidenced by the popularity of Yelp and similar online review information. For example, by Fall 2020, Yelp reported 32 million visits to its platform from unique devices.[5] Second, quality measures from online user reviews are available essentially in real-time. This is in contrast to all other large-scale hospital quality measures of which we are aware, which tend to capture quality based on data collected over two to three years prior. Third, the information from user reviews appears to be valuable, with several studies now showing that online user reviews are positively related to clinical quality measures (Bardach et al., 2013; Trzeciak et al., 2016; Howard and Feyman, 2017;

[4] For example, the Centers for Medicare and Medicaid Services (CMS) proposed an aggregate star rating as part of their "Hospital Compare" initiative in 2015, but subsequent updates to the ratings stalled in response to industry concerns (American Hospital Association, 2016, 2017). CMS recently announced that they would not update the hospital star ratings for 2021. Additionally, Hospital Compare and other measures such as physician report cards may provide a fragmented measure of hospital quality and may not capture relevant aspects of patients' overall experience of care.

[5] See https://www.yelp-press.com/company/fast-facts/default.aspx

Trzeciak et al., 2017). Further, there exists legitimately new information in online reviews that is not fully captured in existing hospital quality measures such as Hospital Compare (Ranard et al., 2016; Campbell and Li, 2018; Perez and Freedman, 2018).[6]

With more accessible information, such as that from online user reviews, it becomes more feasible to detect various effects of quality disclosure. For example, work from Scanlon et al. (2002) and Pope (2009) shows that people are more responsive to *overall* ratings, as opposed to more granular measures. Dranove and Sfekas (2008) similarly find that consumer satisfaction scores are the driving factor behind Medicare beneficiaries' responses to plan report cards. More recently, Chen (2018) finds that online reviews affect patient choices when selecting a physician. These studies provide compelling empirical evidence that patients are responsive to accessible, aggregate measures of quality and are particularly responsive to measures that capture patient satisfaction.

In examining the effects of ratings on prices, our study also contributes to the literature on hospital pricing. Much of this literature 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 Dranove, 2003). Our contribution to this literature is to highlight another likely avenue by which hospitals can increase 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 affordability of health care services is critical for future policy

[6] Dellarocas (2003) explains how web-based reviews provide "word-of-mouth" information but from a broader network than traditional "word-of-mouth" communication. We should therefore expect online reviews reflect some combination of new and existing quality information.

in this area.

2.2 Theoretical Motivation

To provide a formal theoretical framework in which to analyze the relationship between information and prices, 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

$$p_{ij} = \arg\max_{p_{ij}} \left(\Delta\pi_{ij}^{H}\right)^{b_{ij}} \times \left(\Delta\pi_{ij}^{M}\right)^{1-b_{ij}}. \tag{2.1}$$

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

$$\pi_i^{H}(\boldsymbol{p}, \boldsymbol{\theta}) = \sum_n D_{in}^{H}(p_{in} - c_i), \text{ and} \tag{2.2}$$

$$\pi_j^{M}(\boldsymbol{p}, \boldsymbol{\theta}) = D_j^{M}(\theta_j - \eta_j) - \sum_h D_{hj}^{H} p_{hj}. \tag{2.3}$$

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 as follows:

$$p_{ij}^* D_{ij}^H = b_{ij}\left[\Delta D_j^M(\theta_j - \eta_j) - \sum_{h\neq i} p_{hj}^* \Delta D_{hj}^H\right] + (1 - b_{ij})\left[c_i D_{ij}^H - \sum_{n\neq j} \Delta D_{in}^H (p_{in}^* - c_i)\right]. \quad (2.4)$$

The first term on the right side of equation 2.4 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?

We envision online reviews entering this framework and affecting price in at least two possible ways. First, consider the health care decision maker, which we take to include patients, their family and friends, their doctors, and anyone else who influences hospital selection. If a higher quality rating increases the probability of selecting that hospital, then the insurer's willingness to pay to keep that hospital in their network also increases. This improves the hospital's bargaining position, which enables them to negotiate higher prices. Second, online reviews may provide a novel measure of the patient perception of quality, as suggested by Ranard et al. (2016), Campbell and Li (2018), and Perez and Freedman (2018). To the extent that insurers directly value quality and perceive user reviews as a source of new information, then user reviews may again increase the insurer's willingness to pay and ultimately increase price. This mechanism would be akin to pay for performance programs seen in Medicare, in which insurers directly incorporate quality measures into the final payment.[7] We investigate the demand response mechanism empirically in Section 2.6.

The term of a negotiated payment contract between a hospital and insurer is typically three to five years, 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."[8] 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.

Ultimately, the extent to which prices respond to changes in user reviews is an empirical question. We also suspect heterogeneous effects of reviews on prices depending on the relative ranking of quality information. In particular, a marginal increase in ratings may affect

[7]See https://www.forbes.com/sites/brucejapsen/2017/02/02/unitedhealth-aetna-anthem-near-50-value-based-care-spending and https://www.forbes.com/sites/brucejapsen/2017/08/17/employers-accelerate-move-to-value-based-care-in-2018 for further discussion of value-based care structures in private insurance markets.

[8]See Cooper et al. (2019) for additional discussion regarding hospital/insurer contract types.

hospitals at the low end of the rating distribution differently than those at the high end of the rating distribution.

2.3 Data

Our analysis relies on two main sources of data: the population of hospital reviews on Yelp and the Healthcare Cost Report Information System (HCRIS) maintained by the Centers for Medicare and Medicaid Services (CMS). We supplement these data with county-level information from the Area Health Resource Files and hospital-level data from the American Hospital Association (AHA) Annual Surveys.

2.3.1 Yelp

Of the various platforms for user-generated content, Yelp was one of the first to become a popular source of information. The site was launched in 2004 and rapidly began to amass reviews and visitors to the site.[9] Just three years after it launched, Yelp reached one million reviews.[10] Yelp's success piqued the interest of Google, who offered to acquire the company for $550 million in December 2009; however, after Yelp declined this offer, Google entered the online review business themselves.[11] While Yelp continues to be an extremely popular website, it has since been usurped by Google, which is now the most widely used review site.[12] Nonetheless, over our sample period, Yelp was the most prominent source of user-generated review content and is a well-suited setting for a study of online hospital reviews.

[9] See https://www.eater.com/2014/8/5/6177213/yelp-turns-10-from-startup-to-online-review-dominance.

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

[11] See https://www.nytimes.com/2017/07/01/technology/yelp-google-european-union-antitrust.html.

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

Business profiles on Yelp can be created by a consumer, the business owner, or Yelp itself. Once a business profile exists, users can create a profile, submit unstructured comments, and assign a star rating between 1 and 5. People can explore the reviews by searching for a business in a search engine (e.g., Google) or directly on the site. Users first encounter a summary of the business, including the number of reviews, summary rating, business location, and contact information. Visitors to the site can click on this summary page to access the full profile, which contains each individual review with its respective star rating. The order and presence of the reviews are determined by Yelp's proprietary algorithm. For instance, reviews that the algorithm identifies as spam or potentially inauthentic are retrievable under the link "other reviews that are not currently recommended."[13] These reviews are not factored into the aggregate rating and are not included in our analysis.

Using the hospital name and address from the AHA data, we collected all hospital profiles available on the platform. We obtained each review, including the text, star rating, date, and user ID corresponding to each review on a given hospital profile. Using these individual reviews, we are able to construct aggregate summary ratings at a given point in time. We use year-end observed rating as the basis for our analysis. Details on the data collection and cleaning processes are explained in 2.8.1.

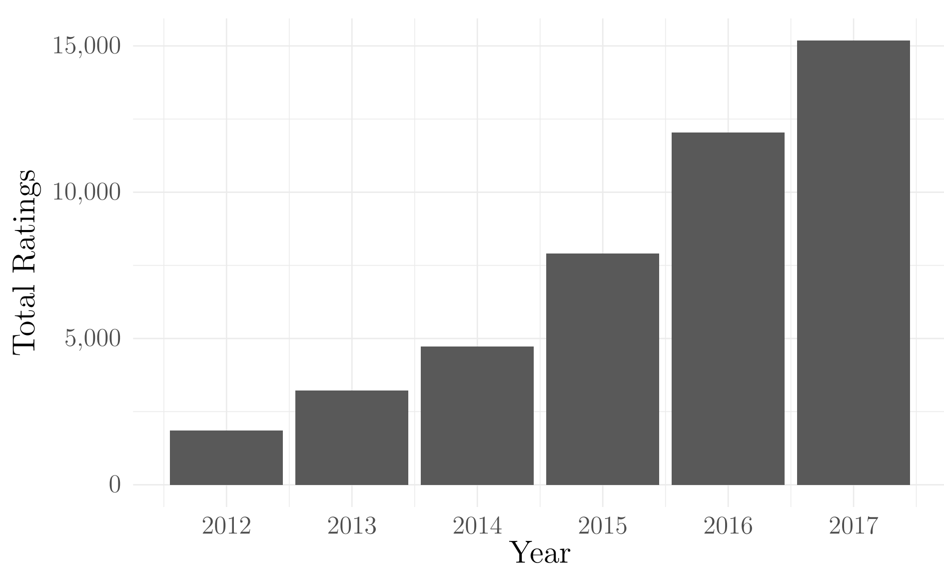
Note that Yelp does not simply display the cumulative average of the star ratings a hospital receives but instead rounds the average star rating to the nearest half star between one and five. For example, if a hospital has a total of three reviews with 1, 2, and 5 stars, the cumulative average is 2.67. A visit to the site would instead see the rounded rating (down in

[13] Luca and Zervas (2016) finds that fraudulent reviews on the platform are generally caught by this algorithm.

this case) of 2.5. The star rating that the user would see based on the rounded cumulative average is what we refer to as the observed rating. Importantly for our empirical analysis, the observed ratings are relatively stable over the course of several months. For example, the median hospital received two new reviews per year in 2012 and four new reviews per year in 2017.

We are able to collect reviews since the platform was first launched in the mid-2000s; however, the representation of hospitals was relatively low in the early years of the site. Until 2012, less than 25 percent of hospitals in our sample had a Yelp presence. As such, we exclude observations prior to 2012 in order to focus on the years in which the platform was sufficiently mature. The presence of hospitals on Yelp grew steadily through 2017, the end of our sample period, at which point 49 percent of the hospitals in our sample had a Yelp presence. Figure 2.1 shows the cumulative number of hospital reviews on the platform by year.

Figure 2.1: Total Hospital Reviews on Yelp by Year



Notes: The bars represent the cumulative number of hospital reviews by year, with over 15,000 reviews by the end of our sample period.

The growing number of reviews and hospitals on the platform indicates that people are engaging with this forum by sharing their experiences, but it does not indicate whether people actually consult reviews when selecting a hospital. We investigate this by analyzing the review text, where we find that up to 10% of the reviews in our sample explicitly indicate that the reviewer considered online reviews in selecting a hospital.[14] This supports the claim that online reviews are relevant to health care decision makers and likely reflects a lower bound of the proportion of patients that use this information since patients may consult the platform without mentioning it in their reviews or consult the platform and ultimately not leave a review.

2.3.2 Healthcare Cost Report Information System (HCRIS)

CMS maintains the HCRIS data and requires that all Medicare-certified hospitals submit a cost report annually. Borrowing from Dafny (2009), Schmitt (2018), and Lin et al. (2020), we use HCRIS data to construct 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. We eliminated observations with price outliers at the 5th and 95th percentiles, and we deflated 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.[15] Specifics on the variables used to construct the price measure are detailed in 2.8.2.[16]

[14] Refer to 2.8.1 for more details.

[15] To construct our price measure, we are able to back out Medicare payments, but the data do not enable us to remove Medicaid payments.

[16] 2.8.4 provides additional analysis regarding the robustness of our results to the presence of outliers. We find that our qualitative results are not sensitive to the presence of price outliers.

Throughout our analysis, our outcome of interest is log price in the following period, i.e. $\ln(price_{t+1})$. For simplicity, we refer to this variable as "price" throughout. Additionally, note that the measure of time in our analysis is calendar year, but that the cost reports are compiled by hospital fiscal year. This 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.[17] We examine the sensitivity of our results to these timing considerations in 2.8.5, in addition to falsification tests when considering prices at earlier time periods. The results of that analysis are consistent with our initial findings.

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 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 ratingsn. 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.

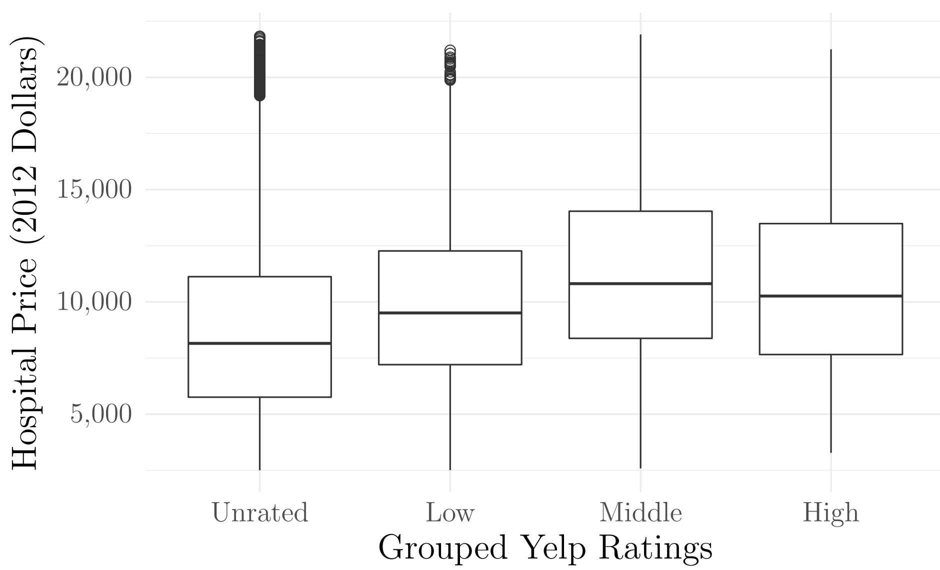
Our delineation of the star rating groups follows from the natural cut-points observed in the data. Given the observed average rating of 2.9, a middle rated hospital falls in a group slightly above the average (i.e. 3 or 3.5 star rating), and a high rated hospital is at least a standard deviation above the average (i.e. 4, 4.5, or 5 stars).[18] Figure 2.2 shows hospital

[17]For example, consider a hospital with fiscal year ending 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 $t+1$.

[18]2.8.4 provides further discussion of our rating group choices along with estimation results of our main specification modified to include more granular rating groups. Additionally, Section 2.5.3 includes results

price in relation to the rating groups, which indicates that the average price among the unrated hospitals is lower than that of the rated hospitals. Further, the average prices of the high and middle rated hospitals, which are \$10,832 and \$11,323, respectively, are higher than the average of the low rated hospitals, \$9,982.

Figure 2.2: Hospital Price by Rating Group



NOTES: Hospital prices are deflated to 2012 dollars. High rated hospitals 4, 4.5, or 5 stars, middle rated have 3 or 3.5 stars, and low rated rated have 2.5 stars or below. The unrated hospitals do not have Yelp profiles.

2.3.3 Additional Data Sources

In addition to the data from Yelp and HCRIS, we incorporate county-level characteristics from the Area Health Resource Files. These variables include population, unemployment and poverty rates, rate of uninsured, and median income. Additionally, we include hospital quality information from CMS's Hospital Compare data. These measures consist of readmission and mortality rates for heart failure, pneumonia, and acute myocardial infarction

for different thresholds for middle and high. Using these thresholds, the results are less precise but have commensurate point estimates.

(AMI). We also use the AHA Annual Survey data, which provides various hospital characteristics. 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 Table 2.1. 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.

2.4 Empirical Approach

We are interested in estimating the effect of consumer reviews on hospital prices. Central to our research design is a feature of Yelp's rating algorithm, wherein a continuous underlying score is rounded to the nearest half-star increment. Hospitals with ratings just above or just below any half-star threshold are therefore of comparable underlying quality (as measured by the continuous score) but ultimately received two different star ratings on the website.

Given this feature, a regression discontinuity (RD) design is seemingly a natural starting point, but several features of our application deem an RD design inappropriate in this case. 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, demonstrated in Figures 2.3a and 2.3b, is particularly problematic as it suggests a violation of the continuity assumption required for the RD design. This contrasts with studies that use Yelp reviews in other settings, such

Table 2.1: 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.).

as restaurants, where there are more businesses, more reviews, and the outcome of interest can be measured more frequently (Anderson and Magruder, 2012). We instead exploit the plausibly exogenous change in ratings due to rounding as an instrument for the observed rating category in a two-stage least squares estimator.[19] We employ the following regression specification, where Equation 5a is the second stage and Equations 5b and 5c are the first stage:

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

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

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

where $High_{it}$ 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.[20] 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.[21] 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.[22] In the first stage Equations

[19]Results when employing an RD design are imprecise but qualitatively similar to our preferred specification outlined in this section. Those results are omitted for brevity but available on request.

[20]We do not include the continuous score in our primary specification, but we do include it in Section 2.5.3 were we add the continuous score, with no change in our results.

[21]Section 2.5.3 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.

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

5b and 5c, RH_{it} and RM_{it} are indicators for being rounded up into a high or a middle rating, respectively. We discuss details of the instrument construction in the following Section 2.4.1.

Each new review that a hospital receives presents the possibility that it will move into a different observed rating and possibly be rounded up into that rating. As such, ratings may fluctuate if hospitals receive numerous reviews but will be relatively stable if reviews are infrequent. In the hospital context, we find the latter. Over the study period, the median number of new reviews in a given year is 3. This implies an extended period of time over which decision makers are potentially "exposed" to the exogenous rounding, thereby allowing an opportunity for changes in online reviews to disseminate to health care decision makers and ultimately affect hospital prices.

There are two points to clarify about our outcome of interest. The first is that we focus on the price in the following year to allow for a lagged response to rating information. Regardless of the mechanism by which ratings impact price, they would manifest through negotiations between hospitals and insurers. Thus, we construct our estimating equation to allow for such a lag. Additionally, we use the natural log of price in all of the results presented in this analysis. We opt for log prices due to the wide range in the outcome variable, as shown in Figure 2.2. We find qualitatively similar results when estimating this equation in levels rather than logs.

Before estimating Equation 5a, we detail the construction of our instrument and investigate its appropriateness in our application. Details of this analysis are presented throughout the remainder of this section. We reserve formal first stage and reduced form analyses for our main results in Section 2.5. Additionally, Section 2.5.3 assesses the sensitivity of our main results to a bevy of potential concerns.

2.4.1 Instrument Construction

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.[23] Section 2.5.3 presents alternative results with a bandwidth of 0.10; the results are commensurate with our main estimates, with a higher magnitude on the high rating group. Given the fact that a 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.[24]

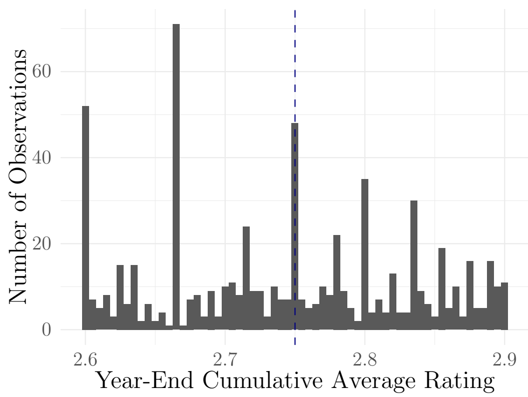
Figures 2.3a and 2.3b show the distribution of the average rating around the middle and high thresholds, respectively. Figure 2.3a contains some spikes but appears to be more uniformly distributed than Figure 2.3b. The distribution in Figure 2.3b is particularly dense at 3.67 and 3.75, which is likely because a hospital can have an aggregate rating of 3.67 with just

[23]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%.

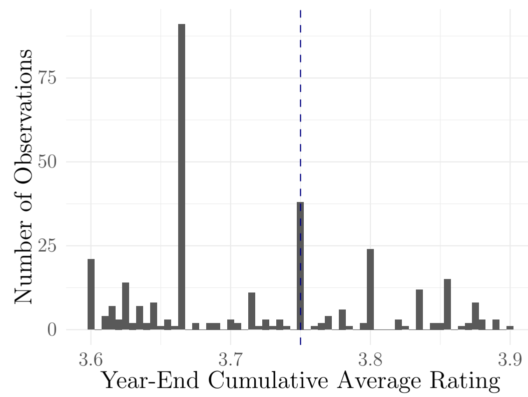
[24]We nonetheless examine the sensitivity of our results to potential violations of the exclusion restriction in 2.8.4, with little qualitative change in our findings.

3 total reviews and 3.75 with just 4 reviews. The uniformity of the distribution shown in Figure 2.3a 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 2.3: 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.

2.4.2 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.[25] 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.[26]

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.[27] The software takes into account a variety of aspects about both the reviewer and the review content when determining whether or not to recommend a review.[28] 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

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

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

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

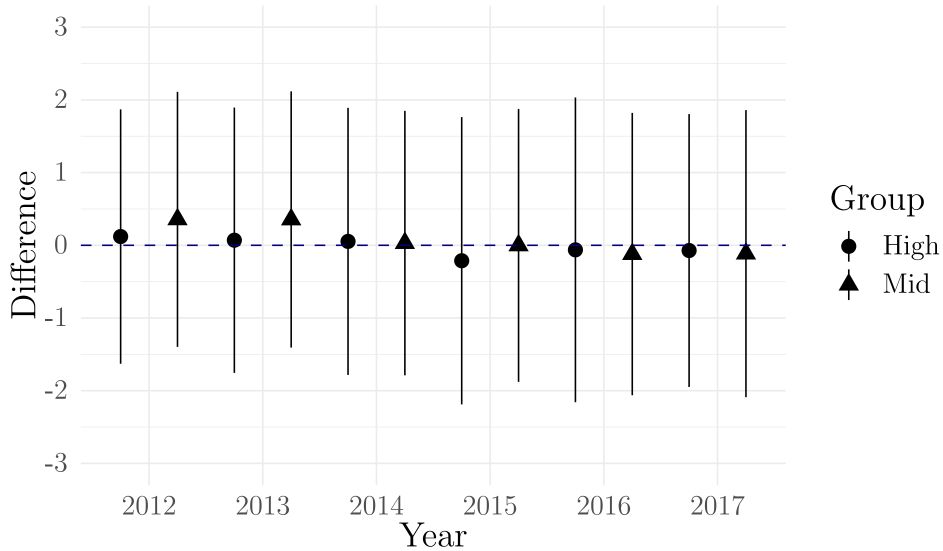
[28] This source outlines Yelp's practices regarding review recommendation: https://www.yelp-support.com/article/Why-would-a-review-not-be-recommended.

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 2.4, 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

Figure 2.4: Difference between Last Rating and Prior Cumulative Average for Rounded Hospitals



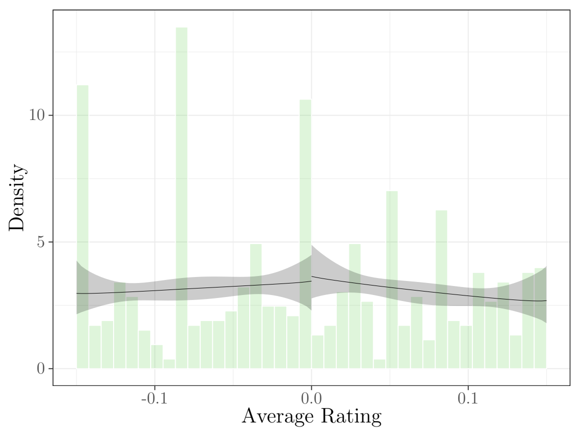
NOTES: The points represent the mean value for the difference, and the interval lines represent one standard deviation.

et al. (2018). The results of the density tests are reflected in Figures 2.5a and 2.5b. 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, in 2.8.3, we consider the same density test when requiring a minimum of five reviews to be considered a "rated" hospital, in which case we find no statistical evidence of sorting at either threshold.[29]

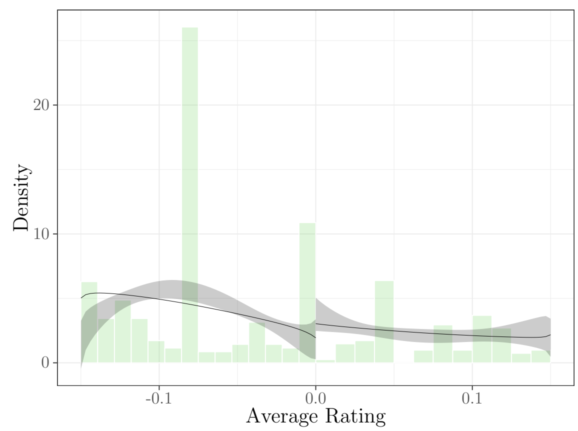
The data and institutional details of the Yelp platform therefore suggest that hospitals

[29]In Section 2.5.3, 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 2.5: Manipulation Tests at the Middle and High Thresholds



(a) Middle Threshold



(b) High Threshold

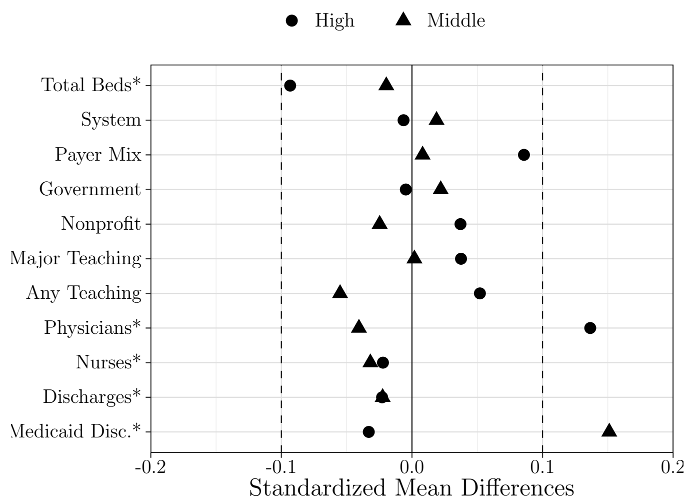
NOTES: The graph depicts the density tests presented in Cattaneo et al. (2018). The x-axis shows the average rating for a hospital at year-end. The density estimates are on the y-axis. The light green bars show the histogram of the average ratings. The bandwidth is 0.15. At the middle threshold there 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, as shown in Section 2.5.3, 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.

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 characteristics that also affect prices.

2.4.3 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 2.6 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 2.6: 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 those 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 2.1. The symbol * denotes variables that are shown per capita.

2.5 Results

We begin by discussing the results for the first stage and "reduced form," followed by the presentation of our main IV results. We then provide an overview of alternative specifications and robustness.

2.5.1 First Stage and Reduced Form Results

Tables 2.2 and 2.3 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

Table 2.2: 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

effects.[30] 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 2.2, which are based on Equations 5b and 5c. The first panel in Table 2.2 summarizes the relationship between the high rating and being rounded into high and middle, followed by

[30] 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 2.3: Reduced Form Results of Rounding Instrument

	(1)	(2)	(3)	(4)
Price				
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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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 present the reduced form results—i.e., the regression of price on the instruments—in Table 2.3. 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 2.3, 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.

2.5.2 Main Results

Our primary IV results are presented in Table 2.4. 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 experience a half-star higher rating due to the rounding mechanism. Even in our most saturated specification (column 4), we see that relative to low-rated hospitals, the average price for an inpatient stay is higher the following year at a high rated hospital. 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

Table 2.4: Main Results

	(1)	(2)	(3)	(4)
Price				
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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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).

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) find a 3 – 5 percent increase in hospital prices after vertical integration. Lewis and Pflum (2017) find 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 Dranove, 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.

2.5.3 Specification Tests and Sensitivity Analyses

We consider the sensitivity of our estimates in Table 2.4 to several potential concerns. First, we note that we are able to precisely reject underidentification and reject the null hypothesis of weak instruments.[31] 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. 2.8.4 details this analysis. 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). We detail this analysis in 2.8.4.

Several falsification tests also lend confidence to our results. These are detailed in 2.8.5, where we consider future clinical quality measures and lagged prices as outcomes. From that analysis, we find no evidence of a relationship between those variables and our rating categories, which 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.

Finally, we expand on our main results—namely column (4) of Table 2.4—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

[31] The Kleibergen-Paap LM statistic shown in Table 2.4 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.

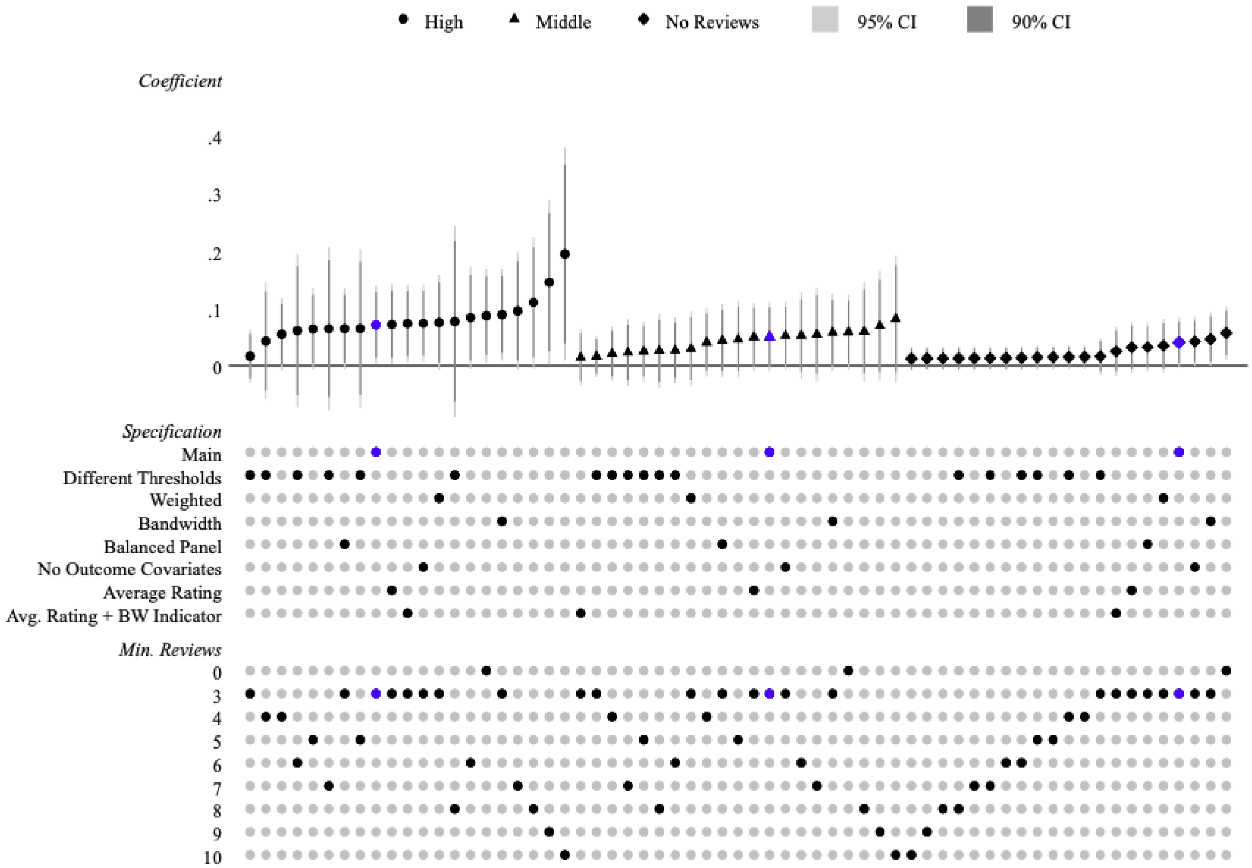
"rated" hospital, and different threshold values for defining middle and high rated hospitals. Figure 2.7 presents the coefficient estimates on the high and middle rating groups, along with the "no reviews" variable for these specifications. They are shown in tandem with our main results, which are presented 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 2.7.

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 2.7, 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

Figure 2.7: 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.

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.

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 present another set of results in 2.8.4 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. Lastly, we include the "Bandwidth" specification, which changes the 0.15 bandwidth used through the remainder of the paper to 0.1. The results are robust to this change, finding slightly higher, statistically significant point estimates for each coefficient of interest.

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.

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 2.7, 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.

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 2.7, 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.

2.6 Potential Mechanisms

The results in Section 2.5 and the related sensitivity analyses provide consistent evidence that higher ratings translate into higher hospital prices. In this section, we build on these results and use AHA and HCRIS data to investigate the underlying mechanism. We indirectly test the demand mechanism using a partitioned regression. Then we build upon that analysis by first examining the review data to understand engagement on the platform and then by directly testing the demand mechanism using a discrete choice model. These analyses provide insights on the potential mechanisms underlying our main results.

2.6.1 Partitioned Regression Analysis

For online reviews to affect prices through demand, patients or insurers must have some choice among hospitals in the market. Without any such choice, there is no difference in the outside option for the hospital in response to a change in information. A demand response should therefore depend on the level of competition in the market, and as such, we would expect heterogeneous effects across hospitals based on the competitiveness of their respective markets. For instance, a hospital that is one of very few hospitals in its referral region should experience little change in its bargaining position in response to a high rating compared to a hospital that faces numerous competitors.

To investigate, we use the specification shown in column four of Table 2.4 (i.e., the regression of price on ratings with year, county, and hospital fixed effects, controlling for county and hospital characteristics and hospital quality measures) but with the sample partitioned into quartiles based on the number of hospitals in the hospital referral region (HRR). The results for this analysis are shown in Table 2.5. Focusing on quartiles 3 and 4, we see that our results are driven by hospitals with at least 13 hospitals in their respective HRR in a given year. This is consistent with the underlying economic theory which posits that for online reviews to matter, physicians and patients must have some hospitals from which to choose. The final column shows the results when we estimate using observations in either the third or fourth quartile (the above-median markets). As we would expect, we continue to see that our main results are driven by the markets where information would be plausibly more relevant in the negotiation process.

2.6.2 Discrete Choice Analysis

We can further explore how online reviews allow for higher prices by directly estimating the effect of online reviews on hospital selection. For such an effect to exist, decision makers must incorporate online reviews into their decision making process. A closer examination of

Table 2.5: Partitioned Regression by Hospitals in HRR

Quartiles	1	2	3	4	3 & 4
Number of Hosps. in HRR	1-7	8-13	13-24	24-65	13-65
Price					
High Rating	0.0145	-0.0160	0.111**	0.135*	0.112**
	(0.0505)	(0.0567)	(0.0544)	(0.0746)	(0.0475)
Middle Rating	0.0323	0.00216	0.0478	0.0990	0.0614
	(0.0481)	(0.0418)	(0.0577)	(0.0626)	(0.0458)
No Reviews	0.00422	0.0238	0.0146	0.104**	0.0491
	(0.0312)	(0.0412)	(0.0450)	(0.0504)	(0.0348)
Fewer than 3 Reviews	-0.00595	0.0257	-0.00151	0.0823**	0.0331
	(0.0304)	(0.0358)	(0.0397)	(0.0414)	(0.0294)
Observations	2263	2012	1769	1837	3661

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. Stars indicate the following: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

our Yelp data suggest that this may be the case. For example, below each review narrative, visitors to the site have the ability to react to a review using three separate buttons titled "useful," "funny," and "cool." These buttons also display the number of votes corresponding to each adjective. In our sample, 74% of the reviews were classified as useful. Among these reviews, there were an average of 4.9 votes for this adjective. Similarly, 34% were "funny" and 26% were "cool," with an average of 2.3 and 2.2 votes, respectively. These reactions suggest at least some level of engagement with the platform for the specific purpose of collecting information about hospitals.

We also analyze the text using natural language processing techniques to determine whether or not reviewers mention using the online reviews prior to their hospital choice. We find that between one and ten percent of Yelp reviewers explicitly state that they considered online reviews in selecting a hospital. 2.8.1 details this analysis. We consider this to be a lower bound since patients can consult reviews but choose not to leave a review of their own or choose not to mention their search process in their review. While this does not identify the extent to which reviews inform hospital choice, it provides further support to the plausibility of this mechanism.

Finally, we test the demand mechanism more explicitly using a discrete choice model where the Yelp reviewer's utility depends on the hospital's rating group, hospital characteristics, and clinical quality metrics. We model the utility for reviewer i from receiving care at hospital j as:

$$\begin{aligned} u_{ij} &= v_{ij} + \varepsilon_{ij} \\ &= \beta_1 High_j + \beta_2 Mid_j + X_j' \alpha + \varepsilon_{ij}, \end{aligned} \tag{2.6}$$

where v_{ijt} follows from the model specified in Equation 5a and includes indicators for a hospital's rating group at the time of reviewer i's review, hospital characteristics, and clinical quality measures. In contrast to Equation 5a, here we omit the "too few" rating category because, by construction, these are rarely selected in our data. The vector X_j contains hospital characteristics and clinical quality measures analogous to those included in our main results. Lastly, we assume that the error term, ϵ_{ij}, is i.i.d. Type I extreme value. This results in the common logit form for the probability of patient i selecting hospital j. We assume that the reviewers chose the hospital that they review, and because patients therefore only appear in these data if they choose a hospital, there is no outside option. We estimate the model using standard maximum likelihood.

In line with our empirical approach outlined in Section 2.4, we use the plausibly exogenous rounding on the platform to instrument for the rating groups using a control function approach (Petrin and Train, 2010). To implement, we first estimate the following equations,

$$High_j = \gamma^H_{RH} RH_j + \gamma^H_{RM} RM_j + X'_j \psi^H + \epsilon^H_j, \tag{2.7}$$

$$Mid_j = \gamma^M_{RH} RH_j + \gamma^M_{RM} RM_j + X'_j \psi^M + \epsilon^M_j, \tag{2.8}$$

where the hospital's rating group is a function of the instruments, RH and RM, and all of the right hand side variables included in Equation 2.6. We then include the residuals, $\widehat{\epsilon}^H_j$ and $\widehat{\epsilon}^M_j$ in the second stage. This enables us to recover consistent estimates of β_1 and β_1. The second stage equation is

$$u_{ij} = \beta_1 High_j + \beta_2 Mid_j + X'_j \alpha + \widetilde{\varepsilon}_{ij}, \tag{2.9}$$

where $\varepsilon_{ij} = \eta \epsilon_j + \widetilde{\varepsilon}_{ij}$, and $\widehat{\epsilon}^M_j$ and $\widehat{\epsilon}^H_j$ are estimates for ϵ^M_j and ϵ^H_j.

We define a reviewer's choice set in this model through community detection algorithms, which use patient flows to identify groups of hospitals that draw from common FIPS codes.[32] To implement, we approximate annual patient flows using CMS's Hospital Service Area Files, which contain the total annual inpatient hospital claims for fee-for-service Medicare patients by zip code, which we aggregate to the FIPS code level.

We limit the sample to choice sets where the reviewer has at least three hospitals on Yelp

[32] 2.8.6 provides additional detail on these methods. Please see https://github.com/graveja0/health-care-markets for an excellent resource that explains how to construct these markets.

in their respective market in order to focus on markets where Yelp is likely a viable source of information. Lastly, the hospital rating groups are based on the ratings that would have been available on Yelp at the time that the individual wrote her review. Given the limited flow of reviews for any given hospital, it is unlikely the hospital's aggregate rating changed between when the reviewer could have consulted the site to inform her hospital choice and when she left the review.

Table 2.6: Discrete Choice Model Results

	(1)	(2)
High Rating	0.163**	0.155*
	(0.081)	(0.083)
Middle Rating	0.346***	0.284***
	(0.054)	(0.055)
Residuals (High)	-0.517***	-0.554***
	(0.098)	(0.100)
Residuals (Mid)	-0.424***	-0.357***
	(0.060)	(0.060)
Total Beds	-0.001***	-0.001***
	(0.000)	(0.000)
Non-Profit	-0.168***	-0.193***
	(0.032)	(0.033)
System	0.013	0.044
	(0.034)	(0.035)
Major Teaching Hospital	0.184***	0.227***
	(0.035)	(0.036)
Any Teaching Hospital	0.228***	0.219***
	(0.028)	(0.029)
Case Mix Index	0.211***	0.137**
	(0.055)	(0.056)
30-Day Mortality (Heart Failure)		-0.039***
		(0.009)
30-Day Readmission Rate (Heart Failure)		-0.016**
		(0.008)
30-Day Mortality (Pneumonia)		0.060***
		(0.006)
30-Day Readmission Rate (Pneumonia)		0.003
		(0.009)
30-Day Mortality (AMI)		-0.087***
		(0.009)
30-Day Readmission Rate (AMI)		-0.068***
		(0.012)
Difference in Predicted Choice Probability		
High v. Low	0.0114**	0.0115**
	[0.00047, 0.02258]	[0.00052, 0.02262]
Middle v. Low	0.0239***	0.0238***
	[0.01828, 0.0284]	[0.01822, 0.02826]
Num.Obs.	13530	13530

Notes: The first section of this table presents the coefficient results of the model. The second section presents the differences in predicted choice probability between the high and middle groups and the excluded group. The confidence intervals are based on 250 bootstrap replications. Statistical significance is indicated as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.6 presents the results. The coefficient estimates indicate that utility is increasing in high and middle ratings. Note that the high rating estimates are less precise than those for the middle group, which is not surprising given that high-rated hospitals appear less frequently in reviewers choice sets. The results are robust to the inclusion of various clinical quality measures. Using these results, we calculate the difference in predicted choice probability for a high or middle rated hospital compared to a low rated hospital. The bottom of Table 2.6 presents these results, which shows that in comparison to a low-rated hospital, a middle rating spurs a 2.4 percentage point increase in choice probability and a high rating results in a 1.1 percentage point increase. This is based on a mean predicted choice probability of 0.1499 for a low rated hospital, 0.1737 for a middle rated hospital, and 0.1613 for a high rated hospital.[33] We calculate confidence intervals for these differences using 250 bootstrap replications and find that both differences highly are statistically significant.

This model adapts our main specification to a discrete choice framework, but because the Yelp data do not have reviewer characteristics, we are unable to include measures of distance or the interaction terms typically used in these models to allow for a rich substitution pattern. Despite these limitations, the results provide compelling support for the underlying demand mechanism and lend confidence to the overarching conclusions of this paper. Future research on patient choice or private insurance contracts may more directly identify the magnitude of the demand response from higher online ratings.

[33]Not that while the probabilities within a given choice set add up to 1, these average values do not because the choice sets tend to have more than 3 hospitals.

2.7 Conclusions

Online reviews provide a modern, accessible source of information to consumers that is potentially more familiar than traditional hospital quality measures. It is also possible that these ratings are informative to insurers, especially if insurers are interested in patient satisfaction. In either case, there is reason to believe that online reviews may impact hospital prices. Our empirical results confirm this prediction, providing evidence that in light of higher ratings, hospital prices increase. The magnitude of this increase is relatively small compared to changes in market structure, but economically meaningful nonetheless. Our findings are further supported by a battery of alternative specifications and robustness checks. Our results are also policy relevant given numerous efforts to make hospital quality information more accessible. As we find, such efforts may have the unintended consequence of facilitating price increases even for hospitals of similar underlying quality.

While our analysis provides strong evidence of price premiums for higher-rated hospitals on this platform, plenty of questions remain. This analysis made use of the star ratings on Yelp, but there is a wealth of data available in the text of each of the reviews that could provide further insights on the mechanisms underpinning this relationship. Additionally, as other review platforms become more popular or information sources such as Hospital Compare gain traction, there may be new opportunities to investigate the connection between quality disclosure and price. Lastly, while we document evidence of a loss in consumer welfare due to higher prices from quality ratings, consumer welfare may also increase via more informed health care decisions. Quantifying the net welfare effects of quality ratings is an important question for future research.

2.8 Appendix

2.8.1 Yelp Data

Using the AHA Annual Survey database, we were able to match exactly 2,935 hospitals to Yelp profiles with reviews. In these cases, the name associated with the profile exactly matched the name listed in the AHA data. In order to ensure that the exact match profiles are associated with the correct hospital, we implemented the cleaning process described below.

Data Cleaning

1. Upon reading in the exact match data, we eliminated any observations that did not have an address associated with their Yelp profile. This takes our total observations to 2,904. Note, the address is always provided in the AHA data, and we need to match the address in the Yelp profile to that in the AHA data in order to ensure the profile is describing the proper hospital.

2. We modified the Yelp addresses so that the street, direction, and number abbreviations (for example, 'E' for 'East', 'St' for 'Street', or '1st' for 'First') follow the convention in the AHA data.

3. We then extracted the first part of the string (can be a word or number) from both the AHA and Yelp addresses. Most of the time this variable is the street number, but in some cases a street number is not listed, resulting in a word for this variable. We do the same process for the second and third words of both addresses. This results in

6 new variables – the first, second, and third word (a set of digits or letters separated by a space) for both the AHA and Yelp addresses.

These steps create uniformity among naming conventions, which allows us to automate more of the matching process. Using the following steps, we filter observations that are correctly matched and eliminate the remaining profiles. At each step, we review the observations selected by our criteria to ensure that the matches are appropriate.

4. We extracted 1,687 observations in which the AHA address and Yelp address match exactly, leaving 1,217 observations to be analyzed.

5. We extracted 333 observations where the first three words of the addresses match. This, for example, handles cases where the address is the same but one has an additional directional term at the end of the address (i.e. NE for 'North East').

6. Next we added 240 observations where the first word (often times the street number) matches and some combination of the second and third words from the address match. This handles cases where the street number matches, but one address states 'South Main Street' and the other is 'Main Street'.

7. We then flagged for manual review the remaining 258 observations where the street numbers match or where this is an '&' or the word 'and' in the name. We opt to review these manually because when the street number matches but the street does not, it often is due to cases where the hospital has its own street name but it is connected to a larger street or highway. Further, the use of '&' or 'and' typically indicates that the address is a cross-street, where both addresses are likely describing the same hospital.

8. We also flagged for manual review observations where the street number is different but the subsequent address information is the same. We check these manually to ensure that we are not including doctor's offices located in the same complex as the hospital, but we also want to be careful not to eliminate real hospital profiles.

9. We eliminated observations where none of the first three words or numbers of the addresses matched, and then flagged for manual review any remaining observations.

Manual Review of Observations

We flagged 258 observations for manual review. We began by inspecting the addresses to see if they appeared to be similar but slipped through our prior sorting process (i.e. cases where the address was listed in the AHA data as 'Ridgeview Road' and on Yelp it was written as 'Ridge View Road.') In cases where the street number was the same for both addresses but the street name differed, we used Google Maps to search each of these addresses to determine whether or not they were referring to the same location. We sorted the remaining profiles manually by finding both addresses in Google Maps to determine whether they were referring to the same location. If not, we dropped the observation.

Approximate Matches

Part of the process for collecting the Yelp data included searching for hospital profiles that had approximate matches to the AHA data instead of exact matches. In the case of approximate matches, we collected data on all of the profiles where the profile name matched with the exception of one word. We conducted a process similar to the one described above to ensure correct matches in these data. However, there were very few that were referring to the hospital in question (many were veterinary hospitals, hospital cafeterias, and physician practices). Due to the small quantity of remaining hospitals in this data, we did not use the approximate match data in our analysis. We simply mention it here to provide more information on the data collection process.

Evidence of Decision Makers Using Online Reviews

In order for online reviews to affect prices, health care decision makers must use this information. We can begin to investigate this by analyzing the text of the reviews to determine if reviewers mention using reviews to inform their decisions. To do so, we identify all of the reviews that have either "read" or "see" (or both) *and* any of the following words: review, rating, star, yelp, google.[34] Nine percent of reviews meet this criteria. This captures 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." The criteria here are relatively loose and may include reviews that do not indicate the specific behavior of interest. Thus we implement 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." One percent of the reviews meet this criteria. We read a sample of 50 reviews and found that each explicitly mentions consulting online reviews.

We conclude from this text analysis that between one and ten percent of persons on the Yelp platform considered online reviews in selecting a hospital. This is not to say that only

[34]We first preprocess the review text to impose all lower-case text.

10% of potential patients use this information. Within the data, the 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.” This comment indicates that this person consulted the reviews, but the verbiage slips through the search criteria. Further, it is of course 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, the fact that some reviewers state that they read reviews in advance, lends confidence to the idea that online reviews are relevant to health care decision makers.

2.8.2 Variable Construction

We construct hospital prices based on the following formula and variables, which come from HCRIS Data. Due to the multiple versions of the cost reports (1996 and 2010), the variable locations are listed for both formats. This price measure, which is intended to capture the average price for a non-Medicare inpatient stay, is defined as follows:

$$\text{non-Medicare price} = \frac{\text{inpatient charges} \cdot \big(1 - \text{discount factor}\big) - \text{Medicare payments}}{\text{total inpatient discharges} - \text{Medicare discharges}}. \tag{2.10}$$

The inpatient charges and discount factor variables combine multiple values from the HCRIS dataset. Inpatient charges is the sum of three variables from Worksheet G-2, Parts 1 & 2: hospital general inpatient routine care services revenue (1996: line 1, column 1; 2010: line

1, column 1), total intensive care type inpatient hospital services revenue (1996: line 15, column 1; 2010: line 16, column 1), and inpatient ancillary services revenue (1996: line 17, column 1; 2010: line 18, column 1). This provides a proxy for hospital charges. The discount factor comes from Worksheet G-3. It is the ratio of contractual allowances and discounts on patients' accounts (1996: line 2, column 1; 2010: line 2, column 1) to total patient revenues (1996: line 1, column 1; 2010: line 1, column 1). Note that because we are using a measure of charges, we need to discount those values to create a measure of what is actually paid by non-Medicare payers.

The remaining variables come directly from the HCRIS data. The first is Medicare payments from Worksheet E, Part A: total Medicare payments (1996: line 18, column 1; 2010: line 61, column 1). We subtract Medicare payments, because we are only interested in the the price for private insurers. The data do not separate out Medicaid payments and therefore we cannot exclude those from the inpatient charges variable. Worksheet S-3, Part 1 contains inpatient discharges (1996: line 1, column 15; 2010: line 1, column 15), which comprises all inpatient discharges from the hospital, and thus includes patients from all payer types. Lastly, Medicare discharges comes from Worksheet S-3, Part 1: Medicare discharges (1996: line 1, column 13; 2010: line 1, column 13). We subtract Medicare discharges from the denominator because, again, we are only interested in the price for private payers, so we need to eliminate Medicare patients from this calculation.

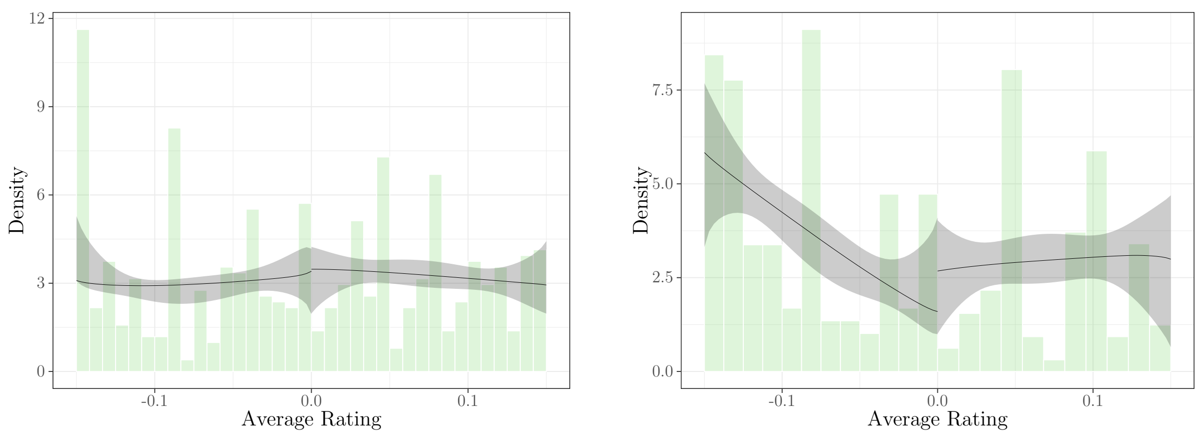
2.8.3 Sorting at Various Review Counts

Section 2.4.2 in the main text presents the manipulation tests from Cattaneo et al. (2018) at each threshold for the data in our main specification, i.e. where we require a minimum of three reviews for a hospital to be considered rated. In this case, we find statistically significant evidence of sorting at each of the thresholds, which we argue is a mechanical aspect of the data as opposed to evidence of intentional sorting by the hospitals. In order to test that assertion, we recalculate this test for a higher level of minimum reviews. Figure 2.8a shows the tests for the high and middle thresholds, respectively, where the minimum review requirement is 5. In this case, there is no longer evidence of sorting. The p-value is 0.9536 at the middle threshold and 0.6087 at high. These results, in conjunction with the institutional details described in Section 2.4.2, lead us to conclude that the sorting apparent in this test is not driven by hospital behavior. Lastly, our discussion of manipulation around the threshold is prompted by our concern about how this may impact our results. Hence, in 2.5, we present alternative specifications where the minimum number of required ratings ranges from 4 to 10, and we find that our main results fall at the lower end of these estimates. In other words, when we impose a minimum of 5 reviews—which alleviates any issues with the manipulation tests—we find qualitatively similar results with earlier point estimates. This remains true as we increase the minimum number of reviews.

2.8.4 Sensitivity Analysis

In addition to the alternative specifications presented in the main text, here we present analyses of violations of the exclusion restriction, the importance of outliers, and alternative

Figure 2.8: Manipulation Tests at High and Middle Thresholds with 5 Review Minimum



(a) Middle Threshold (b) High Threshold

Notes: The graphs depict the manipulation tests from Cattaneo et al. (2018). The x-axis shows the average rating for a hospital at year-end. The light green bars show the histogram of average ratings. The density estimates are on the y-axis. In both panels, the bandwidth is 0.15. Neither figure shows statistical evidence of sorting. 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.

rating groups. Each of the sets of results demonstrate that our results are robust to these concerns.

Violations of the Exclusion Restriction

We are interested in investigating how sensitive our analysis is to assumptions made on the instrument. To begin, we analyze the bandwidth assumption. Recall in our main analysis, the bandwidth around each threshold is 0.15. This means that to be considered rounded into the middle group, a hospital would need an average rating between 2.75 and 2.90. The specification titled "Bandwidth" modifies the instrument so that the bandwidth around the high and middle thresholds (3.75 and 2.75, respectively) is 0.10. As detailed in Section 2.4, the intuition behind the instrument is that hospitals directly on either side of the threshold are of comparable underlying quality, and the only difference for the observations above the threshold is that they are rounded into a higher rating. If our results were driven by observations furthest away from the threshold, this would call into question the underlying assumption. As we can see in the figure, these results tell a similar story, just with a larger magnitude on the coefficient for high rated hospitals, which provides support for this assumption on the instrument.

We further scrutinize the strength of our instrument using the estimation method proposed in Conley et al. (2012). The methodology allows for IV estimation under flexible (i.e. plausibly exogenous) conditions. Consider the following regression equation:

$$Y = X\beta + Z\gamma + \varepsilon, \tag{2.11}$$

where Y is a vector of outcomes, X is a matrix of endogenous variables, and Z is the matrix of instruments, that we assume are uncorrelated with the error term. This means that in a standard IV setting, we assume that γ is precisely equal to zero. Conley et al. (2012) replaces this exclusion restriction, with an assumption on the support or distribution of the correlation to allow for non-zero values. To implement, practitioners must specify minimum and maximum values for the correlation between the instruments and the error term. This approach is then implemented by using these minimum and maximum values to estimate the equation twice, and then taking the union of the resulting confidence intervals. Further, instead of only estimating using the bounds of the specified support, this estimation can be conducted numerous times by taking equally spaced points within this support and then reporting the corresponding union of confidence intervals.

To implement this approach, we must specify the set of support conditions and the number of points on the support to calculate bounds. In our analysis, a violation of the exclusion restriction would mean that being rounded has a direct effect on prices, as opposed to working through the ratings. Recall that the hospitals that are rounded into a higher rating group have a lower underlying cumulative average rating than their corresponding star rating, i.e., the underlying quality measure for any rounded hospital is lower than non-rounded hospitals in the same rating group. Hence, we would not expect any additional positive impact on price from being rounded that is not already captured by the star rating. Rather, we would expect that if the exclusion restriction does not hold, the violation is negative due to the lower underlying quality amongst rounded hospitals relative to the non-rounded hospitals with the same rating.

As such, we set the lower bound of the support as the negative value of the reduced form coefficients in column 1, the least saturated specification (i.e. -0.0219 for high, and -0.0575 for middle). We specify the upper bounds as zero for both instruments. Further, we estimate this relationship for 50 points on the support. Under these assumptions, the confidence intervals are [0.0013, 0.211] for the middle group and [0.0011, 0.229] for the high group. Similarly, if we conduct the same analysis but use the negated coefficients from column 4 of table 2.3 (i.e. -0.0230 for high, and -0.0214 for middle), we get the following confidence intervals: [0.0013, 0.1348] for the middle group and [0.0011, 0.1706] for the high group. This indicates that even under the relaxation of the exclusion restriction, there is strong evidence of a price premium for hospitals that do not have a sufficiently low rating on this platform.

Importance of Outliers

There are well-known concerns about outliers in an instrumental variables setting. Further, in the research analyzing hospital prices, there are issues with outliers, particularly when using an estimate for price based on the CMS cost reports. We take the following steps to address concerns regarding outliers.

At the onset, we drop price outliers at the 5th and 95th percentiles. This is commensurate with approaches taken in Dafny (2009), Schmitt (2018), and Lin et al. (2020), as detailed in the main text. As previously discussed, we also analyze price changes over time and use the the natural log of prices in the following year as our main outcome of interest.

More formally, we address the issues posed by outliers in instrumental variable analyses by using an additional econometric estimator. To do so, we rely on the estimator proposed in Freue et al. (2013) and implemented in Stata by Desbordes and Verardi (2012). This estimator differs from traditional IV estimation because it uses a robust multivariate location and scatter S-estimator to provide increased robustness in comparison to traditional IV. S-estimators are useful in this setting because under certain regularity conditions, they reach maximal break down point regardless of the dimension of the data (Freue et al., 2013). Practically speaking, this methodology as implemented in Desbordes and Verardi (2012) identifies outliers as the observations that have a robust Mahalanobis distance over a certain threshold (0.99 is the standard and is what we use in our analysis); these observations are then given a weight of zero and remaining observations receive weights equal to one.

Our results for this estimator are shown in Table 2.7. The specification includes hospital and year fixed effects, along with all of the covariates and hospital quality measures. Note that this estimation technique drops approximately 25% of the observations that are included in our main sample, as those observations are found to be outliers. While this produces a less precise estimate for the high rating group, it is qualitatively consistent with the results shown in Figure 2.7. The estimate for the middle rating group is quite similar to that of our main results, and is actually more precise. Taken together, these results provide further support for the presence of a price premium for hospitals with sufficiently high ratings.

Table 2.7: Robust Instrumental Variable Results

Price	
High Rating	0.0338
	(0.0342)
Middle Rating	0.0509*
	(0.0301)
No Reviews	0.0392*
	(0.0205)
Fewer than 3 Reviews	0.0279
	(0.0185)
Observations	6058
Kleibergen-Paap LM Statistic	98.00
Kleibergen-Paap F-stat	142.7

Notes: Robust standard errors clustered at the hospital level are in parentheses. Stars indicate the following: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

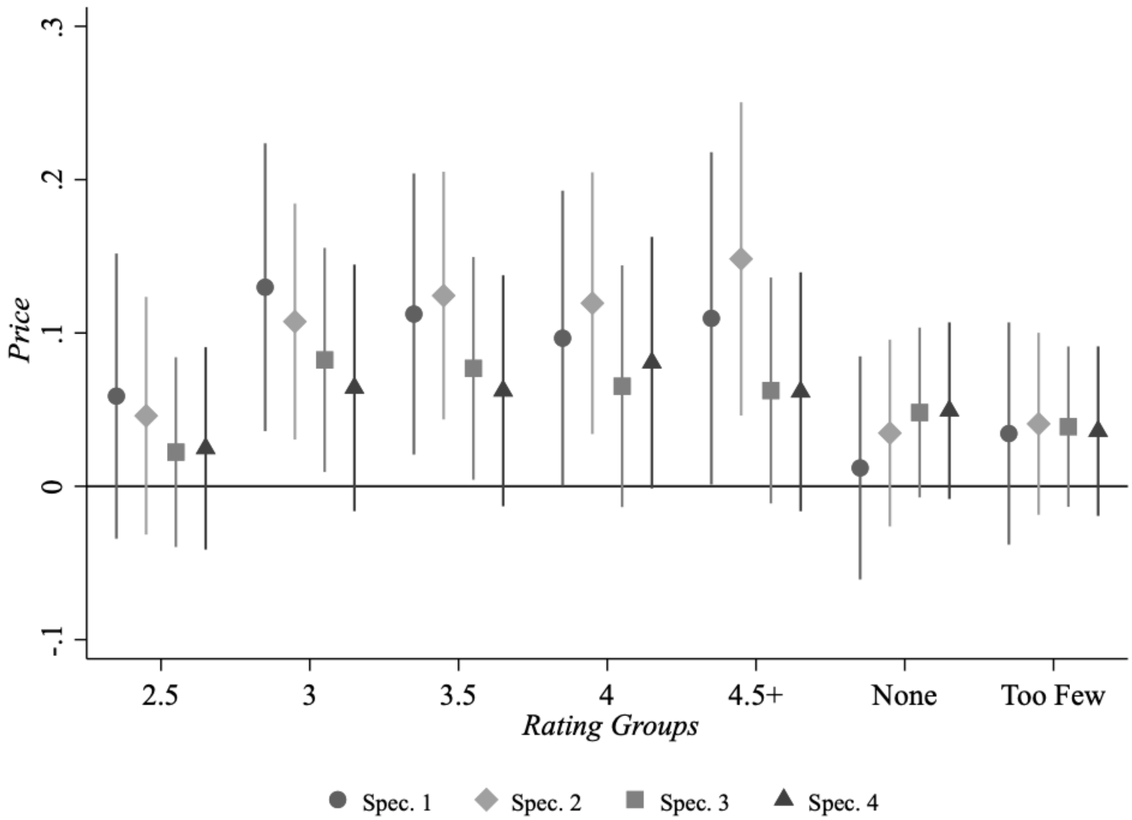
Relevance of Rating Groups

Recall in Section 2.3, we detailed our selection of the middle and high rating groups, elaborating on the complications that we face if we attempt the analysis with each of the individual rating levels. To provide some clarity about our decision to create these rating groups, we estimate a specification that is less granular than using every rating group individually, but more granular than the rating categories we use throughout our analysis. Here we maintain the "no reviews" group, which consists of hospitals that do not have a Yelp profile, along with the "too few" group comprising hospitals with fewer than three reviews on Yelp.

We define "low" as hospitals with ratings of 2 or lower. Additionally, we define "4.5+" as hospitals with year-end ratings of 4.5 or higher. Observations that do not fall into one of the groups already mentioned are included in the analysis using a dummy variable for their year-end rating. In other words, hospitals with ratings of 2.5, 3, 3.5, or 4 are not collapsed into a broader rating group. This results in a set of 8 indicators, one for each of the following categories: no reviews, too few, low, 2.5, 3, 3.5, 4, and 4.5+, where the low group is excluded in our analysis. The "none" and "too few" groups are presented here as well to coincide with the results presented throughout the main text. Lastly, the instrumental variables for the 2.5 through 4.5+ rating categories are analogous to the main results. For example, the 2.5 rating category comprises observations with an average rating between 2.25 and 2.74, and the corresponding instrument is an indicator equal to 1 for observations with an average rating of 2.25 through 2.4.

The results are shown in Figure 2.9. Specifications 1-4 shown in the legend are analogous to those shown in Table 2.4, where specification 4 is the most saturated, including hospital and county characteristics, hospital quality measures, and year, county, and hospital fixed effects. We see that across all specifications the coefficient on hospitals with 2.5 stars is small and insignificant. Note that the average rating for a hospital with a Yelp presence is 2.9, as shown in Table 2.1. As such, it follows that we do not find a significant effect for hospitals

Figure 2.9: Results with More Granular Rating Groups



Notes: The four specifications correspond to those detailed in Table 2.4. The confidence intervals associated with each point estimate are for robust standard errors clustered at the hospital level. The outcome, *price*, follows the same definition as the main results.

slightly below average.

We see that for the 3 through 4.5+ rating groups, the results are consistently positive, but lose significance in the most saturated specification.[35] This result informs our selection of the rating groups used in our main analysis: our "middle" rating group throughout the paper consists of hospitals with 3 or 3.5 stars, and the "high" group encompasses the 4 and 4.5+ groups. By placing the cut-off for middle at 3, we require those hospitals to have marginally

[35]In specification 3, the joint significance of 3 and 3.5-star ratings has a p-value of 0.084, and the 4 and 4.5+ groups have a p-value of 0.236, indicating that neither group is jointly significant. Again, in the most robust specification, neither group is jointly significant (p-values of 0.259 and 0.151 for the middle and high group, respectively).

better than average ratings, and then the high rated hospitals are sufficiently better rated than average.

2.8.5 Falsification

Figure 2.10 presents the results of our falsification analysis. The coefficients for the high, middle, and no reviews groups are shown with circles, triangles, and diamonds, respectively. The outcome of interest corresponding to each of the coefficients is indicated by the black dot in the bottom panel titled “Outcome”. All estimates adopt the specification in column 4 of Table 2.4 (i.e. year, county, and hospital fixed effects, hospital and county characteristics, and hospital quality measures). In general, we selected future hospital quality metrics and lagged prices as our outcomes of interest because these outcomes should be unaffected by the ratings at time t.

Figure 2.10a presents the specification with clinical quality measures as the outcome of interest. Specifically, we use mortality and readmissions data for AMI, heart failure, and pneumonia in $t+1$. [36] We argue that these clinical quality measures are well-suited for a falsification test because, while clinical quality may be correlated with price and may also affect Yelp ratings, we have no reason to suspect that an increase in reported ratings at time t would impact clinical quality measures at time $t+1$. Looking at the results in Figure 2.10a, we see that across the board the coefficients on our rating categories for these measures are economically small and statistically indistinguishable from zero.

Figure 2.10b presents our falsification tests related to prices. The specification here follows

[36] We exclude quality measures at time t in this analysis, although the results are not sensitive to this exclusion.

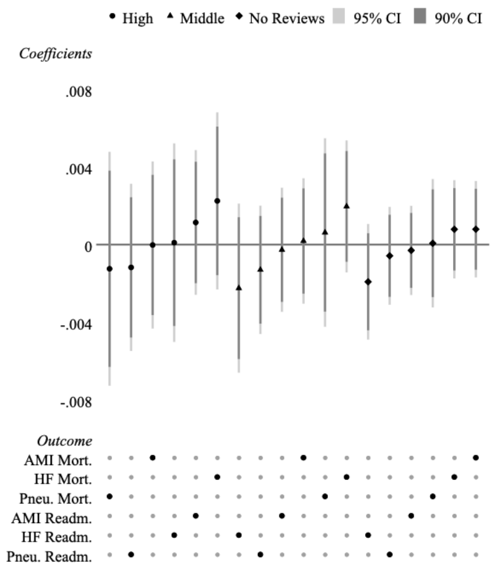
exactly from our main results, the only difference being the change in the outcome variable, which is now log price at time t, $t-1$, and $t-2$. As discussed in Section 2.3, because the hospital fiscal year can end earlier than the calendar year, we would not expect a rating change in time t to impact the time t price (where that price may be based on a fiscal year that ends months earlier than the calendar year-end). Thus, we include price at time t along with two additional price lags. To the extent that a higher rating allows hospitals to leverage their bargaining position to negotiate higher prices with insurers, we would not expect a significant relationship between ratings and past prices, given other controls. Referring to Figure 2.10b we see that the relationship between each of the three explanatory variables and the various price variables is statistically no different from zero in each case.

Given the results shown in Figure 2.10, it is clear that we are not finding results where we would not expect them, generating further confidence in our main results. Building on the discussion presented in Sections 2.5 and 2.8.4, these falsification tests provide convincing evidence that our main results are not fully spurious.

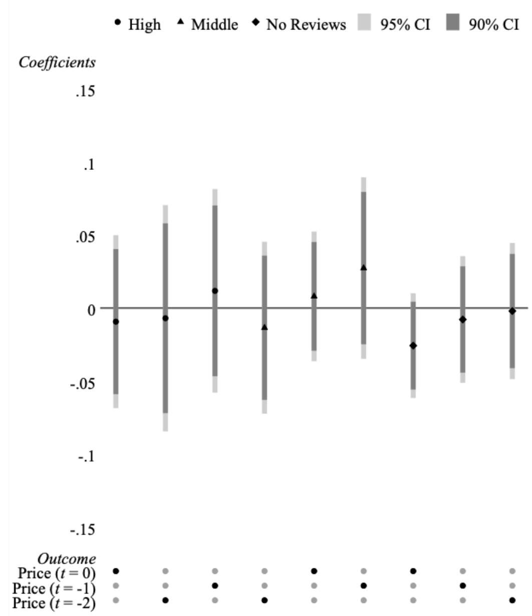
2.8.6 Community Detection Methods

Community detection algorithms provide a novel way to define hospital markets based upon observed hospital choices. For this analysis, I use publicly available data from the CMS Hospital Service Area Files, which include annual summaries of the number of inpatient discharges for Medicare fee-for-service patients by zip code, and we aggregate these data to the FIPS code level. The methodology is implemented using an adjacency matrix that relates patients from common areas (counties, zip codes, etc.) to the hospitals that they choose.

Figure 2.10: Falsification Tests



(a) Clinical Quality Outcomes



(b) Price Outcomes

Notes: The coefficients for high ratings are shown with circles, those for middle ratings are shown with triangles, and no reviews are shown with diamonds. The outcome of interest is indicated in the bottom panel. Each of the mortality and readmission variables are quality metrics for $t+1$.

More specifically, the first step is to create a bipartite matrix that relates FIPS codes and hospitals using zeros and ones. Here, one indicates that people from the respective FIPS code went to the corresponding hospital. This, however, fails to consider the volume of patients. To address this, I use a minimum share of 0.15, which then only connects hospitals and FIPS codes where at least 15% of that hospital's overall admissions come from that FIPS code.

By multiplying the bipartite matrix by its transpose, I create the unipartite adjacency matrix that is symmetric and indicates the number of hospitals that were selected by a sufficient portion of people in both FIPS codes. The algorithm takes this matrix and identifies markets based on common hospitals between FIPS codes. There are a variety of community detection algorithms, but for this analysis, we use the Louvain approach because it had the highest modularity score (0.972), which means that it was the best at dividing the network of hospitals into separate markets. This market definition had a total of 772 markets across the country.

Chapter 3

Medicaid Expansion and Patient Experience

Through public policy and market forces, hospitals face incentives to prioritize patient experience of care which may drive strategic, non-clinical quality investments. Empirical evidence of this behavior, however, has been limited by the absence of data on non-clinical quality and the endogenous nature of quality improvements. Using hospital Yelp reviews to capture patient experience and features of hospital quality along with expanded Medicaid eligibility to isolate a shock to hospital finances, this paper overcomes these challenges. Through an interaction-weighted two-way fixed effects approach, the analysis finds that Medicaid expansion had a substantial effect on hospital finances and patient satisfaction. Hospitals in expansion states experienced ratings that were on average 0.3 to 0.4 stars higher compared to non-expansion states. Analysis of the review text provides additional insight into the elements of care that drive these ratings. The study provides new evidence about the dimensions

of quality upon which hospitals may work to differentiate themselves.

3.1 Introduction

Health care providers face incentives to prioritize patient experience of care. Whether through policy initiatives such as the Value Based Purchasing Program or broader mechanisms, like increasing demand or improving their bargaining position in negotiations with private insurers, hospitals are likely to divert resources to dimensions of care that are of particularly salient to patients (Chandra et al., 2016; Dranove and Jin, 2010; Dafny and Dranove, 2008; Dranove and Sfekas, 2008; Garthwaite et al., 2020; Zhe Jin and Sorensen, 2006). While understanding how hospitals invest in these features—like amenities and non-clinical aspects of quality—has implications for the efficiency of health care delivery, there is limited empirical evidence on strategic quality investment in health care. This study provides new insights into this relationship by addressing the question: do hospitals prioritize investment in patient-centered quality?

Ideally, I would address this question by conducting an experiment in which some hospitals receive an influx of cash and directly observe how they spend it compared to the hospitals that receive no intervention. While practically this is infeasible, using a policy shock and novel patient experience data, I am able to mimic two important features of the ideal experiment.

First, I exploit the Medicaid eligibility expansion of the Affordable Care Act (ACA), which provided an exogenous shock to hospital operating environments. This allows me to disentangle quality investments that are due to a plausibly exogenous change in a hospital's financial position from investments driven by endogenous factors such as prior quality investments or changes in the level of competition in a hospital's market. There are two specific ways that expanded Medicaid eligibility could have driven hospitals to prioritize investment in dimensions of care that may affect patient experience. The first is by directly increasing revenue. Some existing studies find that hospitals in expansion states experienced increases in Medicaid revenue and decreases in uncompensated care, generating overall higher revenue (Nikpay et al., 2015; Blavin, 2016; Rhodes et al., 2020). There is also contrasting work that does not find evidence of revenue increases as a result of this policy (Moghtaderi et al., 2020). Even in that case, this policy change may have decreased a hospitals risk of losing money on uncompensated care, thereby changing the perception of the hospitals operating environment and freeing up cash to redirect toward patient-centered quality investments. In either case, state-level variation in Medicaid eligibility provides a valuable context in which to investigate hospital quality investment.

The second feature of the ideal experiment would be to directly observe hospital investment decisions. This unfortunately does not pan out in reality. Existing data lack the granularity to identify specific dimensions upon which hospitals invest, and therefore, have been ill-equipped to further our understanding of hospital investment. I address this challenge using online hospital reviews from Yelp (Ranard et al., 2016). Yelp reviews are a modern measure of patient satisfaction and the experience of care at a given hospital. Through star ratings, they provide an accessible, real-time metric to measure hospital quality from the patient perspective. Through narrative comments, they highlight the dimensions of care that were particularly important to a patient's experience, providing an innovative way to capture aspects of quality—such as such as food, valet, and other non-clinical features—that are not found in other data. While imperfect, the Yelp data provide a measure of patient experience and valuable new information on the the specific dimensions of care that drive that experience.

Using this plausibly exogenous change to a hospital's financial standing and hospital Yelp reviews, I assess how shocks to hospital budgets impacted patient satisfaction and investment in patient-centered quality. Given the staggered implementation across states, an intuitive estimation strategy is to use a two-way fixed effects model with leads and lags. Recent findings in the econometric literature, however, caution against naively implementing this approach, as the coefficients on the leads and lags may be be contaminated by the effect of treatment in other periods. To avoid this, I rely on the interaction-weighted approach proposed in Sun and Abraham (2021); I use this estimator across all outcomes of interest.

First, I test the underlying idea that Medicaid expansion had substantive effects on hospitals by considering as outcomes Medicaid discharges, uncompensated care, and net patient revenue. Following expanded Medicaid eligibility, hospitals in expansion states experience higher Medicaid discharges, lower uncompensated care costs, and higher net patient revenue. These results show that hospitals experienced changes in their operating environments that likely drove windfall profits, which enhanced their ability to invest in patient-facing dimensions of quality.

Based on this idea, I then turn to analyzing the effect of Medicaid expansion on patient experience, as measured by Yelp star ratings. Two years after Medicaid expansion, hospitals in expansion states received star ratings that were 0.3 to 0.4 stars higher, on average, than hospitals non-expansion states. While this result provides compelling evidence of how these hospitals differentially improved patient experience—arguably through investments in dimensions of quality that drive patient experience—it cannot specifically speak to the underlying mechanism. However, using natural language processing on the review text, I identify which reviews speak to non-clinical features of quality. I find that in the years following Medicaid expansion, hospitals in expansion states experience a significant increase in reviews that positively mention amenities relative to the control hospitals. Taken together, these results provide evidence that suggests that hospitals use the exogenous improvement to their financial standing to differentially advance their investment in non-clinical quality.

This study contributes to existing research on hospital investment and the characterization of quality in health care. Various studies highlight the incentives that providers face to prioritize measures of patient satisfaction, and therefore, invest in features of quality that might improve these metrics (Chandra et al., 2016; Dranove and Jin, 2010; Dranove and Sfekas, 2008; Garthwaite et al., 2020). Garthwaite et al. (2020) presents a model in which hospitals make investments in quality to increase private revenue. They test their model predictions using various measures of quality–one of which is patient satisfaction. This paper builds on their findings by providing further empirical evidence to this previously speculative notion.

Additionally, research finds that the correlation between clinical quality and patient satisfaction is weak, indicating that aspects of the hospital experience other than clinical quality drive patient experience (Ranard et al., 2016; Perez and Freedman, 2018). Measures of patient experience are relatively new, and while researchers have expanded our understanding of what patient experience measures may capture through comparative analyses with other quality measures, we have a limited understanding of the ways in which these metrics relate to the incentive structures at play in hospital markets. This study indicates that providers actively respond to the incentive to bolster patient experience of care.

By using Medicaid expansion to isolate an exogenous change to hospital budgets, this study also builds on the wealth of literature that assesses the impact of expanded Medicaid eligibility.[1] A large body of work has explored how insurance expansion affects health care markets, particularly from the standpoint of the individual. However, as this study shows, the effects of these expansions are not limited to clinical outcomes and can have notable effects on hospital incentives, as well. By exploiting the exogenous change to hospital operating environments spurred by the expansion of Medicaid eligibility in combination with online review data, this study furthers our understanding of how hospitals may strategically invest in features of quality that patients value.

3.2 Data

This paper explores the effect of Medicaid expansion on patient experience using the universe of hospital Yelp reviews. The Yelp platform is well-suited for this study because of its popularity before and after 2014, which is when the majority of expansion states expanded Medicaid. I also use data from the American Hospital Association (AHA) Annual Survey and Cost Report data reported to the Healthcare Cost Report Information System (HCRIS) at the Centers for Medicare and Medicaid Services. In what follows, I first describe the data

[1]See Miller et al. (2021), among others, for more discussion of this literature.

sources independently and then detail the final combined dataset.

3.2.1 Medicaid Expansion

Information on state Medicaid expansion decisions comes from the Kaiser Family Foundation.[2] While some states have expanded Medicaid as recently as 2021, this study focuses on the states that expanded prior to year-end 2018, which includes 31 states and DC. Figure 3.1 shows expansion by state. I aggregate the data to the year level based on the calendar year in which states expanded, even though some states implement the policy later in the year than January 1. To the extent that this affects estimation, it should bias the estimates downward because hospitals in states that expanded later in the year would have less exposure to the policy change, thereby tempering these effects.

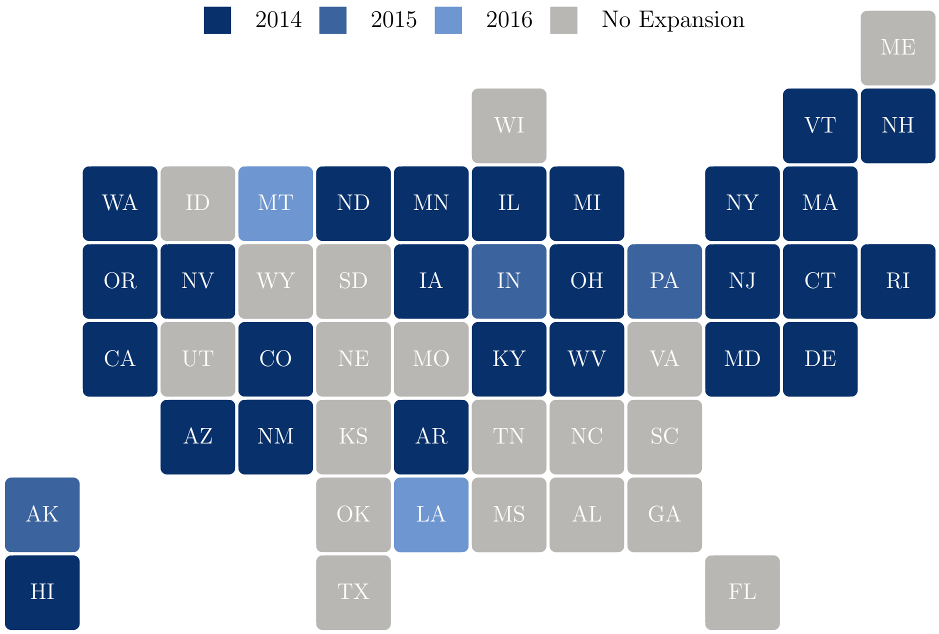
3.2.2 Hospital Data

I use the AHA Annual Survey data for various hospital characteristics. For each hospital-year observation, the data include the hospital's Medicare record number, information on system membership, number of beds, physicians, and nurses, and variables concerning ownership status and teaching hospital designations. I also use the AHA's measure for annual Medicaid discharges. I use Cost Report data for uncompensated care costs and net patient revenue.[3] These three variables—Medicaid discharges, uncompensated care costs, and net patient revenue—comprise the set of outcomes that I use to assess how Medicaid expansion affected hospital operating environments.

[2]Information can be found here: https://www.kff.org/medicaid/issue-brief/status-of-state-medicaid-expansion-decisions-interactive-map/.

[3]The supplemental material details the specific line items of the Cost Reports used to pull these variables.

Figure 3.1: Map of State Medicaid Expansion through 2018



Notes: This map shows Medicaid expansion through year-end 2018. All states that expanded in 2014 did so on January 1, 2014 with the exception of Michigan and New Hampshire, which expanded on April 1, 2014 and August 15, 2014, respectively. In 2015, Pennsylvania expanded on January 1, Indiana on February 1, and Alaska on September 1. In 2016, Montana expanded on January 1 and Louisiana on July 1.

To work with the notoriously noisy Cost Report data, I remove outliers at the fifth and 95th percentiles, along with any observations with missing data.[4] I combine the AHA and HCRIS data using the hospital's Medicare record number, and then, I match these data to the state-year Medicare expansion data. Lastly, I remove from my sample hospitals that do not appear in the data every year, i.e., I impose a balanced panel. This leaves 1,646 hospitals and a total of 11,522 hospital-year observations. Table 3.1 uses pre-treatment data to summarize these variables by state expansion status. The panel of hospital characteristics

[4]Prior to eliminating outliers, I deflate these values to 2012 dollars using the Consumer Price Index for All Urban Consumers: Medical Care in U.S. City Average (CPIMEDSL) from the U.S. Bureau of Labor Statistics.

are used in the estimating the models that include covariates.

Table 3.1: Summary Statistics by State Expansion Status

	Expansion States		Non-Expansion States	
	Mean	St. Dev.	Mean	St. Dev.
Hospital Characteristics				
Total Beds	233.23	189.15	239.64	232.30
Total Admissions	10,803.70	9,950.69	10,360.56	11,234.35
Total Inpatient Discharges	55,645.18	52,996.64	55,868.82	62,548.76
Government	0.15	0.35	0.24	0.42
Non-profit	0.75	0.43	0.54	0.49
Major Teaching Hospital	0.11	0.31	0.06	0.23
Any Teaching Hospital	0.45	0.49	0.40	0.43
System Member	0.64	0.47	0.72	0.43
Physicians	41.77	141.29	25.61	85.40
Nurses	451.53	503.46	449.57	563.08
Outcomes of Interest				
Medicaid Discharges	2,224.53	2,603.22	2,048.19	2,669.45
Uncompensated Care (1,000s)	12,011.10	26,230.59	14,905.35	34,216.65
Net Patient Revenue (1,000s)	250,215.51	308,539.83	199,318.36	253,554.72

Notes: This table summarizes average hospital characteristics by state expansion status. For hospitals in expansion states, i.e., treated hospitals, I present average pre-treatment characteristics. For hospitals in non-expansion states, i.e., control hospitals, I present average characteristics over the entire sample period (2012 - 2018).

3.2.3 Yelp Data

To measure patient experience and features of hospital quality, I use online reviews from the popular rating platform, Yelp. Yelp is a well-suited platform for this study because it launched in 2004, enabling hospitals to amass reviews both before and after the implementation of the ACA. Once a hospital has a Yelp profile—which can be created by a user, or a hospital, or a Yelp employee—then users can post a review. A Yelp review has two required components: a star rating between 1 and 5 and a narrative comment. Note that only registered users may post a review, but anyone online can see the reviews. The reviews were collected using web scraping techniques and comprise the universe of hospital on Yelp through year-end 2018. Among other variables, the data contain both the star rating and the review text, along with the date of the review.[5]

The Yelp data consist of 63,715 reviews from 2,158 hospitals. After limiting this to the hospitals that appear in the hospital characteristics data to ensure that the results are comparable across samples, this leaves 25,893 reviews from 787 hospitals. This means that of the 1,646 hospitals in my sample, 47.8% have a review on Yelp by the end of the sample period. The data are bimodal, with 1-star reviews comprising 48.5% of the data, followed by 5-star reviews at 29.8%. These review-level Yelp data are the basis for the results that assess the effect of Medicaid expansion on patient experience measured by Yelp star ratings. Beyond the patient-driven star ratings, Yelp reviews posses another feature that makes them a compelling data source for this analysis. Specifically, the review text creates the opportunity to use natural language processing to capture specific features of the hospital experience that matter to patients. I discuss this further in Section 3.3.

3.3 Empirical Approach

Using the outlined data, this study advances our understanding of hospital strategic investment behavior by analyzing the effect of expanded Medicaid eligibility on patient experience. Spurred by the ACA, states started expanding their Medicaid programs in 2014. Over my sample period, the largest cohort of expansion states was in 2014, with 26 states. Then in 2015, three states expanded, followed by two more in 2016. Traditionally, economists have estimated the effect of this type of staggered treatment using an two-way fixed effects

[5]Yelp's algorithm hides reviews that are deemed fraudulent, and these reviews, therefore, are not included in this dataset. McCarthy et al. (2020) details the collection and cleaning process for these data.

(TWFE) model with leads and lags, i.e., an event study. Recent advances in the econometrics literature, however, have divulged the shortcomings of this approach. Sun and Abraham (2021) shows that in the context of staggered treatment, the leads and lags in the TWFE model may be contaminated by the effect of treatment in other periods. They propose an alternative, interaction-weighted estimator to address the issue of contamination, which is the basis for the estimating equations in the following subsections. I present the subsequent results in Section 3.4.[6]

3.3.1 First Order Outcomes

The "first order" analysis focuses on the direct effect of Medicaid expansion on hospital operating environments by considering the following outcomes: Medicaid discharges, uncompensated care costs, and net patient revenue. The Sun and Abraham (2021) specification for this analysis is:

$$Y_{ht} = \sum_{e \in C} \sum_{\substack{l=-3 \\ l \neq -1}}^{l=4} \delta_{e,l} \big(\mathbb{I}\{E_h = e\} \cdot D_{ht}^{l} \big) + \gamma_t + \gamma_h + \epsilon_{ht}. \tag{3.1}$$

On the right hand side of Equation 3.1, the term $\mathbb{I}\{E_h = e\}$ is a cohort indicator. Note that the control group is the set of hospitals not treated during this sample period, meaning that the set C includes all observations. The term D_{ht}^{l} is defined as $\mathbb{I}\{t - E_h = l\}$, i.e., an indicator for hospital h being l years away from state-level Medicaid expansion at year t. This term is zero for all hospitals in non-expansion states. I omit the year prior to expansion, i.e. $y = -1$. The terms γ_t and γ_h denote year and hospital fixed effects. I conduct the analysis

[6] The supplemental material includes the TWFE results.

using hospital-year level data. Lastly, in some results, I include a term, X_{ht}, which is a vector that contains as covariates the hospital characteristics outlined in Table 3.1. These include indicators to capture ownership type and teaching hospital designations, counts of beds, nurses, and physicians.

3.3.2 Patient Experience

I use a similar empirical approach to assess the change in experience for patients who receive care in hospitals in expansion versus non-expansion states. Recall, my data contain the star rating and review text that a reviewer left on a given hospital's Yelp profile. I therefore modify Equation 3.1 to reflect the review-level outcome and update the error term accordingly. This results in the following equation:

$$Y_{iht} = \sum_{e \in C} \sum_{\substack{l=-3 \\ l \neq -1}}^{l=4} \delta_{e,l} \big(\mathbb{I}\{E_h = e\} \cdot D_{ht}^{l} \big) + \gamma_t + \gamma_h + \epsilon_{iht}. \tag{3.2}$$

Here, the outcome of interest Y_{iht} is the star rating that reviewer i posted about hospital h in year t. Equation 3.2 is the basis for my main result.

It is worth noting that Equation 3.2 does not contain a vector of hospital covariates, X_{ht}. This is because Sun and Abraham (2021) states that additional assumptions beyond the scope of their paper may be needed to maintain the same interpretation of the treatment effects if covariates are included. Another leader in this difference-in-differences renaissance, Callaway and Sant'Anna (2021), allows for covariates, but unlike Sun and Abraham (2021), it rules out the use of fixed effects, which are important in this analysis. It is well understood in the hospital literature that when analyzing hospital quality or prices, it is often the case that the variation in these measures cannot be explained by other hospital characteristics. This means that we cannot rely on a set of hospital covariates, and instead, we condition on baseline, unobserved hospital quality using fixed effects. Therefore, in Section 3.4 my preferred specification is the Sun and Abraham (2021) model (hereafter "SA") without covariates, but I also present results with covariates included to rule out concerns about the results being driven by the exclusion of some obvious covariate.

3.4 Results

The following subsections present the results for the estimating equations outlined in Section 3.3, starting with the first order outcomes and finishing with patient experience and non-clinical quality measures. For each outcome, I present a the SA estimates with and without covariates, where the preferred estimates are those without covariates. These estimates are plotted using black circles and are the basis for the subsequent discussion.

3.4.1 First Order Results

To reliably estimate the effect of Medicaid expansion on patient experience, I first quantify the extent to which this policy change affected hospital operating environments. This requires exploring various outcomes that are likely to be mechanically related to this policy shock, which could subsequently spur hospitals to reallocate investment toward dimensions of care that drive patient experience.

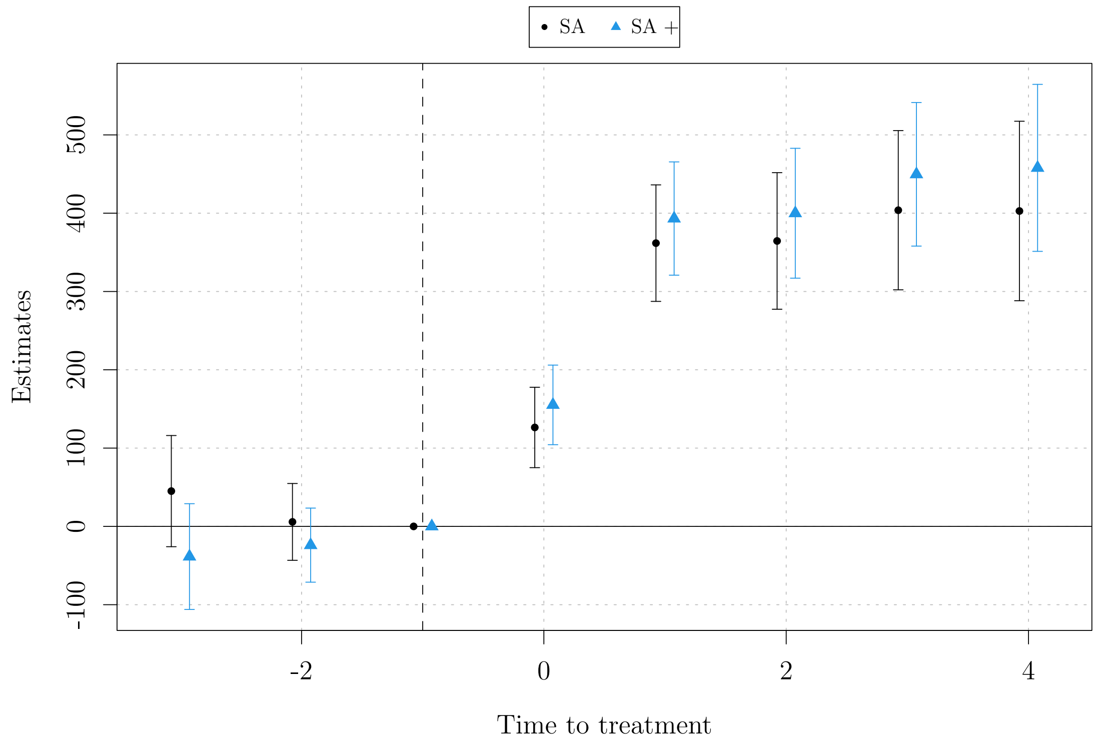
The first outcome I consider is Medicaid discharges, which I expect to increase in hospitals in expansion states relative to those in non-expansion states as more people are eligible to receive Medicaid coverage. This bears out in Figure 3.2, which shows that in the year of expansion, treated hospitals saw an increase of 125 discharges, and by three years post-expansion, treated hospitals had a differential increase of approximately 400 Medicaid discharges. Therefore, hospitals in expansion states did, in fact, care for more Medicaid patients, which constitutes a change to a hospital's operating environment.[7] However, for this policy change to affect hospital investment decisions, one would also expect a change in certain dimensions of hospital budgets. One example of that is uncompensated care costs.

Figure 3.3 presents the uncompensated care results. In the year of expansion, uncompensated care costs decreased by just over \$2 million for expansion state hospitals. By four years after expansion, this relative decrease grew to \$6.3 million. This result follows other work in the literature, such as Nikpay et al. (2015) and Blavin (2016), which also finds decreased uncompensated care costs. The effect on uncompensated care clearly illustrates one dimension upon which Medicaid expansion affected hospital budgets, but whether or not it improved a hospital's overall financial position depends on some additional complications.

If Medicaid expansion provided coverage to patients who otherwise would have received care but been unable to pay, then this policy would likely both decrease uncompensated care costs and improve hospital budgets overall. However, as Nikpay et al. (2015) points out, when treating uninsured patients, hospitals can consider any difference between what the patient pays and the cost of care as uncompensated care. Yet, if Medicaid reimbursement falls short of covering the cost of care, that difference cannot be classified as "uncompensated." De-

[7]This coincides with other work, such as Miller et al. (2021), which shows that Medicaid expansion did translate into higher rates of Medicaid coverage.

Figure 3.2: Event Study Results: Medicaid Discharges

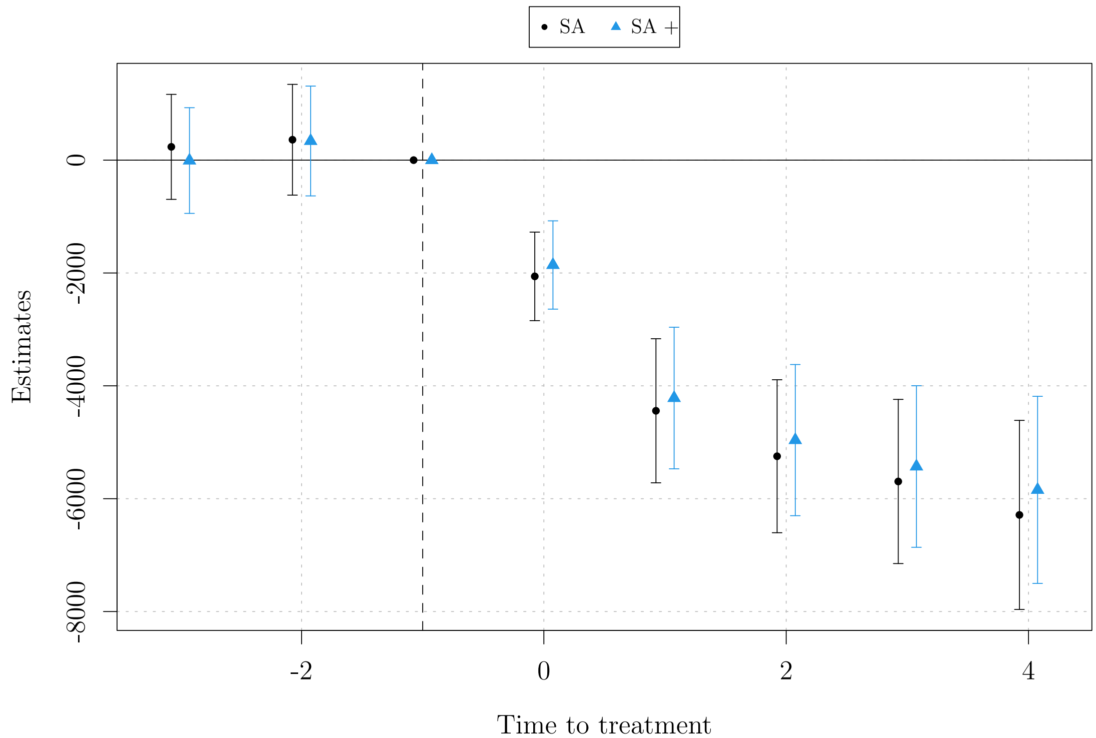


NOTES: The uncompensated care costs enter the regression in thousands, meaning that the coefficient estimates indicate decreases in the millions. The second result at each relative time includes covariates, which are indicated by the "+" following the model title in the legend. I omit observations for relative time −4 because there are only two states with observations in that relative time.

clines in uncompensated care, therefore, would indicate hospital budget improvements only if Medicaid reimbursement exceeded what hospitals would have otherwise received from these uninsured patients. Further, this potential discrepancy would be exacerbated if Medicaid expansion had any crowding out effect on private insurance. Directly investigating each of those effects is beyond the scope of this paper, but in lieu of those analyses, I consider another outcome: net patient revenue.

Net patient revenue is the difference between total patient revenue and allowances and discounts on patients accounts, which includes provision for bad debts, contractual adjustments,

Figure 3.3: Event Study Results: Uncompensated Care Costs



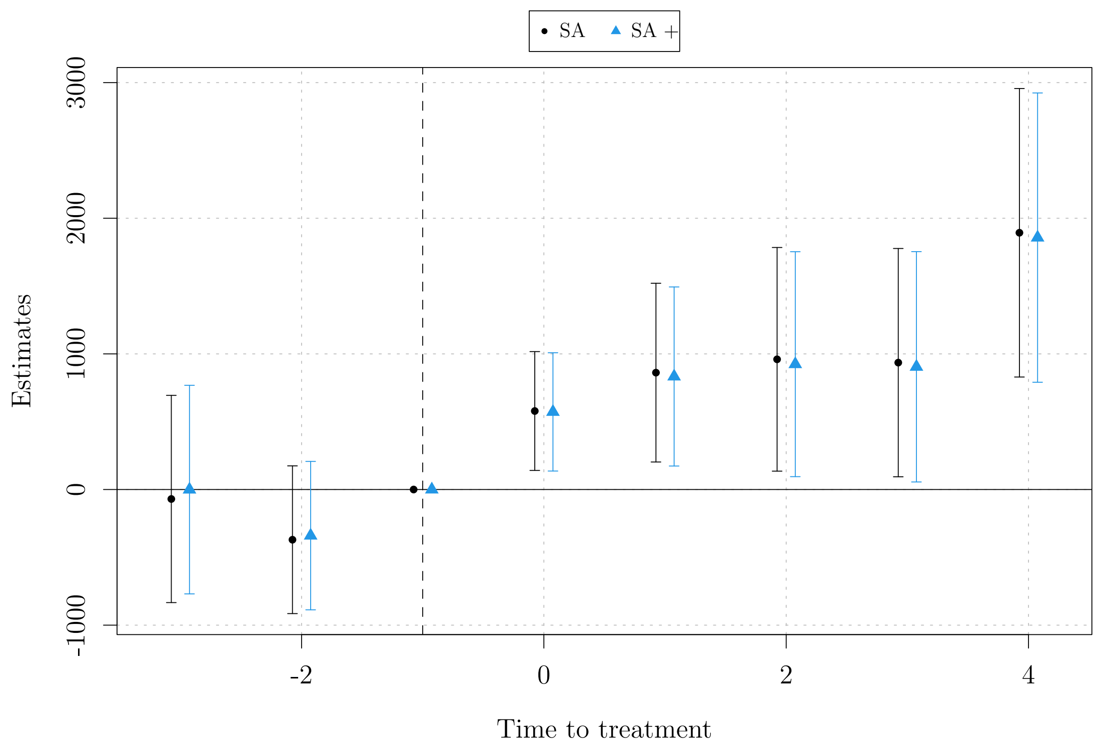
NOTES: Uncompensated care costs are measured in thousands, meaning that the coefficient estimates represent decreases in the millions. The second result at each relative time includes covariates, which are indicated by the "+" following the model title in the legend. I omit observations for relative time −4 because there are only two states with observations in that relative time.

charity discounts, teaching allowances, policy discounts, administrative adjustments, and other deductions from revenue. This outcome isolates the effect of changes in reimbursements that hospitals received and best uncovers changes to a hospital's financial position that would likely facilitate increased investment in non-clinical quality.[8] Figure 3.4 presents the results for net patient revenue per admission. In the year of treatment, expansion state hospitals had a $579 increase in net revenue per patient and in the fourth year after treatment, this increase climbs to $1,893. On average, a hospital in my data set had a total of

[8]Another potentially useful outcome would be "Net Income from Service to Patients," which is net patient revenue less total operating expenses. The challenge, however, is that total operating expenses might encompass expenditures on patient experience that hospitals took on in light of higher revenue.

10,543 admissions annually, therefore, the time-zero point estimate translates roughly into a relative increase of \$6.1 million. This value is, of course, increasing in the coefficient estimates, but given that I expect this revenue increase to affect patient experience through quality investments, it is most relevant to consider the earlier effects.

Figure 3.4: Event Study Results: Net Patient Revenue



NOTES: The net patient revenue value is scaled by total hospital admissions, i.e., the outcome is average net patient revenue per admission. The second result at each relative time includes covariates, which are indicated by the "+" following the model title in the legend. Note that because the outcome is the average net patient revenue per admission, I do not include total admissions on the right-hand side. I omit observations for relative time -4 because there are only two states with observations in that relative time.

3.4.2 Patient Experience Results

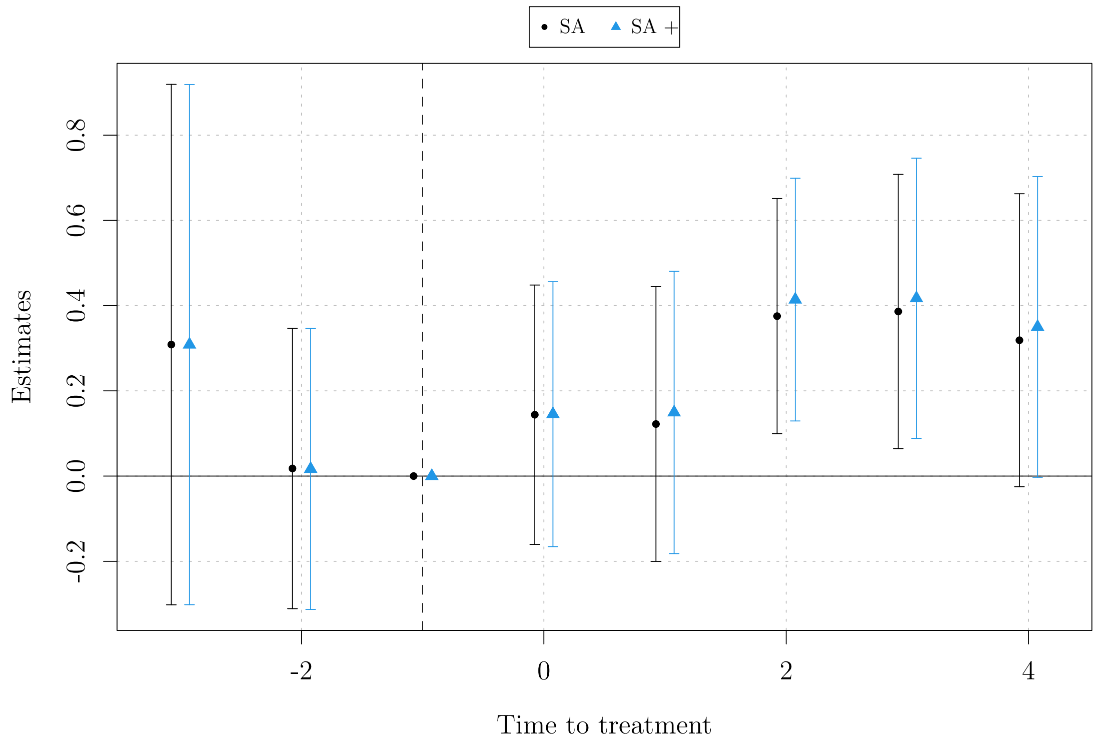
The three first order results provide compelling evidence that the policy change affected hospital operating environments. This suggests a heightened ability for hospitals to make quality investments, potentially in dimensions of care that drive patient experience. I explore this possibility using the hospital Yelp data. The underlying argument is that if hospitals invested in dimensions of quality that are meaningful to patients, then the reviews from patients that received care in expansion state hospitals are likely to reflect a higher level of satisfaction.

I first estimate the event study models using the star rating as the outcome. Figure 3.5 presents the results. In relative times zero and one, I find null effects. However, starting two years post-expansion, hospitals in expansion states have average star ratings that are 0.3 to 0.4 stars higher than their counterparts in non-expansion states. This result coincides with the strategic decision making that one would expect based on existing research, but stops short of clarifying the specific features of care in which hospitals may have differentially invested. To analyze this effect, I use natural language processing techniques on the review text to construct a measure for the presence of non-clinical quality; using this as an outcome, the analysis can uncover differential investment behavior between treated and control groups.

I structure this measure by first preparing the text data to identify which reviews include words that might indicate discussion of non-clinical quality.[9] I create a list of 133 non-clinical quality words that appear in the reviews; nearly 60% of the reviews mention at least one of these words. I present the full list of non-clinical quality words in the supplemental material, but to illustrate, some of the most commonly mentioned words are 'room,' 'food,' 'bed,' 'cafeteria,' 'parking,' 'facility,' 'valet,' and 'bathroom.' However, simply mentioning one of these words does not indicate that the hospital has that quality feature. For example, a

[9] The supplemental material details the steps used to clean the text data. This consists of removing stop words and fixing misspellings.

Figure 3.5: Event Study Results: Star Ratings



NOTES: The second result at each relative time includes covariates, which are indicated by the "+" following the model title in the legend. I omit observations for relative time -4 because there are only two observations in that relative time.

reviewer could say, "the beds here are old and uncomfortable." To avoid considering these mentions in my outcome variable, I impose two additional limitations.

First, I use sentiment analysis to assign a polarity value to each review. The polarity value ranges from -1 to 1, where -1 is entirely negative and 1 is entirely positive. This identifies which reviews are at least marginally better than neutral, i.e. the polarity is greater than zero. Second, I must deal with the fact that there are cases in which the reviewer assigns a 1-star rating, but the review is classified as positive. These represent clear mismatches between the patient's experience and the way that the sentiment analysis categorizes the review. I therefore limit my outcome to the hospitals with ratings greater than 1. With these caveats in mind, I structure my outcome of interest, which is an indicator variable equal to one if the review mentions one of the 133 non-clinical quality words, has a polarity value greater than zero, and does not have a 1-star rating. A total of 47.3% of reviews meet these criteria. This serves as a proxy for whether or not features of non-clinical quality are mentioned positively in the reviews.

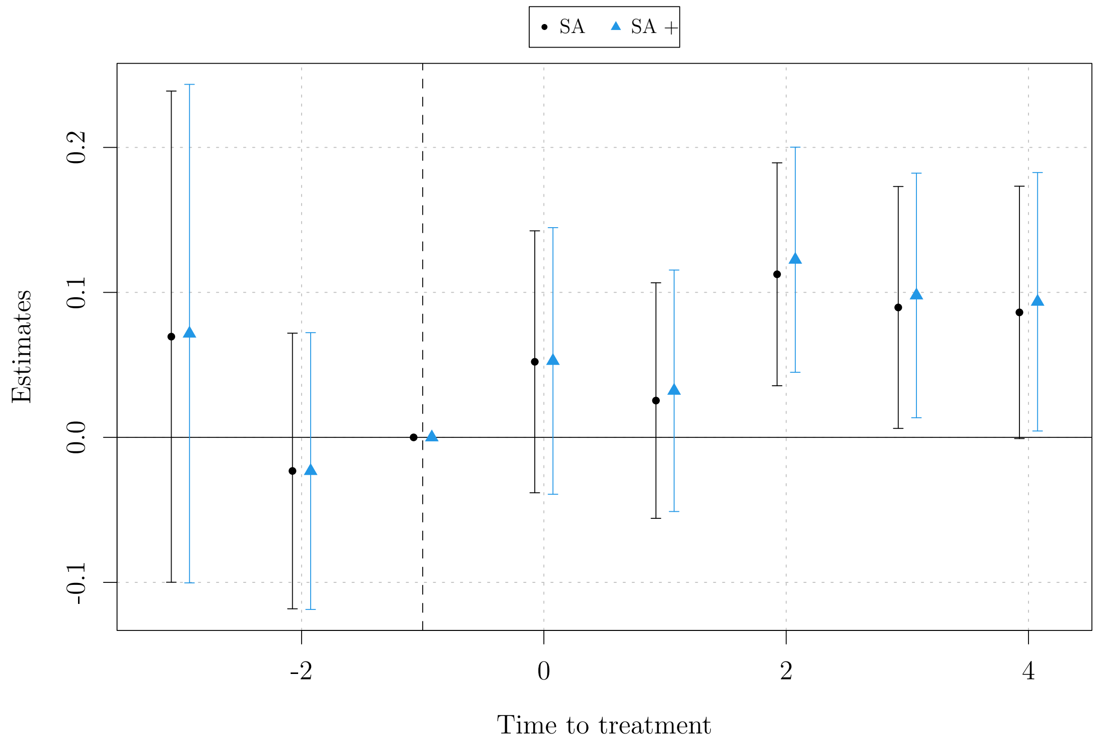
The results with this non-clinical quality indicator outcome are shown in Figure 3.6. Two years after expansion, the results show that hospitals in expansion states are more likely to have amenities mentioned positively in their reviews than hospitals in non-expansion states. This result provides further clarity to the idea that hospitals may be using this exogenous financial shock to increase investment in patient-centered dimensions of care.

3.5 Discussion

This study provides valuable new information to the literature on patient satisfaction and non-clinical quality. Using online review data, I show that hospitals in expansion states improved patient experience as measured by star ratings. Using natural language processing tools and the review text, I show that following Medicaid expansion, reviews in treated states were more likely to positively mention non-clinical features of quality. This suggests that expansion state hospitals differentially increased investment in these dimensions of care.

These results provide a foundation for subsequent research on how these investments relate to productive efficiency. Do investments in non-clinical quality have affect clinical quality and patient outcomes? While this analysis shows that hospitals make strategic investments

Figure 3.6: Event Study Results: Non-Clinical Quality



NOTES: The second result at each relative time includes covariates, which are indicated by the "+" following the model title in the legend. I omit observations for relative time −4 because there are only two observations in that relative time.

in non-clinical quality elements, it is unable to discern if that was at the expense of clinical quality. Theoretically, if hospitals are already producing an efficient combination of clinical and non-clinical quality, then any increase in spending toward one type of quality will come at the expense of the other. If, however, hospitals are not already producing efficiently, then an increase in non-clinical quality may not affect—or may even improve—clinical quality. This remains an open empirical question with important policy implications.

3.6 Appendix

3.6.1 Yelp Data

A detailed description of the steps used to collect and conduct initial cleaning of the Yelp data are contained in Chapters 1 and 2 of this dissertation. I reserve this section for the explicit discussion of the process used to convert the raw review text into a set of indicators to capture elements of non-clinical quality at a given hospital. Using the following steps, I clean the data so that the computer can identify the possible non-clinical quality words within the reviews.

1. I first remove punctuation, numbers, white space, and extraneous letters such as Ã, Â, ã, and â that appear in the process of scraping the reviews. This takes the unadulterated block of review text and returns a cleaner block of text.

2. Next, I take the cleaned block of text and break it down into the individual words through a process known as "tokenizing." This makes it so that the computer can assess each term individually, independent of the surrounding text.

3. I then "lemmatize" each of these tokens, which takes the word down to its simplest form. For example, it takes the various conjugations of a verb and then returns the root word.

4. After lemmatizing, I drop any stop words, such as articles and prepositions, which appear frequently but do not provide any content.

5. Finally, I run a spell check process on the lemmatized words to ensure. Given that

these reviews are posted but not edited, they frequently contain misspellings. By fixing the misspellings, I make the remaining processes cleaner.

After conducting these five steps, I am left with a vector of words corresponding to each review. I use this to create a data frame where the columns are the words mentioned across all reviews and each row represents a review. The data frame is filled with zeros and ones, where ones indicate that the review contained that word.

I take a subset of this data frame that is limited to the columns that refer to the non-clinical quality words listed in the following subsection.

3.6.2 Non-Clinical Quality Words

These are the words used identify which reviews speak to the presence of non-clinical quality.

aesthetic	amenities	amenity	architecture
aromatherapy	arrangement	art	artwork
bakery	bath	bathroom	bathrooms
bathtub	bathtubs	bed	bedding
bedroom	beverage	bistro	blanket
boutique	breakfast	building	cable
caf	cafe	cafeteria	cafeterias
campus	chair	champagne	charm
coffee	coffeeshop	comfort	comforter
computer	concierge	convenience	couch
courtyard	decor	decorate	decoration
deli	design	dessert	dining
dinner	dinnerware	drink	drinks
eatery	enjoyment	facilities	facility
fashion	flavor	flower	food
fountain	fountains	furnishing	furniture
garden	heaven	housekeeper	housekeeping
interior	ipad	jacuzzi	lobby
loveseat	massage	meal	music
musician	netflix	news	oasis
park	parking	performance	photo
photographer	photography	photos	pianist
piano	picture	pillow	pillowcase
plant	playroom	plaza	refreshment
renovation	resort	restaurant	restroom
room	rooms	salon	shower
showering	singer	singing	snack
sofa	soundscapes	spa	starbucks
statue	statues	store	studio

style	suite	taste	telephone
television	terrace	toilet	toiletry
towel	toy	treat	upgrade
valet	waterfall	wifi	window
windows	wireless		

3.6.3 HCRIS Data

I use the the HCRIS data for measures of uncompensated care costs and net patient revenue. For uncompensated care costs, I follow Blavin (2016), which provides insights into matching this value across the 1996 and 2010 versions of the Cost Reports. I first pull the uncompensated care *charges* value. For the 1996 form, this is Worksheet S-10 line 30 (total non-Medicaid, SCHIP, or state and local indigent program charity care and total hospital bad debt charges). For the 2010 form, this is Worksheet S-10, line 20, column 3 (total charity care charges for uninsured and insured patients approved for charity care) minus Worksheet S-10, line 22, column 3 (partial payments by uninsured and insured patients approved for charity care) plus Worksheet S-10, line 26 (total hospital bad debt across all patient types). Then to deflate these uncompensated care *charges* to uncompensated care *costs*, I multiply this by the cost-to-charge ratio. For the 1996 form, this is Worksheet S-10, line 24, column 1, and for the 2010 form, it is the same worksheet and column, but instead, line 1. Locating the "net patient revenue" value is simpler. This is located on Worksheet G-3, line 3, column 1 in both the 1996 and 2010 versions of the form.

3.6.4 Alternative Results

As noted in the main text, my main results are based on the Sun and Abraham (2021) estimator, which I denote as "SA" in the following graphs. This estimator offers an improvement upon the traditional two-way fixed effects estimator (TWFE). For the sake of comparison, here I outline the TWFE specification and present those results next to the main (SA) results. The TWFE specification for the first stage results is:

$$Y_{ht} = \sum_{\substack{l=-3 \\ l\neq -1}}^{l=4} \beta_l D^l_{ht} + \gamma_t + \gamma_h + \varepsilon_{ht}. \tag{3.3}$$

On the right hand side of Equation 3.3, the term D^l_{ht} is defined as $\mathbb{I}\{t - E_h = l\}$, i.e., an indicator for hospital h being l years away from state-level Medicaid expansion at year t.[10] This term is zero for all hospitals in non-expansion states. I omit the year prior to expansion, i.e. $y = -1$. The terms γ_t and γ_h denote year and hospital fixed effects. I conduct the analysis using hospital-year level data. Lastly, in some results, I include a term, X_{ht}, which is a vector that contains as covariates the hospital characteristics.

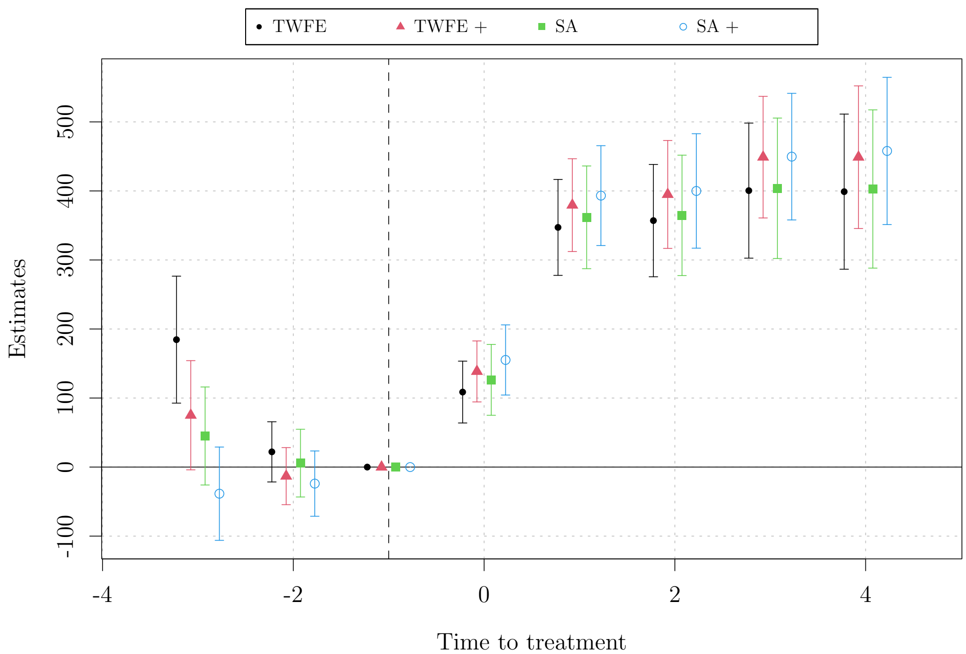
The TWFE specification for the patient satisfaction is as follows:

$$Y_{iht} = \sum_{\substack{l=-3 \\ l\neq -1}}^{l=4} \beta_l D^l_{ht} + \gamma_t + \gamma_h + \varepsilon_{iht}. \tag{3.4}$$

The following plots present results that are analogous to those in the main text, but using the TWFE specification defined above.

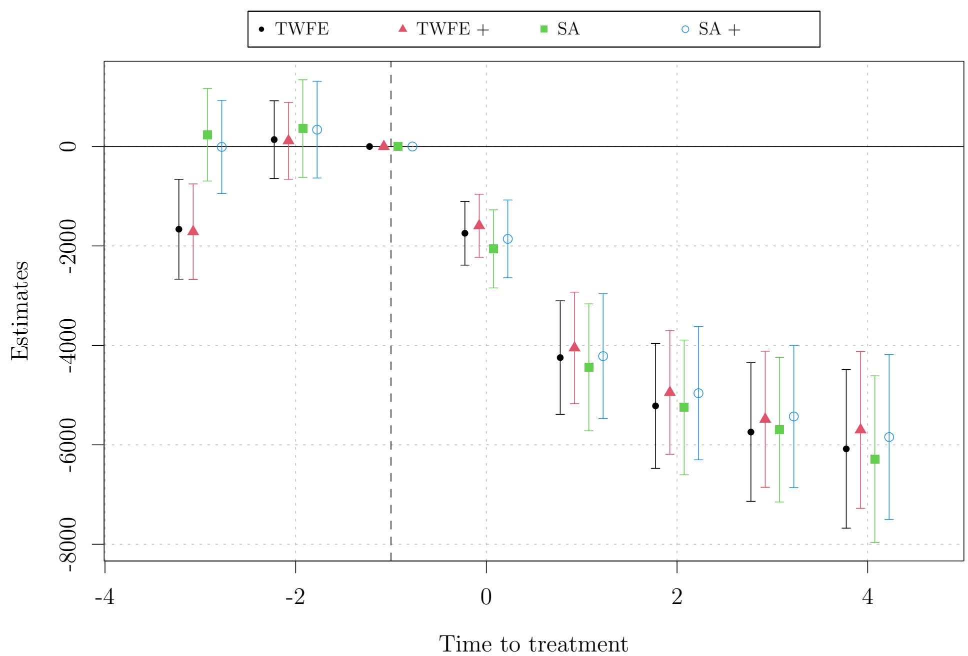
[10] I borrow this notation directly from Sun and Abraham (2021) for ease of comparison between this and the subsequent model.

Figure 3.7: Event Study Results: Medicaid Discharges



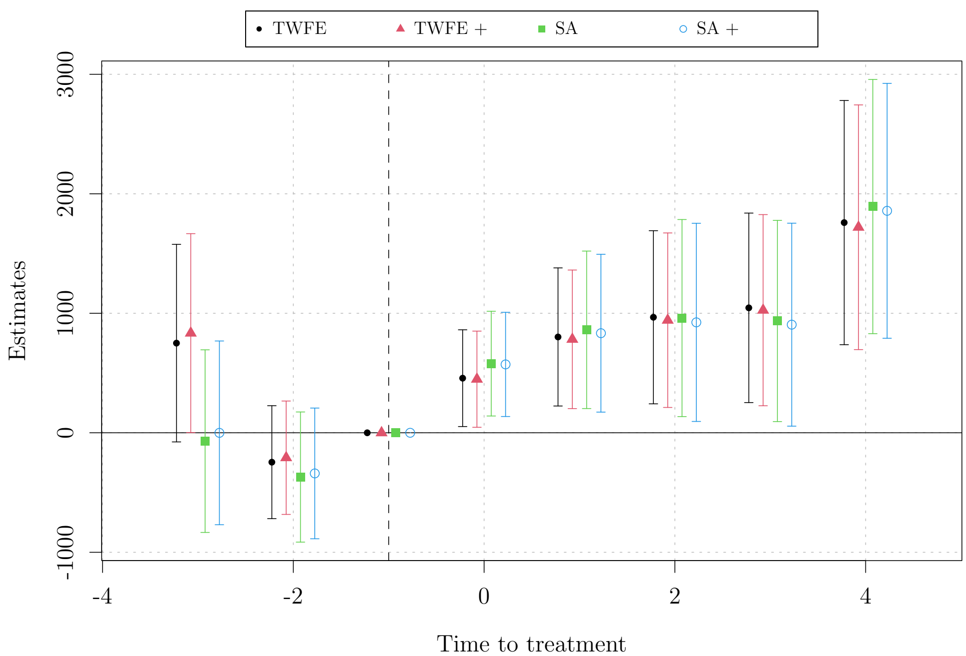
NOTES: The uncompensated care costs enter the regression in thousands, meaning that the coefficient estimates indicate decreases in the millions. The second and fourth result at each relative time includes covariates, which are indicated by the "+" following the model title in the legend. I omit observations for relative time -4 because there are only two states with observations in that relative time.

Figure 3.8: Event Study Results: Uncompensated Care Costs



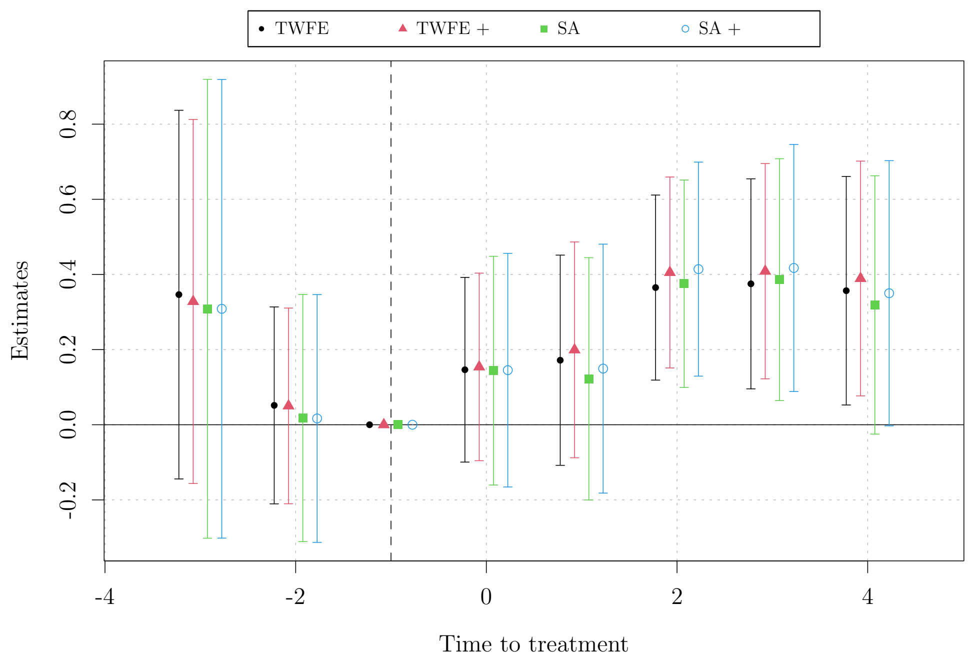
Notes: Uncompensated care costs are measured in thousands, meaning that the coefficient estimates represent decreases in the millions. The second and fourth result at each relative time includes covariates, which are indicated by the "+" following the model title in the legend. I omit observations for relative time -4 because there are only two states with observations in that relative time.

Figure 3.9: Event Study Results: Net Patient Revenue



Notes: The net patient revenue value is scaled by total hospital admissions, i.e., the outcome is average net patient revenue per admission. The second and fourth result at each relative time includes covariates, which are indicated by the “+” following the model title in the legend. Note that because the outcome is the average net patient revenue per admission, I do not include total admissions on the right-hand side. I omit observations for relative time -4 because there are only two states with observations in that relative time.

Figure 3.10: Event Study Results: Star Ratings



NOTES: The second and fourth result at each relative time includes covariates, which are indicated by the "+" following the model title in the legend. I omit observations for relative time −4 because there are only two observations in that relative time.

Bibliography

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.

Blavin F. 2016. Association between the 2014 medicaid expansion and us hospital finances. *Jama* **316**: 1475–1483.

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.

Callaway B, Sant'Anna PH. 2021. Difference-in-differences with multiple time periods. *Journal of Econometrics* **225**: 200–230.

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.

Chen Y. 2018. User-Generated Physician Ratings — Evidence from Yelp.

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.

Chevalier JA, Mayzlin D. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* **43**: 345–354.

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.

Desbordes R, Verardi V. 2012. A robust instrumental-variables estimator. *The Stata Journal* **12**: 169–181.

Dewan S, Hsu V. 2004. Adverse selection in electronic markets: Evidence from online stamp auctions. *The Journal of Industrial Economics* **52**: 497–516.

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

Gaynor MS, Kleiner SA, Vogt WB. 2013. A structural approach to market definition with an application to the hospital industry. *The Journal of Industrial Economics* **61**: 243–289.

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.

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.

Luca M. 2016. Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper* .

Luca M, Zervas G. 2016. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science* **62**: 3412–3427.

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.

Miller S, Johnson N, Wherry LR. 2021. Medicaid and mortality: new evidence from linked survey and administrative data. *The Quarterly Journal of Economics* **136**: 1783–1829.

Moghtaderi A, Pines J, Zocchi M, Black B. 2020. The effect of affordable care act medicaid expansion on hospital revenue. *Health Economics* .

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.

Nikpay S, Buchmueller T, Levy H. 2015. Early medicaid expansion in connecticut stemmed the growth in hospital uncompensated care. *Health Affairs* **34**: 1170–1179.

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.

Reinstein DA, Snyder CM. 2005. The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics. *The Journal of Industrial Economics* **LIII**: 27–51.

Resnick P, Zeckhauser R, Swanson J, Lockwood K. 2006. The value of reputation on ebay: A controlled experiment. *Experimental economics* **9**: 79–101.

Rhodes JH, Buchmueller TC, Levy HG, Nikpay SS. 2020. Heterogeneous effects of the aca medicaid expansion on hospital financial outcomes. *Contemporary Economic Policy* **38**: 81–93.

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.

Sun L, Abraham S. 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* **225**: 175–199.

Trzeciak S, Gaughan JP, Bosire J, Angelo M, Holzberg AS, Mazzarelli AJ. 2017. Association Between Medicare Star Ratings for Patient Experience and Medicare Spending per Beneficiary for US Hospitals. *Journal of Patient Experience* **4**: 17–21. ISSN 2374-3735.

Trzeciak S, Gaughan JP, Bosire J, Mazzarelli AJ. 2016. Association Between Medicare Summary Star Ratings for Patient Experience and Clinical Outcomes in US Hospitals. *Journal of Patient Experience* **3**: 6–9. ISSN 2374-3735.

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.

Wimmer BS, Chezum B. 2003. An empirical examination of quality certification in a "lemons market". *Economic inquiry* **41**: 279–291.

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

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