Quality Disclosure, Demand, and Congestion: Evidence from Physician Ratings

Benjamin L. Chartock*

Department of Economics, Bentley University

November 3, 2022

Abstract

Ratings provide consumers with useful quality information, however, when ratings shift demand to highly-rated sellers, congestion might occur at the top of the quality distribution. Congestion caused by disclosure may be observed in the health care setting, where prices often cannot adjust to reflect varying quality. I study the trade-off between providing quality information for consumers and congestion using a star rating disclosure policy implemented at a large integrated health system in the United States, which requires every physician to have star ratings posted online in a standardized fashion. I identify the effects of physician star ratings on patient volume using a regression discontinuity and difference-in-discontinuity design which leverages the rounding of ratings to discrete values and the fact that I observe ratings before and after their public disclosure online. I find that an increase in a physician's rating increases the number of new patients seen by 2.96 visits per month on a baseline of 5.48 (54% increase). I show that star ratings shift patients to physicians who more often provide medically recommended screenings, counseling, and vaccinations. However, I also show that a higher rating causes patients to wait longer for treatment. New patients wait 2.7 additional days (30.5% longer) for an additional increment of the rating scale and existing patients wait longer as well. I use these findings to compute a revealed-preference estimate of the "shadow price of a star"; I find that patients are willing to wait 3 additional days in exchange for a one standard deviation increase in physician ratings. In the absence of a price, wait times may serve as an equilibrating factor to clear the market.

*I would like to thank Abby Alpert, Ginger Jin, Claudio Lucarelli, and Mark Pauly for helpful comments and also participants at the Conference of American Society of Health Economists, the Midwest Health Economics Conference, and at Indiana University. I also am grateful for helpful discussion from Ellie Prager. This project is supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number 5P20GM121341. Email: bchartock@bentley.edu.

1 Introduction

Quality disclosure can have a profound impact on market outcomes. On the one hand, quality disclosure has been shown to enhance welfare by increasing demand for high-quality products, stimulating competition, and ameliorating adverse selection (Jin and Leslie, 2003). On the other hand, disclosure might lead to unintended consequences, such as causing multitasking problems (Holmstrom and Milgrom 1991; Feng Lu 2012) or inducing inefficient effort on the part of suppliers (Dranove et al., 2003). Although the literature has numerous studies about the effect of quality disclosure on market outcomes, one understudied domain is the impact of quality disclosure on markets with potential congestion effects and wait times. If quality ratings sort consumers to highly rated sellers whose supply is not perfectly elastic, a glut of buyers may seek to purchase from these high-rated sellers if prices cannot adjust to reflect varying quality. One market where this might occur is in health care, where patients often pay the same price for care from any in-network provider regardless of quality. In the absence of a price, wait times may serve as an equilibrating factor to clear the market.

I study this phenomenon in the market for family medicine physicians. This market is a setting where quality ratings are widespread (e.g., ZocDoc.com and Healthgrades.com) and where many consumers search the internet for information before selecting a provider. According to a 2019 University of Michigan National Poll on Healthy Aging, 43% of adults aged 50-80 reported looking at doctor ratings online (Hanauer et al., 2020). While the market for doctors and other medical providers is not the only setting where star ratings are important (other examples are Amazon for retail products, Yelp for restaurants, or Centers for Medicare and Medicaid Services [CMS] Compare for nursing homes), ratings may be particularly relevant in the market for family medicine and primary care because patients typically have a large number of potential providers to choose from and their insurance benefits often force an active choice of a family doctor. This directly contrasts with the choice of a specialist (e.g., cardiologist), where choice sets are often more limited and another factor—referrals—might crowd out the role of consumer-facing quality information such as star ratings.

In this paper, I focus on three primary economic outcomes: quantity demanded, sorting over quality, and congestion spillovers. These three outcomes encompass a range of possible effects of quality disclosure in equilibrium. I study these effects using a novel data set comprised of a combination of electronic health records (EHRs) and the universe of online doctor reviews that was collected and later disclosed by a large, integrated health system in the United States with more than 40 hospitals and nearly 1,500 employed physicians.

I use a regression discontinuity design to estimate the causal effects of an increase in provider rating on new patient visits which leverages the fact that actual provider quality ratings are continuous but are rounded into discrete bins on the health system website. In the spirit of Anderson and Magruder (2012), I exploit the rounding of online ratings, focusing on doctors just above and

just below the rounding thresholds—these physicians have nearly identical underlying scores but different displayed scores. Additionally, and uniquely among papers in the literature that examine the demand response to ratings data, I exploit the fact that the health system collected ratings long before it ever decided to disclose them to the public. Using this distinctive pre- and post-disclosure variation in the information available to consumers, I estimate a difference-in-discontinuities model to capture the effects of quality disclosure.

This health system and the quality disclosure policy that I study have a number of unique attributes that make the setting an ideal laboratory for exploring the impact of ratings. First, the disclosed ratings are highly salient for consumers in this market. Prominent star ratings for doctors are available in a standardized format and are centrally located on each provider's website (an example is found in Appendix Figure A1). In addition, the manner in which ratings are gathered from patients differs from other well-known online sources so these ratings may be of higher fidelity than other star ratings. Ratings disclosed by this health system are calculated from randomly-sent, post-visit surveys that are designed and implemented in consultation with the Agency for Healthcare Research and Quality (AHRQ). In contrast to this standardized survey, any person (patient or not) is able to submit a review of a provider on Yelp or other sites. The random sending of surveys to patients eliminates much of the selection bias that arises due to which individuals are contributing to online ratings. There is also relatively low availability of other sources of online quality data about medical providers (e.g., from HealthGrades, Zocdoc, and Yelp) in the health system's region, which suggests that this quality disclosure represents a major source of information about providers. Lastly, unlike on other websites, these quality scores apply universally to all providers; no provider can opt out of having their rating disclosed or pay for a more prominent placement.

The unique data source is also an advantage of this setting because it allows me to focus directly on the subset of the population most impacted by star ratings: new patients. Using the EHR data, I can identify which patients in the health care system have never before visited a given provider, allowing me to focus directly on the subset of shoppers who are actively searching for physicians but have not yet received a signal via previous consumption. I use the EHR data to construct a volume measure of new patients at the level of a provider–month, which allows me to zoom in directly on the component of health care shopping that might be most impacted by quality disclosure. These data also allow me to explore heterogeneity in the effect of quality disclosure across different provider specialties. This approach is important due to the nature of insurance design. Plans such as health maintenance organizations (HMOs) frequently force members to make active choices about their family medicine providers. These chosen primary care doctors act as gatekeepers via referrals to specialists. Accordingly, I focus on family medicine as the subset of providers who might be most impacted by quality disclosure, but also analyze the effects separately for different types of specialists.

There are several interesting results. First, I find that consumer demand is highly responsive to online digital disclosure of quality scores. In particular, I find that an increase of one interval in the

rating scale in a provider's online profile causes them to see 54% more new patients per month (2.96 new patients). This result is consistent with a number of other studies about the demand response to online disclosure of ratings (Chevalier and Mayzlin 2006; Anderson and Magruder 2012; Hunter 2020). However, I obtain estimates that are larger in magnitude. This can likely be explained by the standardized nature of quality disclosure in this setting and potentially by the paucity of other reputable sources of physician quality information. Second, I find that the effect of quality disclosure is concentrated among family medicine providers (as opposed to other specialties), highlighting the role of referrals in consumer choice of specialist providers. Family medicine doctors are selected from a large choice set relative to other specialists, and it is not surprising that the effects of quality disclosure are large for these gatekeepers of patient health. I also find that the effect of quality disclosure is greatest among the younger population (ages 18-34) as compared to older individuals, potentially because this age group is more accustomed to searching online about product quality more generally. Previous literature has been unable to examine heterogeneity in ratings effect by age.

In addition to these findings about the demand response to quality disclosure, I provide evidence on equilibrium effects. Specifically, I examine equilibrium consequences of disclosure on supply and demand by studying three dimensions of sorting: (1) examining whether information disclosure shifts patients to physicians who supply greater inputs to health, (2) studying whether information disclosure results in market expansion (new patients to the system) or switching (reallocation of existing patients), and (3) investigating whether quality disclosure causes congestion at high-quality sellers.

I first examine whether information shifts patients to physicians who supply greater inputs to health. One common criticism of online disclosure of doctor quality ratings is that stars do not reflect actual provider quality but instead reflect orthogonal concerns such as the quality of the magazines in the waiting room lobby. In contrast to these concerns, I show evidence that the online disclosure sorts patients to providers who more frequently perform medically-recommended inputs to patient health such as vaccinations, screenings, and behavioral health services. Second, I study whether the quality disclosure has market expansion effects or switches existing patients (or both). I find that the quality disclosure largely switches current patients at the health system to higher-rated providers rather than affecting choices of individuals who have never before visited the system, thus suggesting the main margin of action is that disclosure effects established patients in the system.

Finally, I address a previously understudied question about congestion and wait time that is relevant in markets such as health care where prices cannot easily adjust in response to quality scores being released. In contrast to restaurants, for example, which can raise prices in response to an increase in consumer demand, physicians employed by a health system cannot immediately raise prices after receiving a high score (or cannot lower prices after receiving a low score). In this health system, the patient pays the same out-of-pocket price for family medicine irrespective of quality. If

a significant mass of new patients is shifted towards the high-quality sellers after quality disclosure, those sellers will face congestion in the absence of a monetary price which rations the scarce quality (Richards-Shubik et al., 2021). I document that congestion is occurring at high-quality sellers, and that this congestion is affecting both new patients (who wait 30.5% longer for an additional increment of quality score) as well as established patients, who were previously seeing a high-quality provider but now wait longer for appointments *with the exact same provider* due to congestion. This finding helps underscore the winners and losers of quality disclosure and provides the first revealed preference evidence of a willingness-to-pay for provider stars. I calculate that patients are willing to wait 3 additional days for a one standard deviation increase in provider quality, and this wait time serves as a shadow price for quality which rations demand at high-quality sellers.

Taken as a whole, these results paint a multidimensional picture of the economic consequences of online quality disclosure. As markets in health care and beyond increasingly adopt star ratings and quality certification as a means to ameliorate market woes caused by imperfect information, they will face trade-offs between increased ease of shopping for experience goods and congestion at high-quality sellers. The theoretical model which I introduce along with the empirical evidence I uncover suggests that quality disclosure creates a new market for quality even in the absence of differential prices, as wait times can serve as an equilibrating force. This insight is useful for policymakers who are interested in designing, implementing, and evaluating quality disclosure policies, such as those at the Centers for Medicare and Medicaid Services (CMS), because it suggests that increased wait times for highly rated physicians may reflect a market-driven process in the absence of potential capacity adjustments and price variation.

The rest of this paper proceeds as follows. Section 2 reviews the existing literature. Section 3 lays out a model of patient choice and waiting. Section 4 describes the data and institutional setting and Section 5 presents the empirical strategy. Section 6 presents the results, discusses heterogeneity, and institutes robustness checks. Section 7 concludes.

2 Related Literature

In this section, I broadly separate the literature about quality disclosure into two components, the demand-side response and the supply-side response. Before summarizing the literature, I offer a brief introduction to the theory on how incomplete information can cause market failures, a key problem that disclosure policies hope to remedy. A comprehensive review of the economics of disclosure can be found in a survey article by Dranove and Jin (2010).

2.1 Information, Market Failures, and Disclosure

Studies on the relationship between and market outcomes emerged shortly after the development of general equilibrium theory. A finding of general equilibrium theory, the first fundamental welfare

theorem, holds that under a certain standard set of assumptions, such as well-behaved preferences and perfectly competitive markets, the competitive market equilibrium will be Pareto efficient in that it will exhaust all gains from trade. However, an important condition of the welfare theorem that must hold for the results to obtain is the assumption of perfect information. Suggesting that the market for medical care falls short on this dimension, Arrow (1963, p. 951) writes, "uncertainty as to the quality of the product is perhaps more intense here [medical care] than in any other important commodity."

Akerlof (1970) proves that in markets featuring asymmetric information, less-than-efficient levels of trade might occur if one side of the market sorts on quality and the other side cannot readily observe quality *ex ante* but nonetheless knows that low-quality goods will be put on the market first. This creates an adverse selection problem, with consumers who are wary of poor quality products, or "lemons". For disclosure to ameliorate adverse selection, disclosed quality information must be (1) noticed by the market participants and (2) acted upon. Nelson (1970) introduces a useful taxonomy of search goods versus experience goods, with search goods allowing consumers to inspect products for quality prior to consumption while experience goods requiring consumers to learn about quality after purchase. In addition to search versus experience goods, a credence good is a product for which quality may not be observable by the consumer until long after consumption, if ever. The market studied in this paper, physician services, has elements of search, experience, as well as credence goods. Broadly speaking, the more information that is available *ex ante* in a market with quality heterogeneity, the more the product is similar to a search good than a credence or experience good, and disclosure can be used as a lever to moderate if a good is search, experience, or credence type.

The economics literature draws a distinction between voluntary and mandatory information disclosure. With mandatory disclosure, all sellers must post or publish quality information. With voluntary disclosure, it is ambiguous whether firms will choose to disclose the quality of their offering. The theory literature (Grossman and Hart 1980; Grossman 1981; Jovanovic 1982) finds that when disclosure is costless and verifiably truthful, all sellers should voluntarily disclose quality because consumers assume if a seller does not disclose, that seller is low-quality. When disclosure is costly, only sellers with sufficiently high quality should choose to disclose (Jovanovic, 1982). In contrast, Bederson et al. (2018) find that voluntary disclosure might not occur by high-quality sellers due to counter-signaling; in essence, high quality sellers choose not to disclose their quality, sending a signal to buyers that they are such high quality that they do not need to disclose. In the setting studied in this paper, disclosure is mandated throughout the health system, and questions about voluntary disclosure are not applicable.

2.2 Demand-Side Responses to Disclosure

Until fairly recently, there was very little empirical evidence that consumers observe and act upon disclosed quality information. A paper by Mathios (2000) finds that when the Nutrition Labeling

and Education Act required disclosure of fat content on salad dressings, high-fat dressings experienced a significant reduction in sales. Chevalier and Mayzlin (2006), focusing on online reviews, also find that consumers are responsive to disclosure. The authors looked at the same book that sold on both Amazon.com and BarnesAndNoble.com and found that books with a higher review score on one site had higher sales on that same site. By focusing on the same exact book at two online retailers, they cleverly controlled for actual quality of the product.

To measure the effect of information disclosure, most studies rely on panel data methods. For example, in a wide variety of health care contexts, the literature shows that consumers are responsive to disclosure in the form of report card ratings. Studying health plans, Scanlon et al. (2002) show that people avoid health plans with many below-average ratings. The authors controlled for fixed, unobserved plan traits by leveraging a natural experiment when General Motors released plan report cards. Dafny and Dranove (2008) study Medicare HMO report card disclosure and find that consumers switch to high-quality plans independently of report cards (driven by word-of-mouth information), but also that disclosure induces a response to satisfaction scores. This effect is larger when there is large variation in quality. Demand-side responses to quality report cards are shown to occur for hospitals (Dranove and Sfekas 2008; Pope 2009), fertility clinics (Bundorf et al., 2009), and (in a stated preferences experiment) for joint replacement practices (Schwartz et al., 2021).

Identifying the effect of information disclosure on demand-side decisions is complicated by the fact that in almost all settings, rating scores (which are observable to the researcher) are correlated with other factors that are unobservable to the researcher, but observable to the economic agents. One such example is word-of-mouth reputation. These unobserved factors will cause biased estimates in the cross-section, and estimates of the ratings effect on demand will be overstated if publicized ratings are positively correlated with unobservable factors. Of course, the bias could run in the opposite direction, too (e.g., if provider panels are limited in size and high quality providers are full, a form of capacity constraint). Jin and Sorensen (2006) address the omitted variable bias by assessing the demand response to health plan rating disclosure from the National Committee for Quality Assurance, exploiting a data set that includes both disclosed ratings as well as non-public plan ratings. They find that ratings have an effect on patient choice, particularly for first-time decisionmakers. Disclosed information affects only a small number of individuals, but the welfare gains for those individuals are large. The similarities between the Jin and Sorensen study and my research include the presence of both public and non-public ratings data as well as the importance of first-time decisionmakers (in my context, new patients) versus established consumers. Jin and Sorensen also develop a discrete choice framework for estimating the value of information as a function of estimated parameters. Chernew et al. (2008) studied a similar setting of health plan report cards and found a small but significant effect of information on plan choices (average value of a report card to employees was about $20 per year). In contrast to Jin and Sorensen (2006), Chernew et al. (2008) specified a Bayesian learning model to quantify the value of information. They assume patients hold priors about the distribution of quality and update to form a posterior proportional to the prior times the likelihood. They allowed for both continuous or discrete priors

and signals, with discreetness reflecting real-world disclosure methods such as stars.

Another approach to identifying the causal impact of ratings on demand is a regression discontinuity design first initiated by Anderson and Magruder (2012) in a study looking at the market for restaurant services in the context of Yelp.com. The authors find that increasing a restaurant's Yelp score by a half-star (the smallest increment displayed on the website) causes restaurants in their study sample to sell out 19 percentage points more frequently compared to a restaurant without the benefit of a higher rating. That paper not only provides credible estimates of demand effects of ratings in the food service industry, but is also notable for introducing a novel application of regression discontinuity design for the purposes of identifying the effect of ratings on quantity demanded. The authors point out that the underlying distribution of actual, raw ratings for restaurants is continuous, yet the website displays ratings only in discrete, rounded bins. Leveraging this rounding, which is widespread in internet-based rating systems, they used the mass of restaurants just below and just above the rounding cutoff thresholds to identify the causal effect of an increased score on volume, laying the groundwork for the identification strategy used in this paper.

Anderson and Magruder's regression discontinuity design has been applied to a variety of settings where credence and/or experience goods are bought and sold. Some of this has been in the context of health care, where physician quality is heterogeneous and difficult to discern ex ante. For example, in an unpublished manuscript, Luca and Vats (2013) collect ratings from a crowdsourced online doctor platform (ZocDoc) and find that a half-star improvement in a doctor's rating boosts the likelihood that the doctor will have an appointment booked through ZocDoc by 10%. A drawback to this study is that provider participation on ZocDoc is voluntary as opposed to mandatory (in my paper, ratings are required for all doctors in the system). Providers on ZocDoc can additionally choose to pay a subscription to achieve a "verified" status and optimal placement on the webpage, suggesting that there may be unobserved selection into prominent disclosure. In another unpublished manuscript, Brown, Hansman, Keener, and Veiga (2022) look at General Practice (GP) clinics in the English National Health Service (NHS) and find that a half-star improvement for a GP practice increases quarterly enrollment in the practice by 0.13 patients, an implied increase in enrollment of 22%. The Brown et al. (2022) paper is the study most similar to mine. Some important differences, however, relate to the setting. Brown et al. study the causal impact of star ratings in the market for GP practices in England whereas this paper studies the market for doctors and other providers in the United States. The English NHS and the United States health systems differ substantially with respect to autonomy of patient choice at all levels of the health system. For example, GPs in England operate according to geographic catchment areas and only since 2015 have patients who live outside of a GP's practice area been allowed to register with that GP. And health care in Great Britain is marked by long waiting times and failure to provide certain types of treatments (Feldstein, 2007). Furthermore, the GP practices in the Brown et al. paper have an average of 5.9 practitioners per practice, so ratings are not specific to individual providers, while my study focuses on individual providers rather than practices.

The effect of ratings on demand is not limited to health care and restaurants. Hunter (2020) finds that demand for automotive repair services is responsive to online star ratings, and Magnusson (2019) finds that increasing a home furnishing product rating by a half star on Wayfair.com leads to a 5% increase in demand for that item. Both papers use the regression discontinuity from rounded ratings to identify the causal effect.

2.3 Supply-Side Responses to Disclosure

In addition to the demand-side response to quality disclosure, supply-side responses also may have an effect on market performance. Jin and Leslie (2003) find that disclosure of restaurant report cards causes firms to improve product quality. The authors show that restaurants obtaining an "A" relative to a "B" grade causes restaurants to have 5% greater revenue, but also that grade cards cause a 20% decrease in foodborne illness hospitalizations, a decrease not fully explained by consumers switching from low to high hygiene restaurants. This implies that disclosure causes firms to increase quality, a fact that they attribute to reducing adverse selection via disclosure.

However, Dranove et al. (2003) observe that disclosure can have countervailing effects which may be welfare-reducing. Using a difference-in-difference design in a study of heart attack patients and coronary artery bypass graft (CABG) surgeries, the authors found that report cards improved matching of patients to hospitals, increased the amount of CABG surgeries, and shifted this treatment from ex ante sicker to ex ante healthier patients, who derive less of a benefit from the more intense CABG procedure, resulting in higher costs and worse outcomes. On net, the authors conclude that report card disclosure caused doctors to change behavior in a welfare-reducing way. Similar unintended consequences are highlighted by Werner and Asch (2005).

A major concern is that disclosing ratings might incentivize suboptimal behavior on the part of sellers, particularly when quality is multidimensional. Building off the multitasking literature of Holmstrom and Milgrom (1991), Feng Lu (2012) finds that an initiative to report nursing home quality data that discloses some product attributes but not others had the effect of realigning the relative returns across different quality dimensions, leading to improvement on reported quality dimensions but deterioration along other dimensions. Given that patient demand was responsive to this disclosure, the reallocation of effort across tasks might reduce welfare if there is large misalignment between the social planner's objectives and what can be measured (Baker 2002; Gibbons 2010). In the context of a health system disclosing aggregate survey ratings for each doctor, if ratings reflect different quality attributes than what patients actually desire, disclosure could be harmful. In the context of credence goods, where the consumer might have difficulty assessing quality, this problem might be particularly severe. For example, if a patient values medical care and amenities, but faces challenges in observing the medical skill of a doctor, that patient might rate the provider based on only amenities (such as the magazines in the waiting room) and be unable to opine on other elements that enter into their utility. This situation creates a rating score that is misaligned with provider quality. As observed by Baker (2002), the misalignment between

what can be measured by scores and what is valued by consumers may inhibit success of a disclosure policy such as doctor ratings.

Finally, Kolstad (2013) found that cardiologists, when faced with report card disclosure, responded to both financial and non-financial (intrinsic motivation) incentives to increase quality. Using the risk-adjustment model that underlies report cards, Kolstad identified the magnitude of the effect of new information by exploiting the fact that different surgeons gain more or less information about their relative performance compared to substitute surgeons. He concluded that not only does profit motivate reductions in relative average mortality risk, but intrinsic non-pecuniary motivations are relatively large. This result implies that in a model with no immediate differentiation on prices, sellers may still respond to information disclosure because of non-financial determinants of provider utility.

Richards-Shubik et al. (2021) point out that, in equilibrium, prices serve to ration quality when quality is scarce, and in the absence of prices for quality (which may be the case in health care), congestion serves the role of equilibrating the market. They discuss the bias that can result from estimating models of consumer demand that include taste for quality but do not account for disutility from congestion. Studying the market for heart surgery, they found that this bias can be empirically large.

I next present a model about the equilibrium effects of disclosure and turn to the institutional setting studied in this paper.

3 Rationing Demand by Wait List: A Theoretical Model

In this section, I introduce a theoretical model which ties together two related empirical observations that I observe in the data (that demand is responsive to star ratings and that a higher star rating causes a longer wait times, *ceterus paribus*). This model is inspired by Lindsay and Feigenbaum (1984) and introduces a way in which wait times function very much like a price and clear the market when prices are absent.[1] A key feature of the model is attacking the assumption that demand for care is unchanged throughout the wait (Cullis and Jones, 1985) and I link wait time to demand by recognizing that the value of care decays the longer care is postponed. For example, a high-quality doctor might refer a patient with coronavirus symptoms to get monoclonal antibodies, which are helpful if given early but which decay in effectiveness the longer the duration between illness and infusion, whereas a low quality doctor might not refer a patient for monoclonal antibodies at all.

The insight of the model's equilibrium conditions derives from the idea that wait times equilibrate a queue by rising or falling until the number of individuals who join the queue is equal to the number of patients who get treatment in a given time period. I first start by modeling the marginal joiner of a queue.

[1] This intuition of this model is used extensively in the study of the National Health Service in the United Kingdom, where wait lists for elective surgeries are frequent. See Cullis et al. (2000) and Propper (2000), for example.

3.1 Marginal Joiner of a Queue

I assume that patients who are seeking care from a highly-rated family medicine physician might not be able to see that physician right away. The fundamental economic decision faced by the patient when they need care is whether to join the queue and wait to see the highly-rated physician or not. The patient follows the following intuitive cost vs. benefit decision rule: if the present value of the care (when it is eventually delivered) exceeds the cost of joining the wait list, they will schedule an appointment. The binary decision J for a person to join the wait list to see the higher-rated physician is:

$$J = \begin{cases} 1, & \text{if} \quad c < ve^{-dt} \\ 0, & \text{if} \quad c > ve^{-dt} \end{cases}$$

The present value of care is determined by the product of the current value of the care, v, which may include the value derived from a timely referral to a specialist, and an exponential function of the decay rate of demand, d, and wait time, t. The model parameters depend on the differential levels (of cost, value, and decay) between the low and high rated providers. The costs of joining the queue for care are denoted by c (e.g., calling to schedule the appointment).[2] For the ith individual, their value is

$$v_i(d,t) = v_i e^{-dt}$$

Appendix Figure 2 shows the cost-to-benefit tradeoff of a patient adding their name to a wait list for given values of c, v, and d as a function of the wait time t. If the value of joining the queue for care at the date of scheduling an appointment is v_1 and the decay rate is d_1 and costs to join the queue are c, then the critical length of time for joining the queue or not is $\hat{t}_1$. If the wait time t is greater than $\hat{t}_1$, then costs exceed benefits: $c > ve^{-dt}$. So the patient would not add their name to the queue.

As v, c, differ among demanders of care, the critical value $\hat{t}$ will vary. For queue joiners, $\hat{t}$ must be such that the net present value of the benefit exceeds the cost. I next focus on the *marginal joiner*, the individual whose $\hat{t} = t$. Accordingly, for the marginal joiner, expected benefits must equal expected costs: $ve^{-dt} = c$ and we can observe the following first order conditions which follow from differentiation and substitution:

$$\partial v/\partial d = vt > 0$$

$$\partial v/\partial t = vd > 0$$

An increase in the decay rate of the value of care will make someone previously on the margin of joining the queue not join. This is seen in Figure A2 holding v_1 fixed and moving from the

[2]Note that unlike earlier models of queuing, e.g., Barzel (1974), the costs of joining the wait list do not involve physically standing in a line, but merely placing your name on a list.

curve $v_1 e^{-d_1 t}$ to $v_1 e^{-d_2 t}$. Furthermore, holding the decay rate constant at d_2 while increasing the expected wait time from $\hat{t}_2$ to $\hat{t}_1$ increases the marginal queue joiner's value placed on the care from v_1 to v_2.

3.2 Rate of Joining the Queue

Next, given a fixed out-of-pocket price of the medical care (e.g., the patient pays only a pre-set copay for all family medicine), what is the rate of joining the queue? The rate of joining is determined by variation in $\hat{t}$ driven by decay rate d and fixed consumer attributes. As a first step, assume everyone in the population has the same rate d. Then, the only factor that gives rise to variation in $\hat{t}$ in the population is v, the valuation of care at the moment of illness onset. Assume v is distributed in the population according to $f(v)$, which is continuous and has finite range $0 \leq v \leq \bar{v}$. Someone at an expected wait of t_1 must then value the good at v_1 or more to join the queue. The number of people who join the queue per period, as a function of v and N, the population size, is given by

$$h(v) = N \int_v^{\bar{v}} f(v)dv = N[1 - F(v)]$$

and can be converted to t-space by substituting for $v = ce^{-d\hat{t}}$ to get

$$j(\hat{t}) = N[1 - F(ce^{-d\hat{t}})]$$

Which is the number of people for whom the critical delay time (i.e., to join/not join queue) is $\hat{t}$ or greater. Accordingly,

$$j(t) = N[1 - F(ce^{-dt})]$$

is the number of people who would queue at wait time t. Now, I point out the j-intercept:

$$j(0) = N[1 - F(c)]$$

which is the number of people who value the care more than the cost of simply joining the queue. This is also known as the "potential joiners".

The slope of the queue-joining function with respect to t is:

$$\frac{\partial j}{\partial t} = -Nf(v)\frac{\partial v}{\partial t} = -Nf(v)dv$$

This slope is negative which implies as t goes up, the number of queue joiners goes down. The slope of the joining function with respect to the decay rate, $\frac{\partial j}{\partial d}$, does not change at the *intercept* of the joining function because at $t = 0$, there is no change in $j(t)$. However, for a positive t queue time, as d goes up, the number of queue joiners goes down.

3.3 Supply of Family Medicine Rate Given Queue

Beyond whatever exogenous factors influence the quantity supplied (e.g., input cost shifters, regulation, etc.), queues may also influence the rate of supply. Supply at any given time h depends on those exogenous factors $\tilde{w}$ plus the wait time t and we assume that supply is positively affected by wait time:

$$s_h(\tilde{w}, t), \text{ such that } \partial s_h/\partial t > 0$$

The queue size at any given moment h is written as $Q_h = \sum_{k=0}^{\infty}(j_{n-k} - s_{n-k})$.[3] And the *rate of change* in the queue size at any point in time h is written as

$$\dot{Q}_h = j_h(t_h) - s_h(t_h)$$

The expected wait time in period h is t_h, the total number of people waiting in a given time divided by the supply service rate:

$$t_h = \frac{Q_h}{s_h}$$

3.4 Equilibrium and the Implications for the Empirical Setting

This system reaches an equilibrium at t^* when $t_h = t_{h+1}$. This occurs (by definition) when the rate of change in the queue length equals zero, $\dot{Q}_h = 0$.

The equilibrium of this supply and demand system is wait time t^* and queue size Q^* such that $j(t^*) = s(t^*)$; the number of people who would join the queue at wait time t^* equals the service rate (supply rate) at that t^*. And in this state, equilibrium queue size is $Q^* = j(t^*) \cdot t^*$.

This equilibrium is one in which wait times function very much like a price. In contrast to markets with prices, where clearing the market occurs via an increase in the price of the good and the demanders sort by willingness to pay, in this model, *wait times* clears the market by making the medical care less valuable as time in the queue increases. Since there is variation in the population according to initial value v of the care as well as d (the decay rate), the patients seeking care who have high values v and low decay factors d will crowd out those with lower v and higher d.

This model has testable implications. I expect to see longer wait times at higher rated physicians ($t^* > 0$). This also implies that at a given moment in time, the relative number or people in the queue is higher at higher-rated physicians. In my empirical setting, star ratings may causes an increase in demand at highly-rated physicians but at the same time, those physicians do not have an ability to modify their prices in the short run as a response to the disclosure. This model suggests market such as the one I study can be equilibrated by wait times instead of prices. There is an important implication that follows from this model: although an observer might at first believe that

[3] See Lindsay and Feibenbaum section I.B for exposition on normalizing the number of potential joiners in each queue.

an empirical finding of higher wait times for higher quality reflects an inefficient backlog of health care services, instead that same queue might actually be reflective of a market clearing process. In the short run, before high-quality providers can expand capacity or adjust prices, what does the disclosure do? It might lead to the creation of a brand-new "market for quality" that is cleared via a queuing mechanism rather than a price mechanism.

I would expect, as well, that as the short run bleeds into the long run and capacity of physician quality can adjust, the wait times may shrink back to zero. Accordingly, the pair of twin empirical findings that (a) quality rating disclosure reallocates consumers to high-quality sellers and (b) congestion increases at the highly-rated sellers in the absence of prices, might not reflect a market inefficiency but instead reflect a market process in which wait time takes the role of prices in rationing scarce demand.[4] In the following sections, I show that these two empirical predictions do in fact occur. The theoretical model relates these empirical findings to a single economic process.

4 Institutional Setting and Data

4.1 The Large Midwestern Health System

This paper uses data from a large Midwestern Health System ("the health system"), a non-profit integrated health system located in the upper United States. The health system has 46 hospitals (a mix of larger urban hospitals, such as in Fargo, Sioux Falls, Bismarck, and Bemidji, as well as smaller rural hospitals and an acute care children's hospital), more than 200 clinic locations, and nearly 1,500 providers. The health system is known for delivering high quality care in the region: In recent years, U.S. News and World Report has ranked the system's teaching hospital the top hospital in the state. The health system employs the majority of their physicians, and for all of the major insurance providers in the region, if the health system is in-network, patients would have equal access to all health system providers. Importantly, this uniform insurance coverage largely shuts down the role of out-of-pocket price in patient choices conditional on the insured choosing to receive care at the system. The majority of the health system's doctors are compensated on a work relative value unit (RVU) schedule.

4.2 Rating Data

As part of the health system's ongoing efforts to promote patient satisfaction, the system has collected surveys using external consultants ("survey providers"). These national survey providers, Press Ganey and NRC Health, administer post-visit questionnaires related to the patients' subjective experience with their health care provider. The questionnaires are sent out randomly and ask a series of standardized questions based on a survey developed by AHRQ called the Clinician

[4]This implies that policymakers ought not to worry about an increase in short-run congestion when quality ratings are disclosed because that could indicate an equilibrium sorting process.

and Groups Consumer Assessment of Healthcare Providers and Systems (CG-CAHPS). Based on dividing the total number of visits by the total number of submitted surveys, about 5% of total outpatient visits are followed up with a completed survey. Each provider is evaluated according to seven questions, including "Using any number from 0 to 10, where 0 is the worst provider possible and 10 is the best provider possible, what number would you use to rate this provider?"[5]

The answers to each of these questions are linearly transformed to a 5-point scale, and then the arithmetic mean across questions is taken to create a score for each provider for a survey visit. Details of this scaling transformation performed by the health system and their survey provider are available from me upon request.

Data from survey responses (and accompanying provider ratings) date back to 2016. However, until late 2018, rating data were never disclosed on the website, but instead held internally by the health system. On November 2, 2018, the health system launched online quality disclosure through a major overhaul of its website to include ratings and reviews for each doctor. Prior to this date, quality ratings were not available to patients and after that date, visitors to the health system's website see a prominently placed rating in large font (on a scale of 1 to 5 in one-tenth intervals) with corresponding gold star symbols next to a picture of each physician. The website also displays the number (raw count) of reviews. An artistic rendition of what the star ratings look like to consumers is found in Appendix Figure A1. According to the health system's disclosure policy, which is common across the health care industry, doctors with fewer than 30 ratings were not displayed until they reached the 30-rating minimum. For the November launch of rating, to "seed" the ratings with enough data, the health system used a 2-year look-back window to late 2016. The health system regularly updates the ratings for each provider as new survey data arrived, such that, through July 2020, the rating displayed for each doctor reflected the cumulative mean of all ratings to that date, starting from the beginning of the look-back window. In my data, I observe about 500,000 total surveys received by the health system between 2016 and 2020.

Although the values for each patient survey may range from 0 to 5, the vast majority of providers score highly on average and the overall distribution of average provider ratings is quite compressed near the top of the star range.[6] The provider-level ratings have a mean of 4.78, standard deviation of 0.13, and a slight negative skewness. A histogram of the distribution of average ratings is in Figure 1.

For each provider, I have information on listed specialty from the system website, their professional licensing credential (e.g., MD, registered nurse, physician assistant, etc.), provider gender, and a provider identifier (both the national provider identifier [NPI] as well as an internal health system provider identifier). These data come from hospital human resources data and the health system website. Using the entire history of individual patient surveys, I reconstruct the average (mean) raw

[5] The full list of survey questions is found in Appendix A1.

[6] A top competitor in the region also posts star ratings and has a similar distribution of average provider ratings. The competitor does not post star ratings for all providers (unlike the health system I study), perhaps because it is not an employer of most providers.

rating for each doctor at any given day; I then construct what the website displayed historically and verify using the Internet Archive Wayback Machine and internal communication with the health system. This results in a panel at the month level for each doctor containing the raw rating for each doctor on the 15th day of each month (the middle). From the raw, unrounded ratings, I also construct the rounded rating (to the nearest one-tenth), which is the score that is displayed on the health system website. To account for the fact that ratings drift slightly as more surveys are returned, I restrict the panel to include only providers whose rating is displayed at the same value for the duration of the month.[7]

4.3 Electronic Health Records Data

In addition to rating score data, I merge data that comes from a three-year extract of EHRs. The EHR contains de-identified visit data for all patient encounters across all locations in the health system during the three year period from 2017 to 2019. The EHR data contains International Classification of Diseases (ICD), doctor and patient identifiers, location and date of the service performed, and select health and demographic information, such as patient age, gender, zip code, body mass index (BMI), blood pressure, and smoking status at time of visit. Critically for this analysis, from the beginning of this window through August 2019, I have a variable that encodes whether the patient visit was a brand-new relationship between the patient and the provider or an existing relationship. The final months (quarter four of 2019) do not have this new patient visit variable because the EHR system takes some time to calculate and populate this field electronically. For my main analysis, I restrict providers to those practicing the specialty of family medicine according to the health system website; this is the most common specialty in the system (21% of providers) and is a specialty that I hypothesize would permit comparison shopping or consumer search online. The analytic data set comprises a panel of new patient visits at the doctor–month level and includes average rating (the running variable) and displayed ratings for each provider in the system.

4.4 Summary Statistics

Table 1 displays summary statistics for the data used in this paper. The upper panel describes the EHR data; there are more than 12 million total visits across 3 years and about 1 million unique patients. Demographic information available to me in the EHRs is limited. The average patient is 38 years old with a BMI of 27.5, indicating overweight but not obese. We expect patients who interact with the health system to be somewhat less healthy than the average person in the general population, and nothing about this health system suggests atypical patient composition.

[7] Dropping provider-months that display more than one rounded rating per month allows for a sharp regression discontinuity design but means that close to the discontinuity, there is a relatively smaller mass of data compared to further away. Empirical robustness checks in subsequent sections address this issue.

The lower panel of the summary statistics table contains provider-month level summary statistics for the family medicine providers, the baseline cohort for this analysis. These providers have (on average) 178 visits per month and see about 7 brand-new patients per month. These volume measures are skewed such that the mean is larger than the median, meaning there are some providers who have considerably larger visit volume and new patient volume. The average provider rating is a 4.78 and the standard deviation is 0.13. Half of providers have a rating that is rounded up, and the other half have a rating that is rounded down. At the instant quality disclosure was launched, the average count of reviews used to determine the average score of a provider was 228. As more ratings were added as more surveys came in, the average rating count increased to 298.[8] On the website, patients are shown the number of ratings a provider received, and a higher number of ratings could potentially send a stronger signal of quality to patients, all else equal. In total, 55% of family medicine providers are physicians (MDs and Doctors of Osteopathic Medicine [DOs], with the vast majority of these MDs), and the remainder are mid-level practitioners (such as advanced registered nurse practitioners, physician's assistants, etc.). There are 340 unique family medicine providers and the provider-month panel has 2,730 observations.

In Table 2, I assess the correlation between new patient visit volume at a given provider and that provider's online rating. I regress new patient visits per month on the provider's displayed rating score, and I estimate alternative specifications with month-year fixed effects, professional credential fixed effects (e.g., MD vs. PA vs. Nurse Practitioner) and both. In all specifications, I find an *inverse* relationship between rating score and new patient demand.[9] This negative relationship between rating score and new patient visit volume can also be seen in the slope of the points in the binned scatterplot in Appendix Figure A3.

I hypothesize that one driver of this inverse and unexpected relationship arises because high-quality doctors are also frequently near capacity (have full patient panels). If matching with a high-quality family medicine doctor is an absorbing state, then one would expect higher-rated doctors to also be willing to accept fewer new patients because they are already near capacity. Despite the negative correlation I find in Table 2, it is reasonable to believe that patients value quality and that there is not a structurally negative relationship between quality and volume. As a consequence, I approach the question with a causal design in the next section.

[8] The ratings were "seeded" with a 2-year lookback of historical ratings which explains an N larger than 1 on launch of ratings.

[9] I estimate the coefficient on score to be about -16, so a one-star increase is associated with 16 fewer new patients per month. Scaling this by a factor of 1/10, since ratings are displayed on the website in 0.1 intervals, a one-tenth rating increase, say from 4.7 to 4.8, is associated with 1.6 *fewer* new patients per month.

5 Empirical Strategy

5.1 Baseline Regression Discontinuity

I use regression discontinuity methods to compute the effect of an increased provider rating on demand for new patient visits (Thistlethwaite and Campbell 1960; Angrist and Lavy 1999; Hahn et al. 2001; Almond et al. 2010). In particular, the primary empirical strategy is to estimate regression discontinuity and difference-in-discontinuities models, which combines traditional regression-discontinuity estimation with difference-in-differences models (Lalive 2008; Grembi et al. 2016). This discontinuity approach to identification is pursued because although providers' actual ratings are continuous and smooth functions of the data, on the health system website, displayed ratings are rounded to the nearest tenth. For example, a doctor with a 4.749 will be rounded *down* to 4.7 stars, while a doctor with 4.750 will be rounded *up* to 4.8 stars, even though the underlying ratings are very close. Figure 2 outlines this identification strategy. I estimate the number of new patient visits per provider per month approaching the cutoff from the left side as well as the right side. In the figure, Doctor A and Doctor B have similar unrounded survey scores, but because of the rounding, their star rating is displayed differently on the website. The causal effect is the jump precisely at the cutoff; the assumption required for identification is that the other variables that affect new patient volume do not change discontinuously at the rounding cutoff. This is a sharp regression discontinuity design, since all providers with ratings above the rounding threshold are "treated" by being rounded up.

After constructing a panel at the level of provider-month, I estimate two series of regressions. The first series of regressions are based on the classical regression discontinuity estimator:

$$Y_{it} = \beta_0 + \beta_1 \mathbb{1}(\tilde{R}_{it} > 0) + \beta_2 \tilde{R}_{it} + \beta_3 \tilde{R}_{it} \mathbb{1}(\tilde{R}_{it} > 0) + \gamma_c + \varepsilon_{it} \tag{1}$$

where Y_{it} is the number of new patient visits per provider i in month t, $\tilde{R}_{it}$ is the running variable, the standardized raw rating, which runs from -.05 to +.05. I standardize each observation by the distance between the actual rating and the nearest one-tenth cutoff point because there are multiple different rounding cutoffs (e.g., 4.75, 4.85, etc.). This is common practice (Anderson and Magruder, 2012). Accordingly, β_1 is the coefficient on whether the provider's rating was rounded up (the coefficient of interest) and β_2 is the coefficient on the distance to the rounding threshold. Lastly, β_3 is the coefficient on the interaction between the running variable and being rounded up. This allows for alternative slopes to the regression line on both sides of the discontinuity. Also included are cutoff-specific fixed effects, γ_c. I estimate this as a global polynomial of orders 1, 2, and 3. In robustness checks, I estimate the regressions using alternative bandwidths (distances from the cutoff) both by varying the bandwidth size by .005 and use optimal bandwidth construction of Calonico et al. (2014). I weight these regressions based on review count, as higher number of reviews might have an outsized impact on behavior; this is consistent with more ratings leading

to a more precise signal (Magnusson, 2019). Robustness tests in a subsequent section address the economic importance of this weighting.

My preferred specification is a global linear (first-order) polynomial with alternative slopes on both sides of the discontinuity, with cutoff-specific fixed effects and weighting by review count.[10] The linear polynomial is preferred because a visual examination of the binned scatterplot of the running variable and the outcome of interest showed no obvious nonlinear trend, but I report variations by polynomial order and according to global and local linear regression. Standard errors are clustered at the provider level to account for potential error correlation within providers.

5.2 Difference-in-Discontinuities

The second series of estimators I construct are difference-in-discontinuities estimators (Grembi et al., 2016). In addition to the previously mentioned variables, I construct a new variable, $POST_{it}$ that evaluates to 1 if the provider-month observation occurs while the ratings were publicly disclosed online, and evaluates to 0 before they were disclosed.[11] I am able to implement the difference-in-discontinuities estimator because although the health system publicly disclosed provider rating scores only from November 2018 onward, they had been collecting ratings for many years beforehand. The difference-in-discontinuities regression takes the following form:

$$\begin{aligned} Y_{it} = \beta_0 + \beta_1 \mathbb{1}(\tilde{R}_{it} > 0) + \beta_2 \tilde{R}_{it} + \beta_3 \tilde{R}_{it} \mathbb{1}(\tilde{R}_{it} > 0) + \beta_4 POST_{it} \mathbb{1}(\tilde{R}_{it} > 0) + \\ \beta_5 POST_{it} + \beta_6 POST_{it} \tilde{R}_{it} + \beta_7 POST_{it} \tilde{R}_{it} \mathbb{1}(\tilde{R}_{it} > 0) + \gamma_c + \varepsilon_{it} \end{aligned} \tag{2}$$

where just like above, Y_{it} is the number of new patient visits per month. I recover separately the parameters β_1 and β_4; β_1 captures the causal effect of an increased rating on new patient visit volume when information *was not* disclosed, and β_4 captures the relative causal effect of an increased rating score on new patient visit volume when the information *was* disclosed. Again, I include cutoff-specific fixed effects, allow for alternative slopes on both sides of the discontinuity, and weight by count of reviews. As in the previous regressions, standard errors are clustered at the provider level.

[10] I also estimate a model that does not include cutoff fixed effects. Although the literature on rating response, e.g. Anderson and Magruder (2012) includes these cutoff specific fixed effects, I want to ensure that the estimation is robust to not including this fixed effect. According to Cattaneo et al. (2016), the pooled regression discontinuity estimator without cutoff fixed effects can be interpreted as a "double average"; the weighted average across cutoffs of the local average treatment effect for all units facing each particular cutoff value. The weighted average gives higher weights to the particular cutoffs that are most observed in the data in terms of observations.

[11] In these specifications, I drop November 2018, a partially treated month. The disclosure began on November 2, and results are robust to considering this to be a fully treated month.

6 Results

In this section, I show results on market responses to quality disclosure. I present two sets of results about quantity demanded, a baseline regression discontinuity analysis, which identifies a causal effect based on the rounded star rating, and a difference-in-discontinuities analysis that further leverages the time variation in patient exposure to online ratings. Next I discuss heterogeneity in the demand response to rating disclosure along a number of dimensions including provider specialization, patient age and health, and the density of providers in a given geographic area. I then show the effects of an increased star rating on wait times using a regression discontinuity identification strategy similar to what is used when analyzing the demand response but examining individual wait times. Finally, I test the robustness of my results by implementing a number of standard checks from the regression discontinuity literature.

6.1 Information Disclosure and Demand Response

6.1.1 Baseline Regression Discontinuity

In Figure 3, I begin by showing the relationship between the monthly new visits for a given family medicine provider and the distance that the provider's rating is from being rounded up (the running variable), with the distance normalized to zero. This binned scatterplot with 40 equally-sized bins provides a non-parametric way of visualizing the relationship between the running variable and the outcome of interest and assists with evaluating the presence of an effect at the discontinuity. Points to the left of the vertical dashed line represent the conditional mean within a bin for providers with ratings that are rounded down; points to the right of the vertical line correspond to the conditional mean of providers who have a rating which is rounded up. Overlaid on this plot are linear regression lines fit separately for data on each side of the rounding cutoff.

I observe a large and economically meaningful jump in the quantity demanded of new patient visits that takes place precisely at the discontinuity. In Figure 3, providers who have their ratings rounded down see approximately 5.5 new patients per month, whereas precisely at the cutoff, I observe a level increase in the number of additional new patients a doctor sees of approximately 3 new patients.

In Table 3, I provide a regression-based estimate of the causal impact of an increased provider rating on new patient visits. Columns 1-6 of Table 3 present various alternative specifications of Equation 5: linear, quadratic, and cubic in the running variable and allowing for vs. not allowing for alternative slopes on each side of the discontinuity. Based on the absence of a non-linear relationship between the running variable and the outcome variable in Figure 3, my preferred specification is a linear first order polynomial with an interaction between the running variable and the indicator for a provider's rating being rounded up; this is shown in Column 4 of Table 3. The estimated jump persists regardless of whether I assume the relationship between the running variable (distance to

rounding) and the outcome variable (new patient visits) is linear, quadratic, or cubic. I estimate that an increase in a provider's rating causes 2.96 additional patients per month to visit that provider (on a baseline of 5.475, this corresponds to a 54% increase). This causal estimate of the demand response is robust to alternative functional form specifications.

6.1.2 Leveraging Time Variation in Disclosure via Difference-in-Discontinuities

In Figure 4, I show the results of exploiting the unique institutional setting in which the health care system collected ratings for more than two years prior to ever disclosing provider quality scores to patients. I plot two separate series in a single graph: the blue dots represent the conditional mean of the outcome variable, breaking the data into 40 equally-sized bins, for the period of time when the ratings *were* disclosed online and when I have data on new patient visit volume (December 2018-August 2019). In contrast, the red triangles represent the conditional mean of the running variable, but for the "pre-disclosure" time period, from January 2017 to October 2018, when ratings *were not* observed by patients.

The results of Figure 4 are striking. Before online information disclosure of quality scores for providers, a provider whose score was rounded up was expected to see no additional patients per month. This zero-magnitude effect is seen when looking at the red regression line, which shows no meaningful jump in the outcome variable as the threshold is crossed for the pre-disclosure data. However, after disclosure, I observe a large and statistically significant increase in the number of new patients per month for providers with ratings rounded up. This can also be seen by noticing that to the left of the vertical dashed line in Figure 4, the blue dots and red triangles are commingled; in contrast, to the right of the rounding threshold, virtually all of the blue dots are above the red triangles.

I estimate the causal effects that are suggested by Figure 4 by using a difference-in-discontinuities regression and report the results in Table 4. This regression corresponds to Equation 6. The coefficient *Rounded Up* corresponds to the causal effect of an increased quality score in the pre-disclosure period, while the coefficient *Post X Rounded Up* corresponds to the causal effect of an increased quality score during the post-disclosure period. As expected, this effect is estimated as not significantly different than zero when ratings are not disclosed. However, when the ratings are disclosed online, I find an effect size of 4.496 new patients per month (an 88% increase off a baseline of 5.100 new patients per month). This difference-in-discontinuities model serves as a test to validate if other factors outside of online disclosure that also occur precisely as a provider's rating crosses the rounding threshold might causally affect new patient demand. For example, if the internally held but not released ratings were causing patients to see highly rated doctors more, this might be a threat to identification. Regression results from Table 4 serve to bolster and confirm the findings of a large demand response to the disclosure of quality ratings for providers.

6.2 Heterogeneity & Potential Mechanisms

I next explore the heterogeneity that underlies the large demand response to quality disclosure. These heterogeneity analyses will clarify which sub-populations benefit from and which are drivers of the demand response to quality. However, I caution the reader not to make causal conclusions based on these heterogeneity analyses, as unobserved differences across sub-populations inhibit one from making causal connections. Nonetheless, this series of heterogeneity analyses sheds light on some of the potential mechanisms behind the demand-side response to quality disclosure.

6.2.1 Provider Specialization & the Role of Choice versus Referrals

In Table 5, I consider the impact of quality disclosure differentially across provider specialties. The search process by which patients choose providers may differ considerably across the specialty of the physicians. Up to this point, my central focus was on family medicine because patients are frequently required to actively choose their primary care provider. In fact, HMO plans require the active choice of a primary care doctor. Family medicine is also the most common provider specialty in the data, comprising approximately 20% of all of the health system's providers. I now consider the effect of quality disclosure on the quantity of new patient visits at the top five specialties as listed for providers on the health system website (family medicine, pediatrics, internal medicine, cancer, and OB/GYN).

Column 1 of Table 5 shows a 54% increase in the number of new patient visits per month for family medicine doctors (also reported in Table 3). This effect is large and statistically significant. In contrast, however, in columns 2-5 of Table 5, I do not find statistically significant causal effects on the amount of new patient visits for providers with different specialties. None of the coefficients are statistically significantly different from zero at the 5% level, regardless of specialty (pediatrics, internal medicine, cancer, and OB/GYN). This confirms the prior hypothesis that family medicine providers may be those whose demand is most impacted by rating disclosure.

What might explain this heterogeneity across the specialties of providers? One possibility is that at the health system, family medicine providers serve as care coordinators who may create spillovers in terms of future health. If they can shape the trajectory of future patient health, then it might be reasonable for demand to be most sensitive to quality disclosure early on in the chain of care. Buttressing this theory is the fact that insurance design often forces active choices of primary care providers. In contrast, specialists are often found via a referral, in which the primary care doctor (rather than the patient) makes the decision about which doctor to see. This logic is consistent with large rating effects for family medicine but not for other specialties.

Another consideration that might drive the differences across specialties is the variation in the breadth of a patient's choice set. For example, within the specialty of family medicine, it is quite possible that all doctors listed within a geographic region could be chosen by a patient. However, in the case of specialty care for cancer, for example, if a patient needs care for a brain tumor, a

doctor specializing in hematology/blood cancers might not be a valid substitute. Thus, it does not surprise me that I recover a large effect for family medicine but not for other specialties, which are more differentiated within the broad specialty class.

Working against these interpretations is the possibility that there simply is not a large enough sample to identify a causal effect for the other specialties. The provider-month panel for family medicine, the most common specialty, has approximately three times as many observations as the next highest specialty, so the null effects might not be driven by the referral versus active choice hypothesis, but instead driven by sample size limitations.

6.2.2 Older or Younger Patients? Healthy Patients or Sick Patients?

In Table 6, I show estimates of the causal effect of a higher rating on new patient visits separately by the five age groups of adults used by the health system (ages 18-34, 35-49, 50-64, 65-79, and 80+). I find the largest response to quality disclosure is driven by the 18-34 age group (75% more new patients *in that age group* per month in response to an increase in provider rating). In older patients, the demand responsiveness to quality disclosure is lower (although even the 65 to 79-year-old subsample shows a statistically significant demand response to ratings). Note as well that the base rate for new patient visits at a given provider declines with patient age (older patients visit new family medicine doctors at a much lower rate than younger patients).

The overall pattern that the young adults are most sensitive to quality disclosure is consistent with primary care having characteristics of a credence good, where young individuals (with many years ahead of them) are sensitive to quality scores because there may face difficult-to-observe (in the short run) returns to provider quality. The result in Table 6 is evidence that younger patients are sensitive to quality disclosure for providers, potentially more than older patients. Chen (2018) studies the impact of physician Yelp ratings on revenues and patient volume using Medicare claims, but he finds considerably smaller effects than I do. My age heterogeneity analysis can partly explain that difference. Chen's paper uses data on Medicare patients (the preponderance of beneficiaries are age 65+) and combines that data with ratings from Yelp (a website which might be easier for younger rather than older individuals to navigate). One reason that the aggregate effect size I find (Table 3) is larger than what Chen finds in his paper is that I see evidence that a large portion of the effect of disclosure on quantity demanded is driven by the younger population, which he does not systematically study. Additionally, there are differences between the types of information about physicians found on Yelp and found on the health system website (based on AHRQ surveys). In prior studies of demand response to quality disclosure, the ratings are from surveys in which everyone is eligible to participate. In contrast, my setting relies on quality disclosure comprising of scores from a survey sent to a random subset of patients who received care. The differences between my larger results and the smaller magnitude results seen in Chen (2018), Brown, et al (working paper), and Luca and Vats (2013) might be due to the standardized and random nature of the surveys; if this is viewed by patients as more credible, it might induce a larger demand response.

This is consistent with a conversation I had with a health system CEO who said that he chose to publicly disclose quality scores based on AHRQ surveys (such as those studied in my paper) in order to control the information environment in direct comparison to what patients might find if they were to go to Yelp themselves.

In Table 7, I explore the relationship between patient health status and responsiveness to quality score disclosure. First, I separate patients into healthy and unhealthy patients. I do this three different ways: (A) if they ever have a comorbidity diagnosis code that would trigger a flag in a Charlson Comorbidity Index score, then they are categorized as unhealthy, e.g., a diagnosis of COPD, dementia, or cancer, for example, (B) I use obesity/BMI $\geq$30 to separate patients into healthy vs. sick, and (C) if the patient is ever recorded as a smoker.[12]

Columns 1-3 of Table 7 show the responsiveness to quality scores for the healthy patients. Providers whose ratings were rounded up saw 54%, 48%, and 55% more new *healthy* patients per month (where health is defined as no comorbidities, non-obese, and non-smoker, respectively). In contrast, columns 4-6 of Table 7 show the responsiveness to quality scores for the sicker patients. The sicker patients are more responsive to new patient ratings. Providers with ratings that are rounded up see 64%, 71%, and 54% (comorbidity, obese, and smoker, respectively) more *unhealthy* patients per month relative to providers with ratings that are rounded down.

The fact that sicker patients have a larger response to disclosed quality scores is consistent with the Grossman model of demand for health (Grossman, 1972). As an individual's health capital stock depreciates with illness, demand may be more sensitive to the quality of service provided. I note that the demand responsiveness for one category of health (smoking status) is not as stark as the other two (major comorbidities as well as obesity). Perhaps this is because there exists young and healthy smokers, and major comborbidities are often present later in an individual's life.

6.2.3 Do Provider Credentials Matter?

In the United States, family medicine is delivered by providers with numerous types of educational backgrounds and professional credentials. For example, a primary care provider might be an MD, DO, an advanced registered nurse practitioner, or a physician's assistant. Each type of provider credential requires different post-secondary education in order to practice, and consumers may view providers with different professional credentials in a different light.

In Table 8, I explore the impact of professional credentials on the response of patients to increased quality scores. Half of provider-months in the sample are MDs, and the other half are non-MDs. I find that the response to quality scores exclusively takes place among MDs. MDs see a 102% increase in the number of new patients per month that is causally attributed to an increase in a displayed provider score, whereas providers with other professional credentials see only a 6.5%

[12]Because my EHR data has only a primary diagnosis on a patient visit level (and not secondary diagnoses), I compute a Charlson score across all episodes for that patient in the EHR.

increase (not significantly different from zero). The mechanism behind this difference is unclear. Perhaps patients select MDs when they need a different type of care than when they select non-MDs. Given that the MD credential is typically the longest license to attain (in terms of years of formal schooling and residency), it is possible that consumer demand is sensitive to this aspect of provider training.

Another possibility that I suspect is that MDs specialize at more complicated care within family medicine whereas NPs might specialize in more routine care. If patients value high quality ratings more for more complicated care, that could generate the patterns observed in Table 8, with the majority of the causal effect driven by MDs.

6.2.4 Geographic Density of Physicians

I investigate the effect of provider density per capita on the demand responsiveness to ratings. In a model of search for physicians, more information may lower search costs, and provider density per capita may affect search costs, as well. I split the providers in the panel into groups which vary according to number of providers per capita in a given geographic area. Although the actual market for primary care is hard to calculate, I form geographic counts of providers at the county level. This does not, of course, proxy perfectly for actual physician geographic markets. However, I use counties because I can acquire the number of providers not just from the health system but from all physicians using the Area Health Resource File. Both per capita levels of all providers and per capita levels of the health system's providers are computed using 2017 county-level census data (from the Area Health Resource File [AHRF]). I assign a provider to a particular county by taking the modal county from which he or she draws patients, and then compute the number of primary care physicians per capita in each county (according to the AHRF as well as using the health system's physicians only). The distribution of primary care provider density is more or less split into two groups, which I call "low" and "high".

I find that providers working in above-median density counties see a much larger increase in number of new patients per month attributable to ratings (72%, 84%, for the all-physicians [AHRF] and the health system only cuts, respectively). See Table 9. In contrast to the large demand response for providers who draw patients from areas with a large number of family medicine doctors per capita, I do not find a statistically significant causal impact of ratings for providers in the below-median per capita density markets. An important factor to consider is that substitute information about provider quality is not randomly distributed across markets; for example, Yelp or HealthGrades may have substantial presence in large urban environments, but not in smaller rural settings. The presence of endogenous substitute information about quality is a difficult challenge to overcome. I am also hesitant to generalize the results from this heterogeneity analysis because within the health system's geographic area of operation, there may be insufficient variation in provider density across geography. Perhaps the results might differ if I included the nation's largest cities such as New York, Chicago, and Los Angeles. As such, I believe that more research on this question is warranted.

I also test the model of increasing monopoly (Satterthwaite, 1979), which hypothesizes that as physician supply in an area increases, the price of a reputation good may increase as the number of sellers in a market rises (in contrast to the canonical model where prices fall as number of sellers rise). The Satterthwaite increasing monopoly model hinges on the hypothesis that consumer search is less efficient in markets with many sellers. The conclusion of that model follows from two propositions. First, as the number of physicians in a market increases, the amount of consumer information about each physician decreases. For example, in a small town, it is easy to ask around for information about the town doctors, but in large cities, asking around about quality information for all doctors may be prohibitively costly. The second proposition of the increasing monopoly model is that as search becomes increasingly difficult, consumers become less price sensitive. It follows from these two propositions that as physician supply increases, fees for primary care rise.

The distribution of primary care providers in the area resource file four the counties served by the health system falls in three bins, which I call "low", "medium", and "high" density of primary care providers. The distribution of health system physicians (by county) is more or less split into two groups, which I call "low" and "high". I find that the physicians from the "high" number of physician counties do not have as large in magnitude an effect of quality disclosure on quantity demanded as the physicians from lower-count communities (Appendix Table A1). Although Pauly and Satterthwaite (1981) find evidence supporting Satterthwaite (1979), one possible reason that I find a larger response to disclosure in less physician-rich markets its because dense markets already have other unobserved (by the econometrician) sources of information about quality. For example, in larger cities, there may be better complements to the disclosed health system quality ratings (e.g., ratings from Yelp or HealthGrades) compared to smaller counties. The complementarities between the health system's quality disclosure and other sources of physician quality information make it more difficult to evaluate the effect of number of physicians within a geography on the effect of quality disclosure. Without exogenous variation to exploit on the number of physicians in an area, it is hard to tell the causal effects of the number of physicians on consumer search.

6.3 Sorting

In the previous sections, I showed that patient demand is responsive to quality score disclosure. In this section, I discuss the equilibrium consequences of this disclosure by studying the impact of provider rating disclosure on patient sorting. I study three dimensions of sorting: (1) Does the information disclosure shift patients to doctors who supply greater inputs to health? (2) Does the quality disclosure have an effect on brand new patients to the health system, on existing patients, or both? (3) Does the disclosure cause congestion at high-quality sellers? I use this analysis of the effect of ratings on wait times to understand who are the winners and losers of quality disclosure.

6.3.1 Inputs to Health

Many critics of disclosing doctor scores online claim that star ratings are uncorrelated with true provider quality, or worse, that ratings or report cards cause doctors to shift effort towards activity with low medical value but high rating value (such as putting fish tanks in a waiting room in order to receive favorable reviews). Doctors at the health system often complain to their administration about having scores posted online. (The most frequent critics are the low-rated providers.) The concern about providers reallocating effort towards tasks based on alternative performance measures is detailed extensively by Feng Lu (2012) in the framework of a multitasking agency problem. I assess whether this is occurring in my setting by measuring whether highly rated doctors supply greater levels of inputs to health.

The health system uses nine metrics to assess primary care quality; I study whether the highly scoring doctors in the online ratings also score highly on these nine internal quality metrics. The metrics are known as process measures, which is one of three types of performance metrics in the taxonomy created by Avedis Donabedian, the other types being outcome metrics and input metricss (Dranove, 2011). Outcome metrics (e.g., mortality) are challenging to use for evaluating primary care because the effects of primary care may be difficult to observe in the short run, and inputs (staffing ratios, hours of training) may be uncorrelated with actual desired results. Process measures, such as whether the providers use accepted practices and follow guidelines, are certainly not perfect measures of quality, but are nonetheless helpful tools to evaluate whether the providers are supplying commonly-accepted inputs to health. I rely on such process measures.

The nine metrics the health system evaluates are: frequency of BMI counseling, cervical cancer screenings, colorectal cancer screenings, diabetes management care, hypertension management care, mammography, pneumococcal vaccination, and 6- and 12-month depression followups. Doctor performance on these metrics is measured only for clinically eligible patients (e.g., the mammography denominator is based only on women withing the age range of government mammography guidelines). I compare the propensity of a doctor to undertake recommended medical care to their average star rating. The relationships are plotted in Figure 5; the best fit line is plotted over a binned scatterplot of the data.

For all nine of the process metrics, higher-rated providers are also supplying greater inputs to health. Note that the binned scatterplots are tighter and steeper for the cancer screenings and vaccination relative to the BMI, hypertension, and diabetes counseling scatterplots. This suggests a stronger relationship between process metrics and quality score in settings where doctors alone have greater control over inputs to health relative to settings that are more jointly determined by provider inputs as well as patient lifestyle and behavior such as weight and blood pressure. The overall slopes are consistent with Perez and Freedman (2018), who find that best-ranked hospitals had better clinical quality scores than worst ranked hospitals.

In sum, I conclude based on these relationships that in addition to disclosure shifting patients to

higher-rated providers, disclosure is shifting patients to providers who supply greater inputs to health, on average.

6.3.2 Is Disclosure Causing Market Expansion or Switching?

Is the demand response to quality disclosure primarily having an effect on patients who are brand-new to the health system, or is the effect concentrated among switchers, those who choose new doctors but have already sought care from other providers within the health system? I investigate this question to better understand whether quality disclosure primarily causes a market expansion or a reallocation of established patients. It is possible that both occur. To differentiate across this dimension, I use the EHR data to identify brand-new patients to the health system (which I label *de novo* patients) versus established patients (new patients to a particular doctor, but not to the health system). The EHR data extract that I have does not have an indicator for *de novo* patients, but does have an indicator for patients who are new to a particular provider. I use a three-pronged data-driven method to identify *de novo* visits. The visits must be (1) the patients' first recorded visit in the entire extract of the EHR I have access to (2017-2019); (2) flagged as a "new visit" for the particular doctor, meaning even if it is the patient's earliest occurrence in the EHR file, but it is not a "new visit" with that particular provider, it does not count as *de novo*; and (3) after November 2018, which creates a nearly 2-year window in which the patient did not appear in the EHR at all before their first appearance. These rules are meant to prevent as many patients who had already visited other health system doctors from inadvertently getting classified as *de novo*. A patient could have seen a health system doctor in 2015 (before my data window) and had a subsequent first visit with any provider after November 2018, but I think this gap would be unlikely.

The results of this market expansion versus market stealing breakdown are displayed in Figure 6. This figure shows that patients who already had previous contact with the health system, but with different providers, are driving the response to quality disclosure rather than *de novo* patients. In Appendix Table A2, I estimate that the additional new patients a provider sees per month who are switching from other health system providers increases by 2.059 new patients per month (e.g., 60% increase on a baseline of 3.454 found in column 4). However, for *de novo* new patients (those who have never been to any doctor at the health system, I do not observe a statistically significant increase in the number of new patients a provider sees if they have a higher rating due to rounding (Appendix Table A3).

I view Figure 6 and Tables A2 and A3 as suggestive evidence that the response to demand occurs mainly along the margin of switching, causing a reallocation of previously existing patients towards physicians and other providers who are rated more highly in terms of quality scores.

6.3.3 Wait Times and the "Price of a Star"

In this section, I explore the causal effects of quality disclosure on congestion. In doing so, I link my empirical results to the theoretical model by examining wait times. Wait times may play a role in rationing scarce quality because health care is different from conventional product markets in part due to the presence of third-party payors (insurers). Because patients can often face the same price for care from any provider in their insurance network, there is no direct out-of-pocket price that can easily vary in physician quality. This directly contrasts with conventional products, where sellers can immediately raise (or lower) prices in response to a high (or low) quality score when scores are disclosed.[13]

To motivate the possible role that wait times have in equilibrating supply and demand after ratings disclosure, I first focus on conventional product markets as a benchmark. In the case of conventional products, Wolinsky (1983) models an equilibrium where individual sellers set prices in response to buyers' expectations of quality. In that model, Wolinsky establishes a separating equilibrium where each price signals a unique level of quality. In contrast, health care providers do not have any way to adjust prices paid by consumers in the short run after disclosure. Conditional on service line (e.g., family medicine) and insurance plan membership, patients at the integrated health system pay the same amount out-of-pocket and have the same access to the same set of doctors. In sum, at the point-of-sale to a patient, the patient effectively pays the same out-of-pocket price for any primary care provider they see, regardless of the quality rating of a provider. High-quality providers cannot charge patients more based on their high rating (or any other factor). Of course, physicians could always leave the system, but in the short run, the patient does not face a higher price for quality and capacity and entry are fixed.

Does the market have any way to find equilibrium in the absence of a monetary price for differential quality? Richards-Shubik et al. (2021) suggests that congestion (or wait times) play a similar role to prices in such markets. I evaluate this hypothesis by studying wait times, measured in the number of days between when an appointment is booked and when that appointment takes place.

For each outpatient visit with family medicine providers, I compute the total number of days that the patient waited for care (using the EHR data to gather the number of days between when an appointment is entered into the system and when it occurs). I make a few sample restrictions. First, I exclude from the data all visits that occur more than 180 days after they are scheduled, as these represent visits for which patients do not likely care about wait time to see a doctor (there is a small mass of visits that are scheduled exactly one year out). Second, I drop visits that occurred at a walk-in clinic (as the patient might not have a choice of a particular provider); individuals less than 18 years old; visits where the flag for the visit being new to a provider was not present (primarily post August 2019); and visits when the wait time was coded in error as being less than 0 days.

[13] For example, sellers can immediately raise or lower prices in response to changes in ratings on the online tutoring platform www.wyzant.com, where sellers name their own prices and star ratings are a salient part of product search.

To identify the causal effect of ratings on wait time, I exploit both the variation induced by rounding ratings to the nearest tenth as well as the variation in timing of pre- vs. post-disclosure of quality scores to estimate both a regression discontinuity model as well as a difference-in-discontinuity model in the spirit of the identification strategy laid out in Section 5. These models assess whether patients wait longer to see a provider with a higher rating. The regression is similar to the model estimated in Tables 3 and 4, but run at the individual level rather than provider-month level, and I also include a diagnosis code fixed effect (using the primary ICD9 code for the visit) because the patient's type of medical condition when arriving at the doctor might dictate how quickly the provider moves them to the front of the line. For the specifications presented in Table 10, I restrict the bandwidth to 0.025 on both sides of the cutoff of the normalized running variable, and report robust standard errors.

The results in Table 10 show that a higher star rating causes new patients to wait longer to receive care. Column 1 of Table 10 presents the pre-disclosure (placebo) regression discontinuity specification which finds no increase in wait times (statistically indistinguishable from zero). Column 2, the regression discontinuity specification that relies only on post-disclosure data, shows an effect of 2.105 additional days on a baseline of 8.765 (24.0%). Finally, the difference-in-discontinuities (Column 3) shows that a higher star rating causes new patients to wait 2.695 days longer to receive care relative to a baseline of 8.848 days (a 30.5% increase). In Appendix Table A4, I perform a barrage of robustness tests regarding these specifications. First, in the spirit of Imbens and Lemieux (2008) and Eggers and Hainmueller (2009), I test for jumps at non-discontinuity points. I construct two "false placebo" thresholds in the running variable, at -0.025 and 0.025 instead of 0, and find no statistically significant increase in wait time at these placebo points (this is true both during the disclosure period as well as prior to the disclosure period). For the "true" discontinuity (0.00 from the rounding threshold) during the period that the ratings were public online, I find a statistically significant increase in the wait time for new adult patients of 2.450 days; however, there is no statistically significant difference at the true discontinuity during the pre-disclosure period (as expected). To further ensure that I am picking up a causal effect, these robustness test regressions are estimated by first residualizing wait time on cutoff fixed effects and ICD-9 fixed effects and then estimating optimal bandwidth local linear regression (Calonico et al., 2017).[14] The optimal-bandwidth residualized binned scatterplots with local linear regression best fit lines for the pre-disclosure period and post-disclosure period are found in Appendix Figures A4 and A5, respectively, and illustrate an increase in wait times at higher rated physicians occurring when quality ratings are disclosed but not before.

I interpret this finding to represent a "shadow price of a star." That is, new patients are willing to wait 30.5% longer to get care from a physician who has a one-increment increase in their quality score (e.g., the effect of moving from a 4.7 to a 4.8). Furthermore, I can extrapolate this estimate to calculate how much patients are willing to wait for a one standard deviation increase in quality.

[14] I use the *rdrobust* and *rdplot* packages in Stata.

If I make the assumption that the effect size scales linearly as ratings increase, my estimate of a willingness-to-wait of 2.695 wait days for a 0.1 star increase represents a 3.05-day willingness-to-wait for a standard deviation increase in star rating (st. dev = 0.13). I argue that the wait time "shadow price of a star" operates similarly to a traditional price by helping supply and demand clear in this market. This market-clearing role of wait times, in which a higher-rated physician "costs more" in terms of number of days a patient must wait, helps facilitate equilibrium because if patients are heterogenous in their willingness-to-wait (just like patients may be heterogeneous in willingness-to-pay for conventional products), an equilibrium queue may emerge in the spirit of Lindsay and Feigenbaum (1984). Here, sorting occurs on the basis of underlying valuation of quality, and disclosure creates a market for physician quality which did not exist in the absence of ratings.[15]

In addition to examining the effect of a quality score on new patients' willingness to wait for care, I also investigate what happens to wait times for established patients when a provider's quality score increases. I previously showed that an increase in a providers' rating causes more new patients to see that provider. This creates congestion for established patients. In Table 11, I show that established patients wait longer to receive care from a doctor with a higher quality score. Columns 1 and 2 show pre- and post-disclosure regression discontinuity estimates, and column 3 shows an effect of a 1.736 day increase in wait times on a baseline of 12.8 days (12% increase). Because these patients are not shopping for a new provider, I interpret this to be evidence of congestion spillovers: If capacity of family medicine providers is restricted in the short run, since additional patients visit providers due to higher ratings, the established patients face congestion. Clearly, these patients suffer as a result of the quality disclosure. They were already seeing the higher-rated doctor, but disclosure causes them to wait longer for care because newer patients are now sorting to that doctor, as well.

I also explore whether the congestion effects of star ratings differs by the urgency of the patient's medical condition. From a high-level perspective, if patients wait longer to see family doctors with higher star ratings, all else equal, and lower star rating doctors have excess capacity (or "slack") because of this additional volume at higher-rated doctors, it could be inefficient for patients to wait longer for conditions that might end them up in the emergency department. Using the decomposition between productive and allocative efficiency (for example, see Baicker et al. (2011)), I note that it may be efficient from the perspective of the health system for patients to wait longer for a physician with a higher star rating for non-urgent conditions like a checkup but not for urgent conditions. Perhaps the preferences of the patient for a checkup from a higher-rated physician are such that they are willing to wait longer, and there is little efficiency cost to this additional waiting which reflects a revealed preference argument about patient choices. However, if patients

[15] People willing to wait longer may have less acute needs and one possible implication is that quality disclosure with wait times as shadow prices could lead to suboptimal allocation of resources. However, disclosure without price adjustment may lead to more equitable allocation compared to disclosure where prices can respond depending on the relative distributions of willingness-to-pay versus willingness-to-wait in the population.

are waiting longer for care that is urgent in search of a higher star (such as care that would wind them up in the emergency room if not treated quickly), then it might be productively inefficient for these patients to be reallocated or sorted to doctors with excess availability.

I test this by restricting to a subset of cases where patients are seeking care from family medicine doctors where ED care might be needed but is preventable or avoidable. I use a taxonomy of diagnosis codes developed from an algorithm developed by John Billings at NYU Wagner.[16] In Table 12, I show that patients are willing to wait longer for avoidable ED care when star ratings are disclosed (but not before) using the same regression-discontinuity design as before. When stars are disclosed (column 2), patients are willing to wait 2.37 additional days for a higher-rated physician when they are seeking care that the Billings, et. al. algorithm would consider to be urgent where ED care may be needed but is preventable or avoidable. If these patients were simply reallocated to doctors with lower stars who had excess capacity, it may lead to an efficiency improvement from the perspective of the health system.

In conclusion, this congestion effect (and willingness-to-wait for quality) is informative in explaining how quality disclosure operates in markets with limited ability to adjust prices. How is equilibrium reached? Sorting patients based on willingness-to-wait for quality is one way in which this market can reach equilibrium in the absence of a price. The ability of this market to reach equilibrium may be dependent on sorting based on willingness to wait for quality.

6.4 Robustness

In this section, I present a number of robustness checks. I address potential pitfalls relating to the bandwidth used for the regression discontinuity estimates, to the functional form of the running variable, and to the use of local linear regression. I also test for covariate balance. I find that the results are robust to these tests; although my point estimates very minimally across some specifications, the direction and magnitude of my estimates holds up under the barrage of traditional regression discontinuity robustness tests.[17]

6.4.1 Bandwidth

To check that the regression results above are not sensitive to proximity to the cutoff and choices of the econometrician, I vary the bandwidth under which data is included in the regression discontinuity. Because regression discontinuity models are identified locally at the jump in the conditional function of the running variable, data far from the discontinuity can lead to biased estimates (Lee and Lemieux, 2010). However, the more I restrict to a very narrow bandwidth around the discon-

[16] Available here: https://web.archive.org/web/20160313195339/https://wagner.nyu.edu/faculty/billings/nyued-background

[17] In fact, the first robustness check is seen in the presentation of Table 3, where I show that the results of the baseline regression discontinuity model are invariant to linear, quadratic, or cubic polynomial functional form.

tinuity, the less data is available for estimation. Accordingly, adjusting the bandwidth induces a bias-variance tradeoff.

The results hold as I increasingly restrict the bandwidth (see Table 13). I plot the coefficients and standard errors for the baseline specification causal effect as I vary the bandwidth used in estimation from (-.05,.05) to (-.01,.01) in Figure 7. I find that the results are insensitive to adjustment in bandwidth size. (As bandwidths decrease, there is less data on which to estimate, so confidence intervals widen slightly.) However, the overall results are invariant to bandwidth variation. I also plot the optimal bandwidth selected by the routine of Calonico et al. (2014), denoted by the dashed line labeled "CCT."

6.4.2 Manipulation, Density Tests and Alternative Sample Definitions

A concern in regression discontinuity design studies is that there is precise manipulation of the running variable by agents who want to be on a certain side of a cutoff. From a high-level perspective, I do not think this is likely a problem in this setting, since a provider would have considerable difficulty in manipulating their rating to be rounded up or down. Why? Because provider surveys are sent randomly and submitted by only a small number of patients, and a provider would have no way of knowing *ex ante* which patient would receive and ultimately submit a survey. Accordingly, they would have to exert effort on every single patient in order to be on a given side of the threshold (rounded up). Also, providers do not know their own distance from the threshold during the time period I study. (After my study window ended, providers were made known about their current raw underlying rating, but during my data availability, providers had no way of knowing if they were close to being rounded up or far from the threshold.) Nonetheless, to test for manipulation of the running variable, I plot the density of the running variable in discrete bins on both sides of the threshold in the spirit of McCrary (2008).

Appendix Figures A6 and A7 show that there is no discontinuity in the density of the running variable (quality rating on the 15th day of the month) that would suggest bunching on one distinct side of the threshold. Figure A6 plots this histogram for *all* the providers in the data, where Figure A7 plots the density for the subsample of providers who have only a single disclosed score in a given month and do not have multiple scores in a given month. Although the density is symmetric around the threshold in both settings, there is a symmetric dip in the number of providers very close to the threshold in Figure A7. This dip is explained by fact that providers with more than one rating a month (say, who show both a 4.7 and 4.8) are likely to have a closer score to the rounding threshold given that they crossed it.

As an additional robustness check to make sure that the baseline regression results are robust to not dropping the provider-months which cross the rounding threshold in a given month, I plot the regression discontinuity results and report regression tables for the sample where I do not drop these observations (Appendix Figure A8 and Appendix Table A5). The results are quantitatively

and qualitatively similar to the baseline specification.

As mentioned in footnote 10, as an additional robustness check, I estimate the main baseline regression discontinuity model (number of new visits per month) *without* including cutoff specific fixed effects, which results in a coefficient which can be interpreted as a "double average", the weighted average across cutoffs of the local average treatment effect for all units facing each particular cutoff value, giving higher weights to the particular cutoffs that are most observed in the data set. Table A7 shows the estimates from the Rounded Up coefficient of interest for the same six baseline specifications as the cutoff-specific fixed effects model found in Table 3. The estimates are comparable in both magnitude and direction across all specifications.

6.4.3 Covariate Balance

In Appendix Figure A9, I show that based on observable predetermined characteristics, physicians with ratings that are rounded up display no different qualities than those just rounded down. I include these covariate balance tests for four predetermined attributes in the provider–month panel (the probability a physician is male, the probability the provider is an MD, the probability they are employed in a high density of provider market [using the definitions from section 6.2.4] and the elapsed years since that provider started working at the health system). Figure A9 shows covariate balance across each of these available predetermined attributes. Appendix Table A6 shows the regression estimates from these covariate balance tests. Physicians with ratings rounded up seem to be no different than physicians with ratings rounded down based on available predetermined observables.

6.4.4 Weighting & the Significance of Number of Reviews

I also show my results are robust to whether or not I weight the observations by rating count in addition to varying the bandwidths and global polynomials in Table 14. Following the practice of Magnusson (2019), I estimate the baseline specification unweighted, weighted by count of ratings, and weighted by inverse rating count. Weighting by count allows the providers with more precise information signals due to more scores reported on the website to reflect that precision, whereas weighting by inverse count allows providers with fewer ratings (and less precision of signal) to count for more. I find that the results are as expected: count ratings show a stronger causal effect, and inverse count ratings shrink the effect towards the null. Unless otherwise indicated, throughout this paper, weighted estimates are shown, as a higher count of reviews may reflect a higher level of information available to consumers (in the spirit of Bayesian updating).

7 Discussion

7.1 Limitations

In this paper, I use a physician-level star rating disclosure policy at a large midwestern health care system to study the effects of quality disclosure on economically meaningful outcomes such as demand, sorting, and congestion. Using a regression discontinuity design, I find that quality disclosure caused a response in the quantity demanded of highly rated physicians, leading to a 2.96 new patient per month increase caused by an additional tenth of a star. I also find that the demand response was heterogenous across provider specialty and age, among other dimensions, as well as finding that disclosure caused longer wait times at higher rated physicians.

This study is not without limitations, however. First and foremost, I do not have data on many dimensions of physician behavioral response to ratings disclosure that would allow me to identify a supply response on the part of physicians. For example, I am not able to ascertain if physicians substituted to providing different services that patients might demand. A common concern is that patients could reward physicians by leaving high ratings for providing medically unnecessary services, such as prescribing antibiotics for ear infections when antibiotics are not helpful or even harmful (Martinez et al., 2018). Because my data set does not have granular procedure code data about what treatments physicians performed, I am not able to test whether physicians responded to quality disclosure by altering the type or quality of care they provide or by adjusting across different dimensions of quality.

Another limitation to this paper is that I do not have longitudinal data on physician rates of screenings, vaccinations, and counseling services. The analysis displayed in Figure 5 could be more informative about the causal effect of rating disclosure on these services had I been able to construct a panel over time of physician propensity to supply inputs to health. Because I only have a single snapshot of physician screening and vaccination rates to provide these services but ratings fluctuate over time, I cannot estimate regression discontinuity models using these outcomes in the same sense as in other sections of the paper. Furthermore, as is common in papers studying the impacts of family medicine, it is difficult to observe direct health outcomes as compared to specialties such as cardiac surgery, where mortality and adverse events are far more common. Nonetheless, despite these limitations, I show that ratings, which cause changes in demand, also shift patients to doctors who, on average, perform more of these medically recommended services.

Lastly, these results may not generalize to other populations that may differ demographically or in their propensity to use quality information to search for physicians. Although generalizability is a possible concern (the large Midwestern health system cares for a population that is more White and more rural than the United States as a whole), I nevertheless note that this is an ideal population to study the questions posed in this paper. First, the system covers a broad geographic and demographic area (four states with both rural and urban areas). Second, the advantages to

studying the impacts of quality disclosure in my setting, where quality disclosure is mandatory, where patients face the same price for any provider, and where there is unique pre- and post-disclosure data, suggests that my setting is an ideal laboratory for this study.

7.2 Conclusion

In this paper, I provide new evidence on the causal effects of star rating disclosure on demand, sorting, and congestion in markets where prices cannot readily adjust to new information about quality. I leverage a unique institutional environment and a causal framework to show that demand is responsive to medical provider star ratings and that ratings sort patients to higher-quality providers.

I find a 54% increase in new patient visits caused by a provider having their rating rounded up relative to rounded down. I explore the drivers of this demand response by addressing heterogeneity, such as age, health status, and provider type. Younger patients are more responsive than older patients (75% increase in new visit volume by 18- to 34-year-olds relative to 58% by 60- to 64-year-olds), perhaps because the younger patients are more accustomed to seeking quality information on the internet, and sicker patients are more responsive than healthy patients, perhaps due to sicker patients placing a greater value on physician quality. I show that disclosure shifted volume to providers who on average produce greater levels of medically recommended inputs to health (screenings, counseling, and vaccinations), and I show that a higher online rating also causes increased wait times at a provider. New patients wait 30.5% longer for a doctor with a higher rating and established patients wait longer, too (12.6% longer). These results are consistent with my model of congestion effects in which wait times serve as a shadow price for quality and equilibrate the market.

Taking all the evidence together, quality disclosure appears to facilitate an equilibrium outcome in which patients actively look for information about product quality, in which they act on that information by substituting to higher-rated and higher-quality sellers, and select an experience good based on their willingness to pay (wait) for quality. Using the reduced form estimates and extrapolating to a one-standard deviation increase in quality, I estimate the shadow price of a star is that consumers are willing to wait 3 additional days for a one standard deviation increase in quality. I argue that this shadow price facilitates equilibrium market clearing in a setting where price differences are unable to do so.

My results shed light on the complex role that quality disclosure plays in market outcomes, particularly in the market for health care and other insured products where prices cannot immediately vary after disclosure. Many health systems have adopted quality ratings in the past decade, and business leaders (e.g., hospital management) along with policymakers continue to focus on expanding the scope of physician ratings. Understanding the effects of star rating disclosure on such markets is key to designing, implementing, and evaluating policies meant to fix market imperfections by

improving patient access to information about quality. This paper contributes to the growing body of empirical literature on information disclosure by providing novel evidence about information's effect on non-price markets and these results inform scholars as well as policymakers about the equilibrium effects of quality disclosure.

References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Almond, D., Doyle Jr, J. J., Kowalski, A. E., and Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The quarterly journal of economics*, 125(2):591–634.

Anderson, M. and Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563):957–989.

Angrist, J. D. and Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly journal of economics*, 114(2):533–575.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973.

Baicker, K., Chandra, A., et al. (2011). Aspirin, angioplasty, and proton beam therapy: the economics of smarter health care spending. In *Jackson Hole Economic Policy Symposium*, volume 41. Citeseer.

Baker, G. (2002). Distortion and risk in optimal incentive contracts. *Journal of human resources*, pages 728–751.

Bederson, B. B., Jin, G. Z., Leslie, P., Quinn, A. J., and Zou, B. (2018). Incomplete disclosure: Evidence of signaling and countersignaling. *American Economic Journal: Microeconomics*, 10(1):41–66.

Bundorf, M. K., Chun, N., Goda, G. S., and Kessler, D. P. (2009). Do markets respond to quality information? the case of fertility clinics. *Journal of health economics*, 28(3):718–727.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, 17(2):372–404.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G., and Keele, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4):1229–1248.

Chen, Y. (2018). User-generated physician ratings: Evidence from yelp.

Chernew, M., Gowrisankaran, G., and Scanlon, D. P. (2008). Learning and the value of information: Evidence from health plan report cards. *Journal of Econometrics*, 144(1):156–174.

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Cullis, J. G. and Jones, P. R. (1985). National health service waiting lists: A discussion of competing explanations and a policy proposal. *Journal of Health Economics*, 4(2):119–135.

Cullis, J. G., Jones, P. R., and Propper, C. (2000). Waiting lists and medical treatment: analysis and policies. *Handbook of health economics*, 1:1201–1249.

Dafny, L. and Dranove, D. (2008). Do report cards tell consumers anything they don't already know? the case of medicare hmos. *The Rand journal of economics*, 39(3):790–821.

Dranove, D. (2011). Health care markets, regulators, and certifiers. In *Handbook of health economics*, volume 2, pages 639–690. Elsevier.

Dranove, D. and Jin, G. Z. (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature*, 48(4):935–63.

Dranove, D., Kessler, D., McClellan, M., and Satterthwaite, M. (2003). Is more information better? the effects of "report cards" on health care providers. *Journal of political Economy*, 111(3):555–588.

Dranove, D. and Sfekas, A. (2008). Start spreading the news: a structural estimate of the effects of new york hospital report cards. *Journal of health economics*, 27(5):1201–1207.

Eggers, A. C. and Hainmueller, J. (2009). Mps for sale? returns to office in postwar british politics. *American Political Science Review*, 103(4):513–533.

Feldstein, P. J. (2007). Health policy issues. an economic perspective.

Feng Lu, S. (2012). Multitasking, information disclosure, and product quality: Evidence from nursing homes. *Journal of Economics & Management Strategy*, 21(3):673–705.

Gibbons, R. (2010). Inside organizations: Pricing, politics, and path dependence.

Grembi, V., Nannicini, T., and Troiano, U. (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics*, pages 1–30.

Grossman, M. (1972). On the concept of health capital and the demand for health. *Journal of Political Economy*, 80(2):223–55.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461–483.

Grossman, S. J. and Hart, O. D. (1980). Disclosure laws and takeover bids. *The Journal of Finance*, 35(2):323–334.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Hanauer, D., Kullgren, J., Singer, D., Solway, E., Kirch, M., and Malani, P. (2020). National poll on healthy aging: Searching for a good doctor, online.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*, 7:24.

Hunter, M. (2020). Chasing stars: Firms' strategic responses to online consumer ratings. *Available at SSRN 3554390*.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635.

Jin, G. Z. and Leslie, P. (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. *The Quarterly Journal of Economics*, 118(2):409–451.

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

Jovanovic, B. (1982). Truthful disclosure of information. *The Bell Journal of Economics*, pages 36–44.

Kolstad, J. T. (2013). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review*, 103(7):2875–2910.

Lalive, R. (2008). How do extended benefits affect unemployment duration? a regression discontinuity approach. *Journal of econometrics*, 142(2):785–806.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355.

Lindsay, C. M. and Feigenbaum, B. (1984). Rationing by waiting lists. *The American economic review*, 74(3):404–417.

Luca, M. and Vats, S. (2013). Digitizing doctor demand: The impact of online reviews on doctor choice. *Cambridge, MA: Harvard Business School*.

Magnusson, E. (2019). Unboxing the causal effect of ratings on product demand: Evidence from wayfair. com. *Com (October 2, 2019)*.

Martinez, K. A., Rood, M., Jhangiani, N., Kou, L., Boissy, A., and Rothberg, M. B. (2018). Association between antibiotic prescribing for respiratory tract infections and patient satisfaction in direct-to-consumer telemedicine. *JAMA internal medicine*, 178(11):1558–1560.

Mathios, A. D. (2000). The impact of mandatory disclosure laws on product choices: An analysis of the salad dressing market. *The Journal of Law and Economics*, 43(2):651–678.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714.

Nelson, P. (1970). Information and consumer behavior. *Journal of political economy*, 78(2):311–329.

Pauly, M. V. and Satterthwaite, M. A. (1981). The pricing of primary care physicians services: a test of the role of consumer information. *The Bell Journal of Economics*, pages 488–506.

Perez, V. and Freedman, S. (2018). Do crowdsourced hospital ratings coincide with hospital compare measures of clinical and nonclinical quality? *Health services research*, 53(6):4491.

Pope, D. G. (2009). Reacting to rankings: evidence from "america's best hospitals". *Journal of health economics*, 28(6):1154–1165.

Propper, C. (2000). The demand for private health care in the uk. *Journal of health economics*, 19(6):855–876.

Richards-Shubik, S., Roberts, M. S., and Donohue, J. M. (2021). Measuring quality effects in equilibrium. Technical report, National Bureau of Economic Research.

Satterthwaite, M. A. (1979). Consumer information, equilibrium industry price, and the number of sellers. *The Bell Journal of Economics*, pages 483–502.

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

Schwartz, A. J., Yost, K. J., Bozic, K. J., Etzioni, D. A., Raghu, T., and Kanat, I. E. (2021). What is the value of a star when choosing a provider for total joint replacement? a discrete choice experiment. *Health Affairs*, 40(1):138–145.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309.

Werner, R. M. and Asch, D. A. (2005). The unintended consequences of publicly reporting quality information. *Jama*, 293(10):1239–1244.

Wolinsky, A. (1983). Prices as signals of product quality. *The review of economic studies*, 50(4):647–658.

8 Figures and Tables

Figure 1: Distribution of Provider Average Ratings

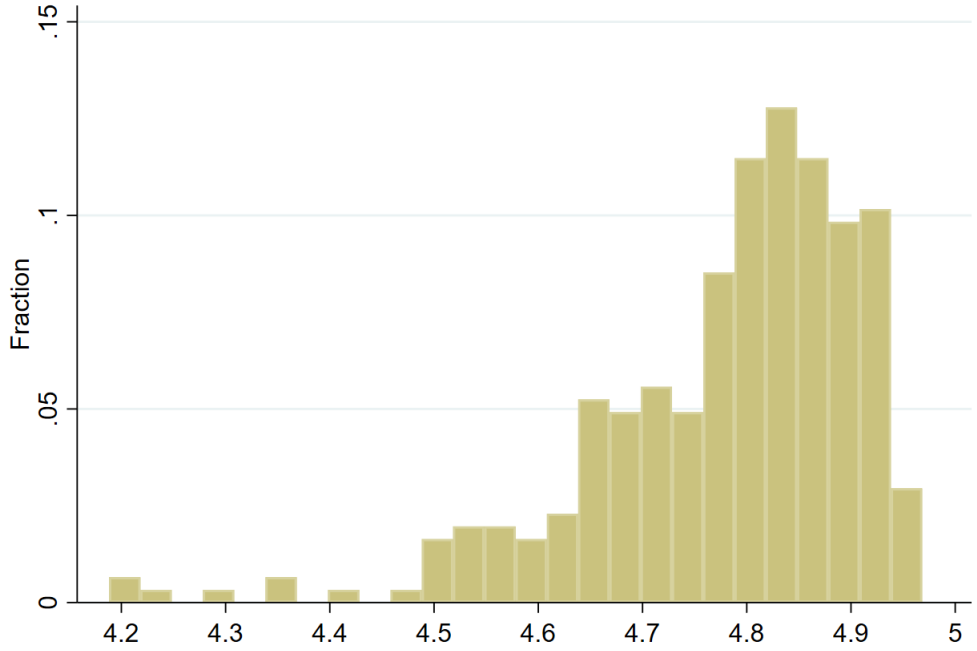
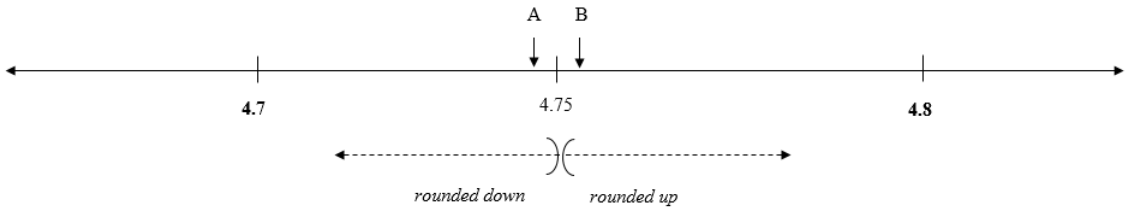
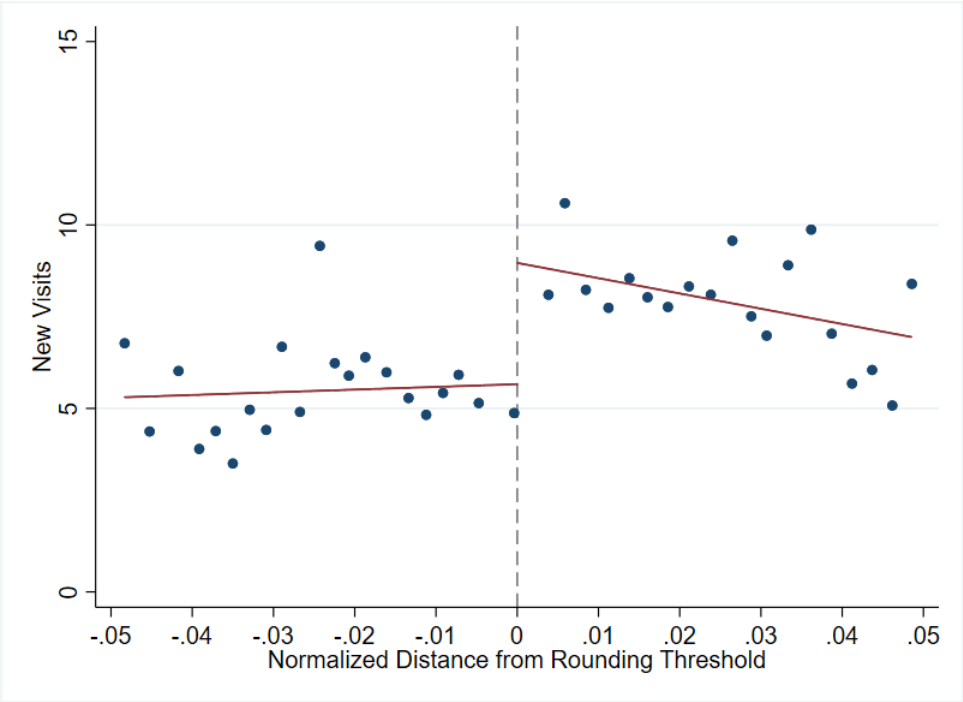


Figure 2: Intuition of Identification Strategy



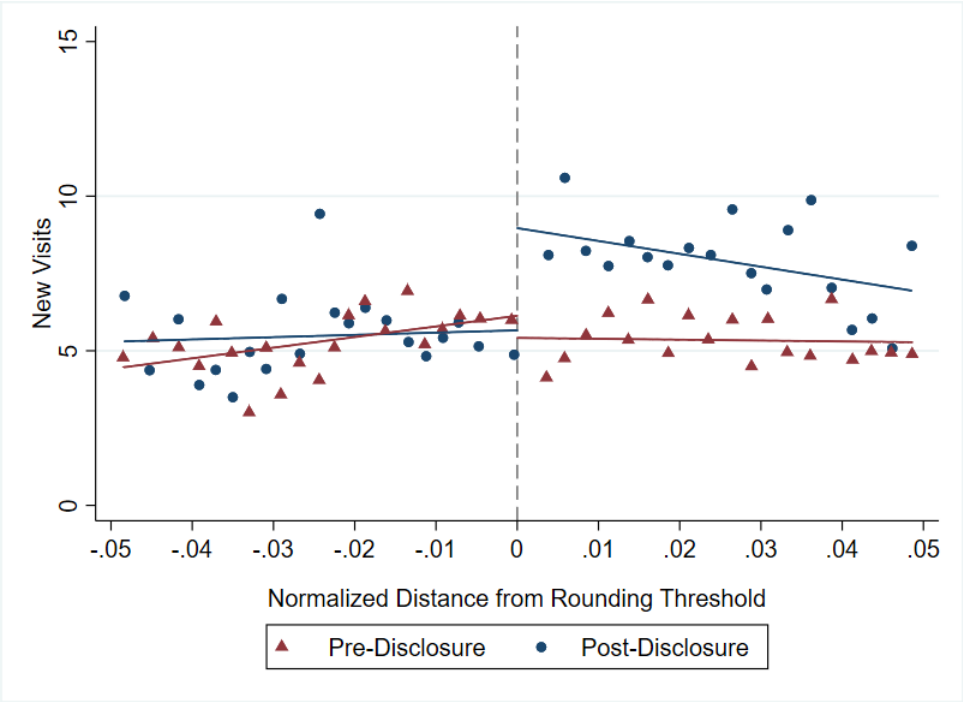
Although physicians A & B have similar raw ratings, the discrete rounding rule causes physician A to be displayed with 4.7 stars and physician B to be displayed with 4.8 stars.

Figure 3: Demand Response to Quality Disclosure



Note: Figure presents a binned scatterplot of the new visits per month at a family medicine provider, given the distance of that provider to the nearest star rating rounding threshold. Distances to nearest thresholds are pooled across the cutoffs and normalized to the nearest threshold and observations are weighted by count of reviews. Superimposed on the binned scatterplot are best-fit linear regression lines on both sides of the cutoff.

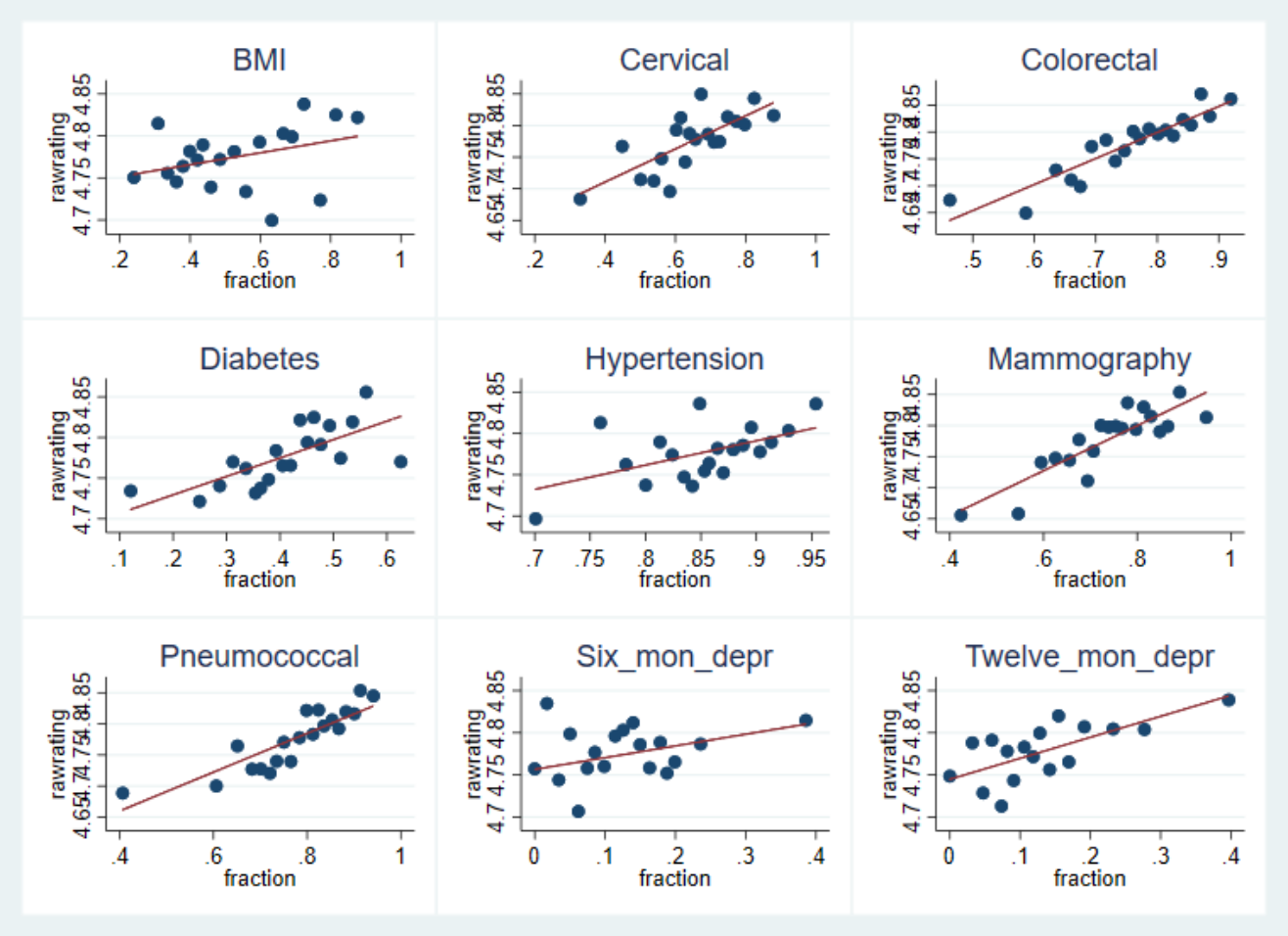
Figure 4: Demand Response to Quality Disclosure, Difference in Discontinuities



Note: Figure presents a binned scatterplot of the new visits per month at a family medicine provider both before the online ratings were disclosed (red triangles) and after online ratings were disclosed (blue dots), given the distance of that provider to the nearest star rating rounding threshold. Distances to nearest rounding thresholds are pooled across the cutoffs and normalized to the nearest threshold and observations are weighted by count of reviews. Superimposed on the binned scatterplot are best-fit linear regression lines on both sides of the cutoff for both pre-disclosure (January 2017 to October 2018) and post-disclosure (December 2018 to August 2019) time windows.

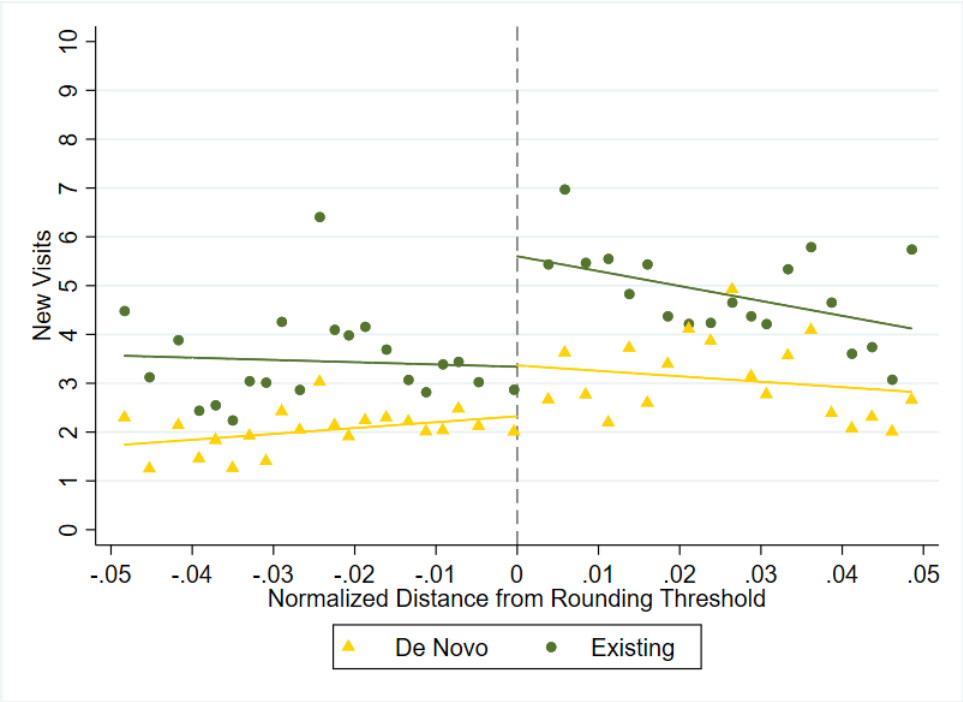
Figure 5: Relationship Between Star Ratings and Health Quality Metrics

(Vaccinations, Screenings, and Counseling)



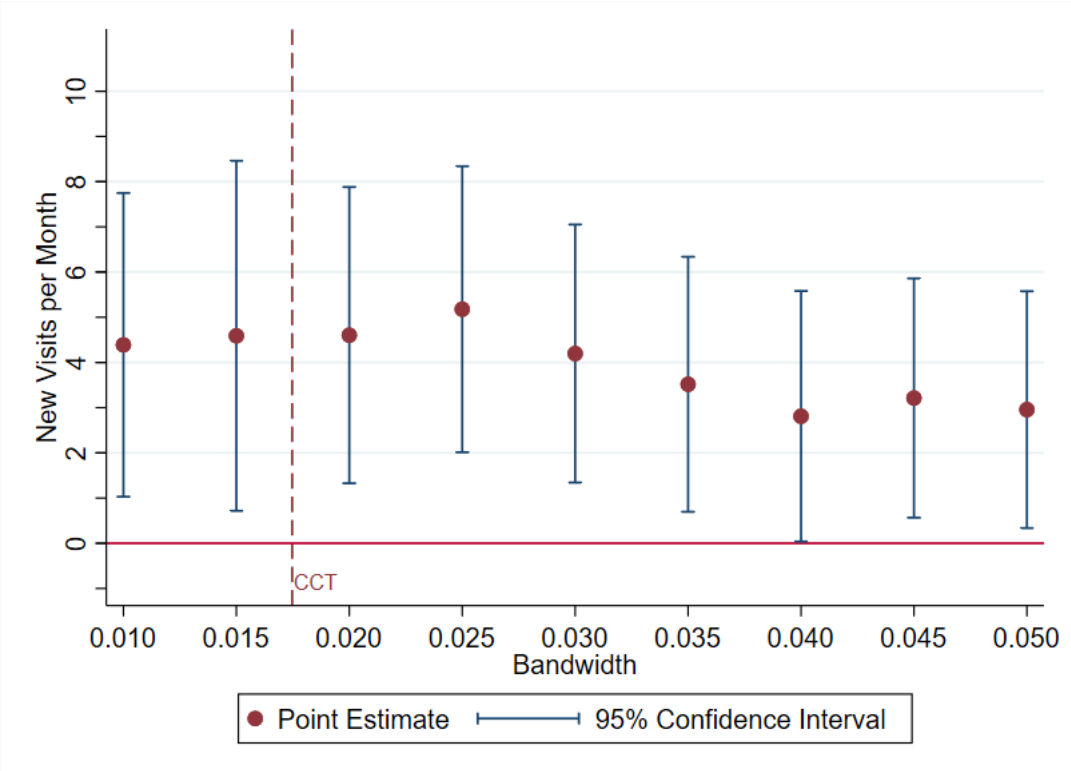
Note: Six_mon_depr and Twelve_mon_depr correspond to 6- and 12-month depression screenings. Fraction (x-axis) corresponds to fraction of the time the provider performs these vaccinations, screenings, and counseling on patients who are indicated for them. For example, the denominator for mammography is only women in the age range recommended by the government for mammography. These quality metrics are used internally by the health system to measure quality of family medicine. I only have one time period of these provider quality metrics available, so I cannot exploit time variation in quality metrics to estimate regression discontinuity models.

Figure 6: Market Expansion vs. Switching



Binned scatterplot of new visits per month at family medicine providers, separately by whether the patient is *de novo* at the health system or already had exisiting exposure to other providers in the health system. Observations weighted by count. Data plots post-disclosure period only.

Figure 7: Effects by Bandwidth



Note: Figure plots effect sizes from the baseline regression specification. Standard errors are clustered on the provider. The red dashed line denotes the mean-squared-error minimizing bandwidth of Calonico, Cattaneo, and Titiunik (CCT).

Table 1: Summary Statistics

Patient Level

	Mean	*Median*	*SD*
Age	38.76	36.86	24.49
BMI	27.51	26.98	8.26
B.P. (systolic)	118.87	119.45	13.83
B.P. (diastolic)	72.06	72.00	9.27
Race = White	0.89		
N (Visits)	12,575,190		
N (Patients)	998,244		

Provider-Month Level

	Mean	*Median*	*SD*
Monthly New Visits	7.34	4.00	10.08
Monthly Visits	178.48	172.00	94.34
Rating Score (continuous)	4.78	4.82	0.13
Rating Count (Dec '18)	228.55	206.50	127.30
Rating Count (Aug '19)	298.28	264.00	171.59
Physicians share (MD/DO)	0.55		
Mid-level practitioner share	0.45		
Distinct providers	340		
N (Provider-Months)	2,730		
height			

Note: Patient level data comes from EHR and provider-month data comes from the EHR merged with the ratings data. Provider–month level data is restricted to family medicine providers only.

Table 2: Outcome: Monthly New Visits (OLS)

	(1)	(2)	(3)	(4)
Displayed Rating Score				
(...,4.5,4.6,4.7,...)	-16.48***	-16.52***	-16.67***	-16.71***
	(3.365)	(3.369)	(3.614)	(3.619)
Controls:				
Month-Year FE		X		X
Professional Credential FE			X	X
Observations	2,730	2,730	2,730	2,730

Note: Standard Errors clustered at the provider level and observations weighted by review count. Restricted to Family Medicine providers. Professional Credential FEs include MD, PA, CNP, APR, DO, and other professional credentials.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Monthly New Visits - Family Medicine

	(1)	(2)	(3)	(4)	(5)	(6)
Rounded Up	2.978**	2.958**	3.850**	2.956**	4.287**	5.550**
	(1.347)	(1.336)	(1.542)	(1.332)	(1.738)	(2.352)
Functional Form:	Linear	Quad.	Cubic	Linear	Quad.	Cubic
Treatment Interaction	No	No	No	Yes	Yes	Yes
Cutoff FEs	Yes	Yes	Yes	Yes	Yes	Yes
Mean Below Threshold	5.475	5.475	5.475	5.475	5.475	5.475
% Change	54.4	54.0	70.3	54.0	78.3	101.4
Observations	2730	2730	2730	2730	2730	2730

Note: Standard Errors clustered at the provider level and observations weighted by review count. Treatment Interaction refers to an indicator permitting different slopes on each side of the discontinuty.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Difference-in-Discontinuities

	New Visits per Month
Post x Rounded Up	4.496***
	(1.244)
Rounded Up	-1.414
	(0.899)
Distance to threshold	19.38
	(20.37)
Dist x Rounded Up	-36.53
	(28.10)
Post	-0.940
	(0.713)
Post x Distance	-46.15*
	(26.96)
Post x Dist x Rounded	0.689
	(45.41)
Mean below threshold	5.100
% Change	88.2
Observations	7762

Standard errors clustered at the provider level.
and observations weighted by count. Restricted to
family medicine providers and specification is
linear with interaction. See text for pre/post dates.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Monthly New Visits - By Leading 5 Specialties

	(1) Family Med	(2) Pediatrics	(3) Internal Med	(4) Cancer	(5) OB/GYN
Rounded Up	2.956**	0.0532	-3.983*	2.055	-2.086
	(1.332)	(1.394)	(2.271)	(3.219)	(2.231)
Distance to threshold	-26.92	17.80	-22.07	-16.42	-50.78
	(24.86)	(28.06)	(61.38)	(94.48)	(102.6)
Dist × Rounded	-35.84	-94.96*	54.79	-113.9	134.4
	(45.82)	(51.64)	(94.15)	(141.2)	(156.8)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes
Mean below threshold	5.475	4.805	5.914	14.664	14.060
% Change	54.0	1.1	-67.3	14.0	-14.8
Observations	2730	983	529	657	499

Standard errors clustered at the provider level & observations weighted by count.
Preferred specification is linear trend plus interaction. Bandwidth (-.05,.05)
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Monthly New Visits - By Patient Age Groups

	(1) Age 18-34	(2) Age 35-49	(3) Age 50-64	(4) Age 65-79	(5) Age 80+
Rounded Up	1.194**	0.688**	0.593**	0.291**	0.0881
	(0.535)	(0.321)	(0.268)	(0.134)	(0.0616)
Distance to threshold	-11.72	-4.922	-7.703	-4.601	-2.570**
	(7.630)	(5.488)	(5.002)	(3.129)	(1.293)
Dist × Rounded	-16.02	-10.95	-5.895	1.034	1.549
	(15.63)	(11.11)	(9.022)	(4.837)	(2.205)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes
Mean below threshold	1.576	1.105	1.020	0.479	0.165
% change	75.8	62.2	58.2	60.8	53.4
Observations	2529	2529	2529	2529	2529

Standard errors clustered at the provider level & observations weighted by count.
Preferred specification is linear trend plus interaction.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Monthly New Visits - By Patient Health Status

	Healthy			Sick		
	(1)	(2)	(3)	(4)	(5)	(6)
	Zero Comorb.	Non-Obese	Nonmoker	Comorbid	Obese	Smoker
Rounded Up	2.867**	1.952**	2.337**	0.357**	1.271***	0.887**
	(1.227)	(0.974)	(0.997)	(0.160)	(0.453)	(0.414)
Distance to threshold	-38.28	-25.32	-34.37*	-4.022	-16.99**	-7.933
	(24.23)	(19.34)	(20.23)	(3.352)	(8.497)	(8.244)
Dist × Rounded	-15.29	-10.86	-7.786	-4.661	-9.095	-12.16
	(42.99)	(33.43)	(36.13)	(5.978)	(16.31)	(13.50)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes	Yes
Mean below threshold	5.303	4.082	4.206	0.558	1.780	1.655
% Change	54.1	47.8	55.5	63.9	71.4	53.6
Observations	2529	2529	2529	2529	2529	2529

Standard errors clustered at the provider level & observations weighted by count.
Preferred specification is linear trend plus interaction.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Monthly New Visits - By Provider Credentials

	(1)	(2)
	MDs	Not MDs
Rounded Up	4.203**	0.506
	(1.981)	(1.838)
Distance to threshold	-11.39	-20.86
	(31.76)	(40.00)
Dist × Rounded	-75.87	-10.09
	(62.48)	(68.77)
Cutoff FEs	Yes	Yes
Mean below threshold	4.120	7.847
% Change	102.0	6.5
Observations	1363	1367

SEs clustered at the provider level
Weighted by rating count. Bandwidth (-.05,.05).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Monlthy New Visits, by Geographic Density of Family Medicine Providers

	(1) Low Density	(2) High Density	(3) Low Density	(4) High Density
Rounded Up	1.927	4.079*	2.166	4.769***
	(1.495)	(2.393)	(1.859)	(1.692)
Distance to threshold	-26.12	-35.75	-49.54	-52.51*
	(37.17)	(36.38)	(41.97)	(30.23)
Dist × Rounded	0.241	-56.49	9.092	-21.20
	(63.18)	(70.21)	(70.04)	(58.62)
Cutoff FEs	Yes	Yes	Yes	Yes
Mean below threshold	5.864	5.705	5.864	5.705
% Change	32.9	71.5	36.9	83.6
Observations	1389	1186	1361	1214

Note: Standard Errors clustered at the provider level and observations weighted by review count. Columns 1-2 compute physician density using all physicians included in the Area Health Resource File, and columns 3-4 use only health system physicians. Density calculations explained in section 6.2.4. Model includes cutoff FEs.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Wait Days for Appointment, New Patients

	(1) Pre-info	(2) Post-info	(3) Diff-in-Disc
Post x Rounded Up			2.695***
			(0.886)
Rounded Up	-0.850	2.105***	-0.847
	(0.612)	(0.704)	(0.593)
Distance to threshold	37.17	-3.134	42.11
	(29.22)	(35.38)	(28.17)
Dist x Rounded Up	-43.02	-71.44	-51.67
	(40.93)	(50.15)	(39.47)
Post			-1.224**
			(0.611)
Post x Distance			-52.69
			(42.68)
Post x Dist x Rounded			1.310
			(60.77)
Cutoff FEs	Yes	Yes	Yes
ICD Diagnosis Code FEs	Yes	Yes	Yes
Mean below threshold	8.896	8.765	8.848
% Change	-9.6	24.0	30.5
Observations	13300	8745	22045

Unit of observation is a patient visit. Restricted to Family Medicine specialty, patients 18+, and dropping new visits scheduled greater than 180 days out. Regression is unweighted and inference is done with robust standard errors and bandwidth is [-.025,.025].

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Wait Days for Appointment, Established Patients

	(1) Pre-info	(2) Post-info	(3) Diff-in-Disc
Post x Rounded Up			1.736***
			(0.240)
Rounded Up	-0.549***	1.163***	-0.547***
	(0.132)	(0.203)	(0.131)
Distance to threshold	42.34***	-36.35***	45.23***
	(6.425)	(10.01)	(6.401)
Dist x Rounded Up	-50.01***	43.14***	-50.76***
	(9.055)	(14.26)	(9.025)
Post			-0.858***
			(0.171)
Post x Distance			-94.30***
			(11.76)
Post x Dist x Rounded			101.4***
			(16.73)
Cutoff FEs	Yes	Yes	Yes
ICD Diagnosis Code FEs	Yes	Yes	Yes
Mean below threshold	13.462	14.513	13.793
% Change	-4.1	8.0	12.6
Observations	448285	205788	654073

Unit of observation is a patient visit. Restricted to Family Medicine specialty, patients 18+, and dropping new visits scheduled greater than 180 days out. Regression is unweighted and inference is done with robust standard errors and bandwidth is [-.025,.025].

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Wait Days for Urgent Conditions

	(1) Pre-Disclosure	(2) Post-Disclosure
Rounded Up	-0.771	2.374**
	(1.090)	(1.096)
Distance to threshold	35.36	-165.9**
	(65.34)	(82.94)
Distance X Rounded	-40.76	132.2
	(71.19)	(99.88)
Cutoff FEs	Yes	Yes
ICD FEs	Yes	Yes
Observations	1124	650

Wait Time (residualized) for conditions indicated by Billings, et al. (2000) to be ED Care Needed but Preventable/Avoidable
Preferred specification is linear trend plus interaction.
Bandwidth (-.05,.05) and robust SEs.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Monthly New Visits - Observations Restriction to Specified Distance from Cutoff

	(1) (-0.05,0.05)	(2) (-0.04,0.04)	(3) (-0.03,0.03)	(4) (-0.02,0.02)	(5) (-0.01,0.01)
Rounded Up	2.956**	2.810**	4.197***	4.603***	4.389**
	(1.332)	(1.409)	(1.450)	(1.663)	(1.699)
Distance to threshold	-26.92	-30.79	-111.0	-83.52	-180.6
	(24.86)	(36.07)	(75.10)	(77.04)	(126.0)
Dist × Rounded	-35.84	-17.80	35.27	-50.41	210.0
	(45.82)	(62.75)	(98.74)	(137.0)	(217.7)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes
Mean below threshold	5.475	5.457	5.901	5.427	5.052
% Change	54.0	51.5	71.1	84.8	86.9
Observations	2730	2204	1611	987	440

Standard errors clustered at the provider level & observations weighted by rating count. Specification is linear trend plus interaction.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 14: Monthly New Visits - Family Medicine: Effect of Weighting by Rating Count

	(1) No Weighting	(2) Weight by Count	(3) Weight by Inv Count	(4) No Weighting	(5) Weight by Count	(6) Weight by Inv Count
Rounded Up	2.978**	2.978**	5.704*	2.943**	2.956**	5.602*
	(1.468)	(1.347)	(3.150)	(1.442)	(1.332)	(3.022)
Distance to threshold	-40.21*	-45.83**	-58.90	-21.62	-26.92	-18.49
	(21.85)	(21.35)	(36.37)	(29.06)	(24.86)	(42.65)
Dist × Rounded				-35.71	-35.84	-78.12
				(57.89)	(45.82)	(99.89)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes	Yes
Mean below threshold	10.856	6.652	8.826	10.856	6.652	8.826
% Change	27.4	44.8	64.6	27.1	44.4	63.5
Observations	2730	2730	2730	2730	2730	2730

SEs clustered at the provider level. Cols. 1-3 are linear trend, 4-6 linear plus interaction.

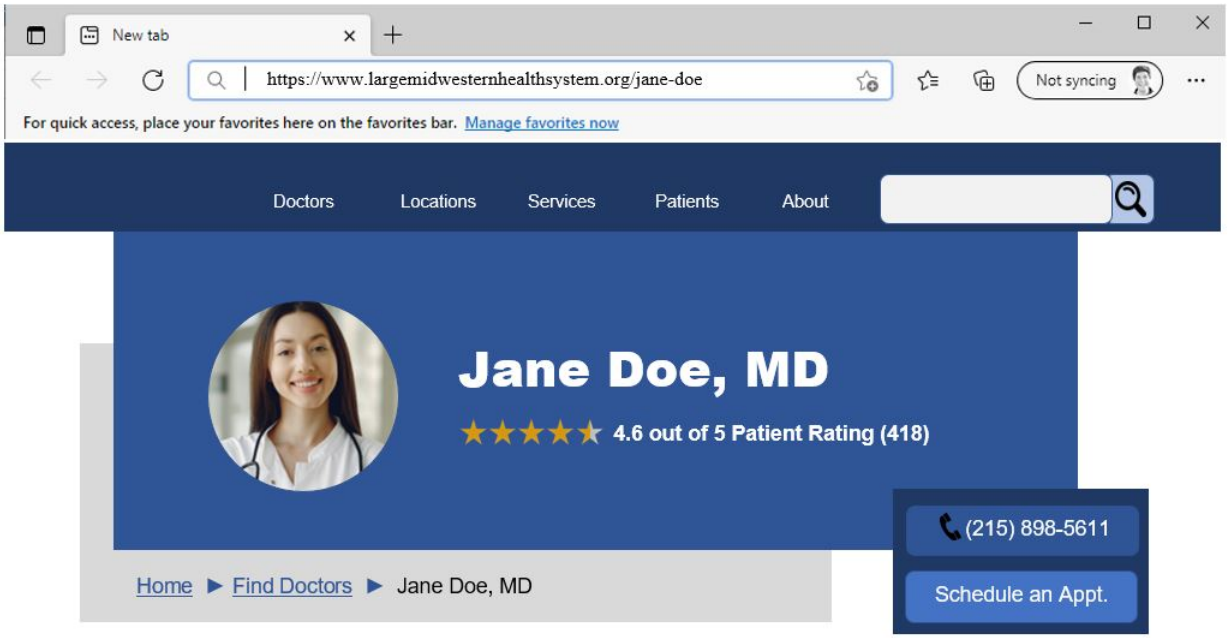
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix

A Example of Provider Quality Score Disclosure and Survey Questions

This figure shows an artistic rendition of what a new patient would see when he or she visited the health system's website to search for a new provider after November 2, 2018. Note the 4.6 out of 5 (ratings rounded to the nearest one-tenth) and N=418 ratings, along with the gold stars. The regression discontinuity design captures the causal effect of increasing a provider's score by exploiting the rounding of raw averages to discrete binned intervals. Prior to disclosure, the website looked the same, but without the star ratings.

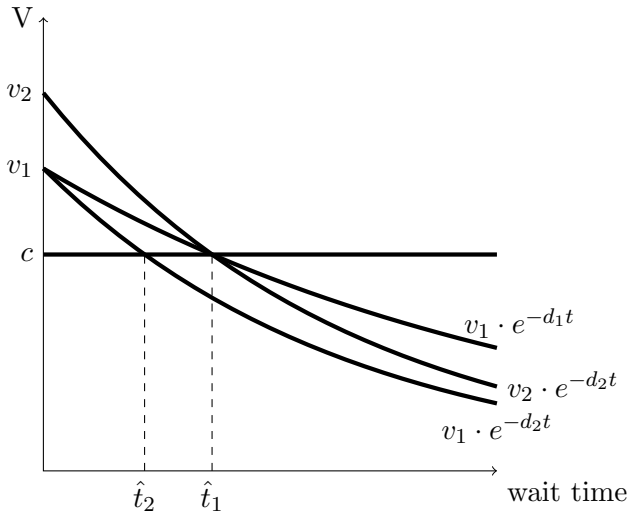
Figure A1: Sample Physician Rating Webpage



Survey Questions:

1. Did this provider explain things in a way that was easy to understand?
2. Did this provider listen carefully to you?
3. Did this provider give you easy to understand instructions about taking care of these health problems or concerns?
4. Did this provider seem to know the important information about your medical history?
5. Did this provider show respect for what you had to say?
6. Did this provider spend enough time with you?
7. Using any number from 0 to 10, where 0 is the worst provider possible and 10 is the best provider possible, what number would you use to rate this provider?

Figure A2: Relationship Between Benefits and Costs of Waiting



These figures show the RD separately for each distinct rounding threshold in the rating scale (See Table 3 for pooled regression with cutoff fixed effects), restricting to the majority of providers with displayed ratings of 4.6 and up. Separate best fit lines are fitted for each Panel (A) shows the relationship between rating and new visit volume before information was disclosed, and Panel (B) shows the relationship after disclosure. Vertical lines indicate rounding thresholds.

Figure A3:

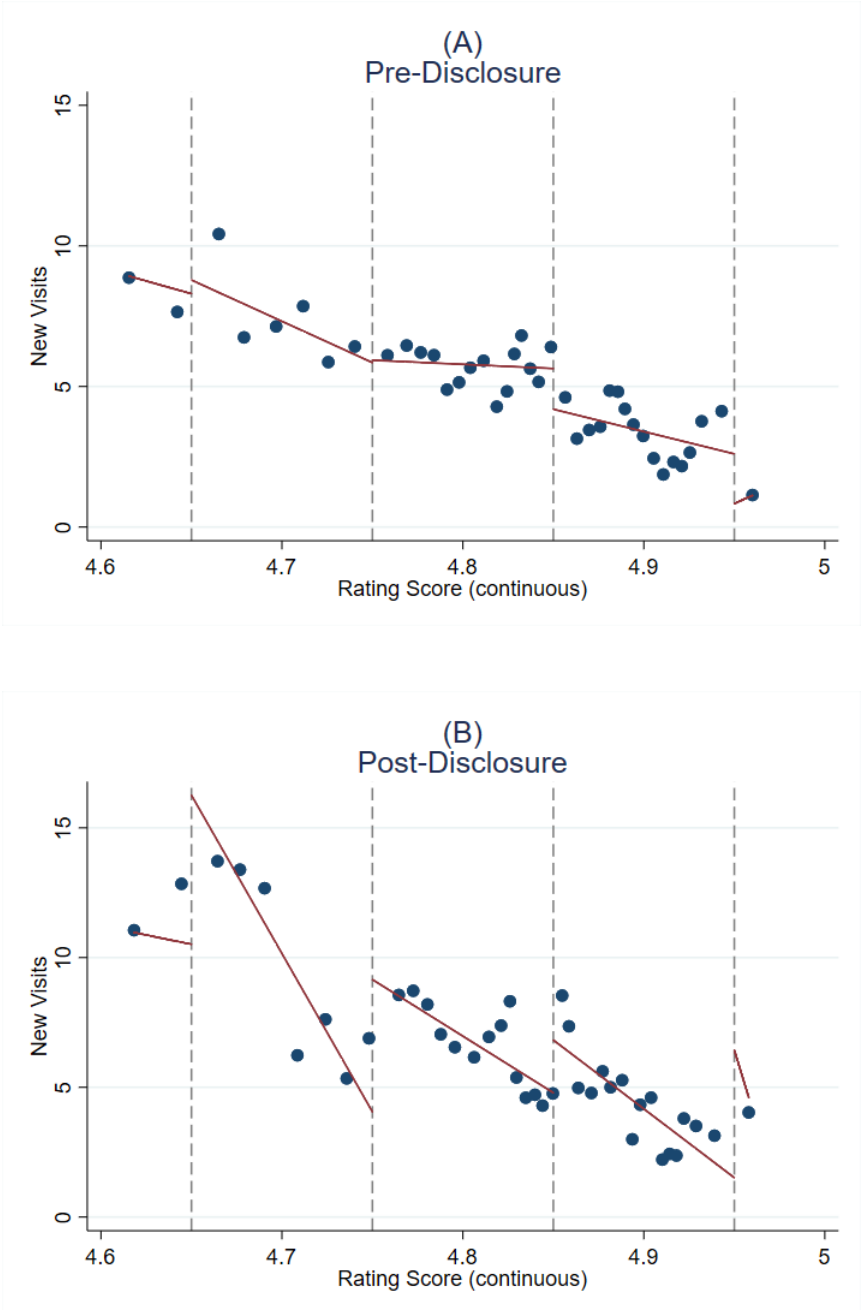


Figure A4:

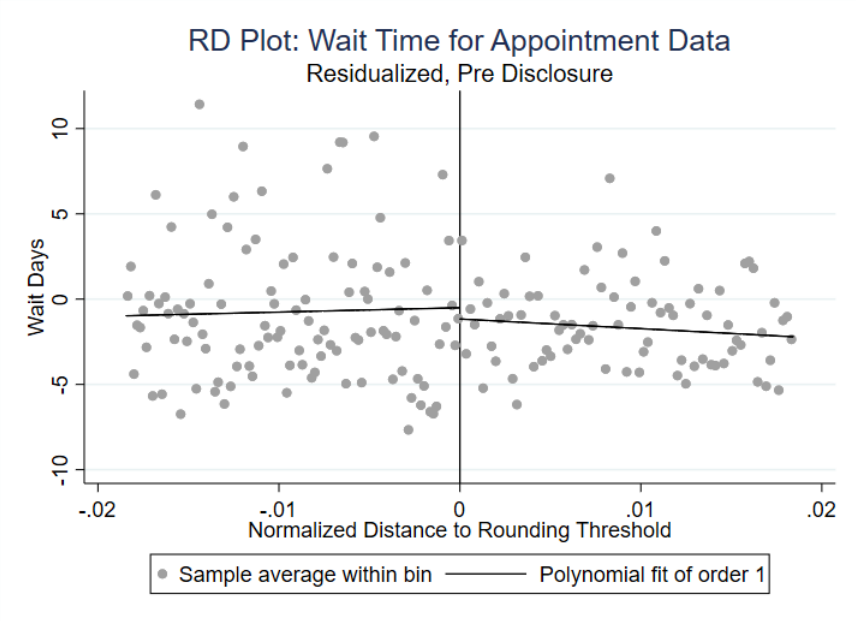


Figure A5:

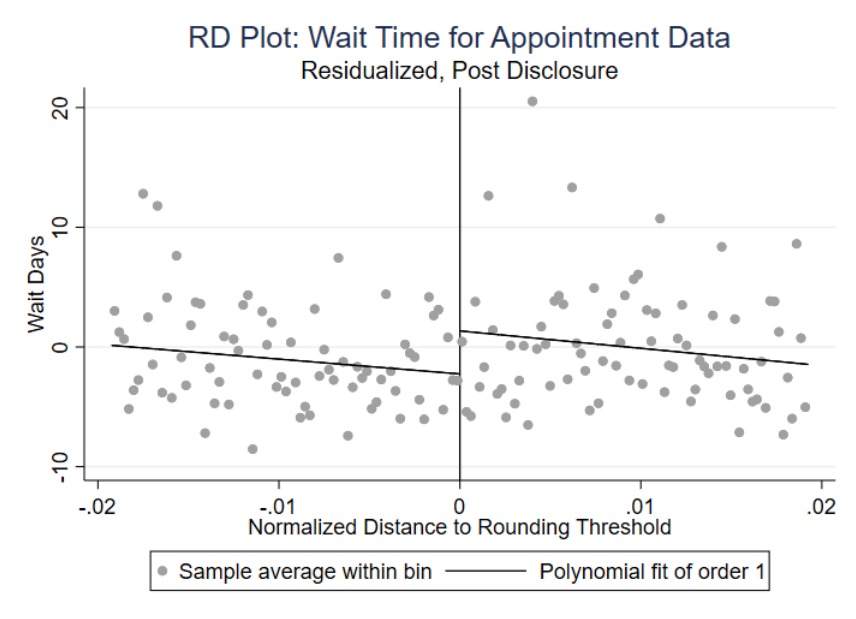
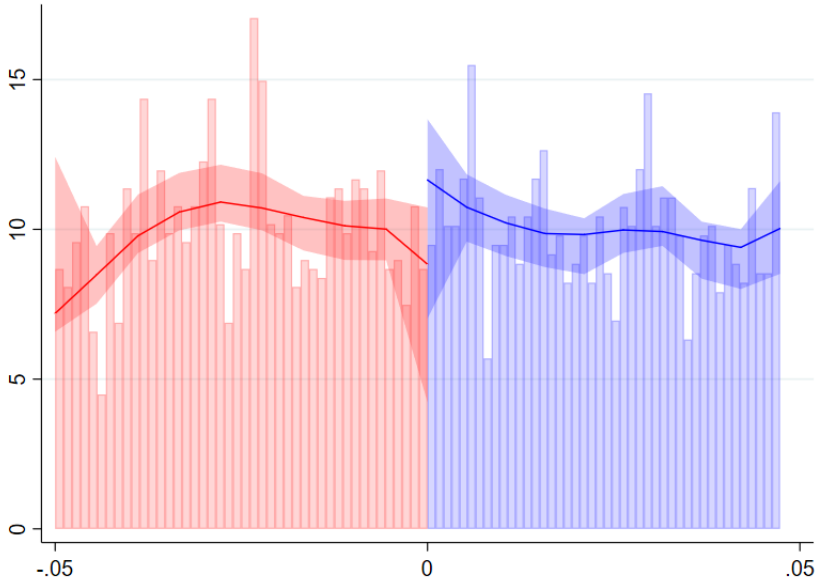
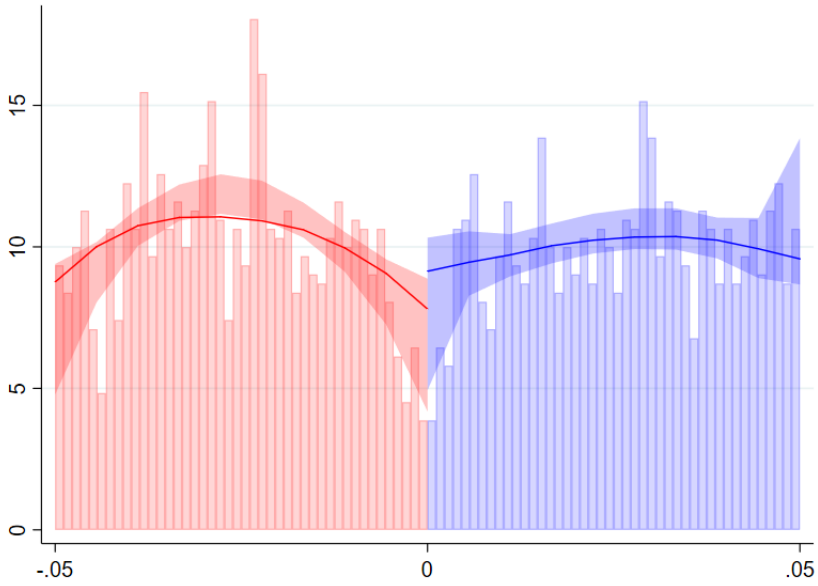


Figure A6: Manipulation Testing Plot



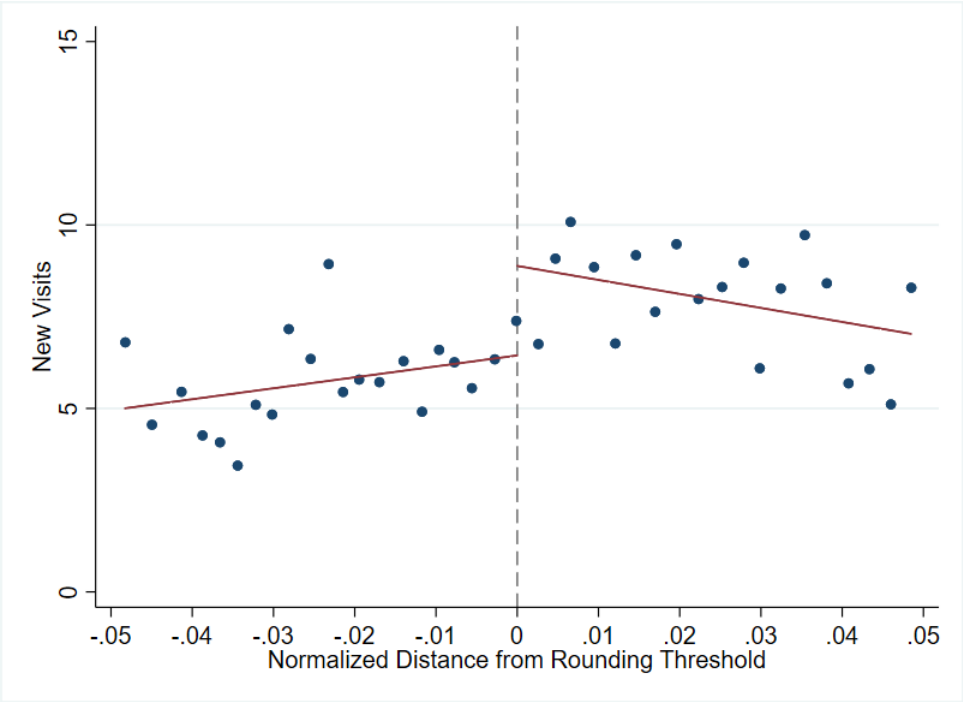
Note: Density test of the running variable, keeping provider–month observations with more than one displayed rating per month

Figure A7: Manipulation Testing Plot



Note: Density test of the running variable, dropping provider–month observations with more than one displayed rating per month

Figure A8: Demand Response to Quality Disclosure



Binned scatterplot, data restricted to family medicine physicians, but not dropping observations with more than one displayed rating per month. Compare to Fig. 3 which drops panel observations displaying more than one rating per month.

Figure A9: Covariate Balance on Baseline Regression (Provider–Month Panel)

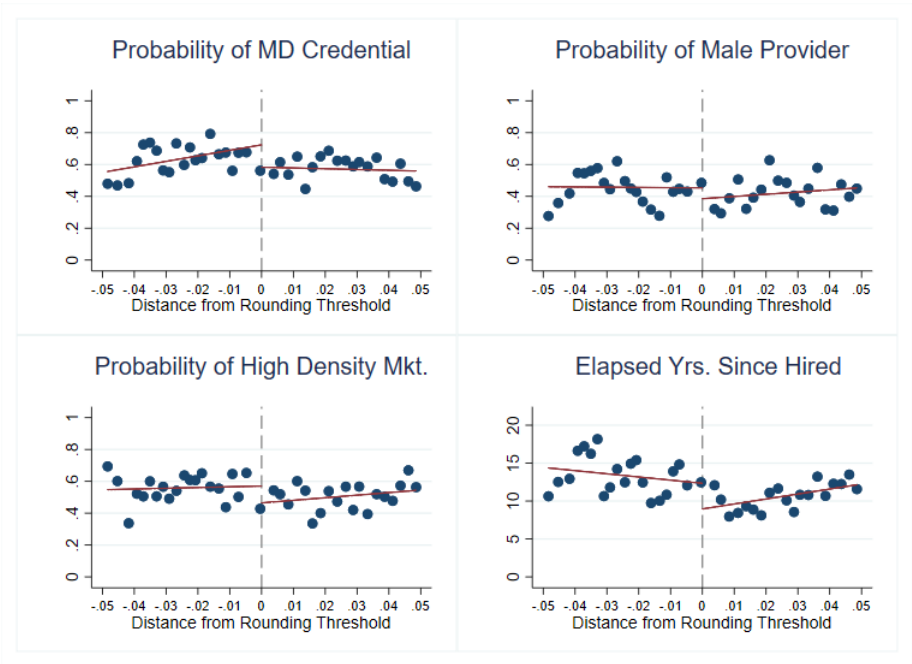


Table A1: Monlthy New Visits, by Count of Family Medicine Providers

	(1) Low Count	(2) Medium Count	(3) High Count	(4) Low Count	(5) High Count
Rounded Up	3.363**	4.788	0.877	2.651	2.330
	(1.353)	(4.742)	(2.658)	(1.981)	(1.738)
Distance to threshold	-57.36*	-74.48	-3.390	-45.84	-10.62
	(31.46)	(72.15)	(47.96)	(39.25)	(35.83)
Dist × Rounded	24.08	10.04	-119.7	48.59	-68.73
	(56.52)	(194.4)	(77.13)	(85.11)	(60.16)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes
Mean below threshold	5.625	6.609	5.798	5.370	6.145
% Change	59.8	72.4	15.1	49.4	37.9
R-squared	0.140	0.375	0.204	0.137	0.143
Observations	1750	231	594	1365	1210

Note: Standard Errors clustered at the provider level and observations weighted by review count. Columns 1-3 compute physician counts using all physicians included in the Area Health Resource File, while columns 4-5 use only the health system's physicians. Model includes cutoff FEs

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A2: DeNovo = No, New Visits - Family Medicine

	(1) Linear	(2) Quadratic	(3) Cubic	(4) Linear	(5) Quadratic	(6) Cubic
Rounded Up	2.070**	2.063**	3.418***	2.059**	3.954***	4.917***
	(0.880)	(0.873)	(1.112)	(0.870)	(1.269)	(1.679)
Distance to threshold	-35.72**	-35.28**	-92.27**	-26.25	-108.6	-330.8**
	(14.99)	(14.59)	(40.46)	(17.75)	(96.48)	(163.5)
Dist × Rounded				-17.94	-64.90	186.0
				(33.11)	(136.2)	(273.6)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes	Yes
Mean below threshold	3.454	3.454	3.454	3.454	3.454	3.454
% Change	59.9	59.7	99.0	59.6	114.5	142.4
Observations	2730	2730	2730	2730	2730	2730

Note: Standard Errors clustered at the provider level and observations weighted by review count.
Columns 1-3 parameterize same slope on both sides of disconinuity, 4-6 do not.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: DeNovo = Yes, New Visits - Family Medicine

	(1) Linear	(2) Quadratic	(3) Cubic	(4) Linear	(5) Quadratic	(6) Cubic
Rounded Up	0.908	0.895	0.432	0.896	0.333	0.633
	(0.647)	(0.641)	(0.581)	(0.641)	(0.625)	(1.000)
Distance to threshold	-10.11	-9.399	10.10	-0.665	-20.15	-69.12
	(8.905)	(8.585)	(21.78)	(9.496)	(45.84)	(82.79)
Dist × Rounded				-17.91	85.54	122.1
				(17.81)	(79.51)	(220.3)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes	Yes
Mean below threshold	2.021	2.021	2.021	2.021	2.021	2.021
% Change	44.9	44.3	21.4	44.4	16.5	31.3
Observations	2730	2730	2730	2730	2730	2730

Note: Standard Errors clustered at the provider level and observations weighted by review count.
Columns 1-3 parameterize same slope on both sides of disconinuity, 4-6 do not.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Residualized new Patient Wait Days , Regression Discontinuity Estimates

Table A4: Local Linear Regression with Optimal Bandwidths

	False Discontinuity 0.025		False Discontinuity -0.025		True Discontinuity 0.000	
	Pre	Post	Pre	Post	Pre	Post
	(1)	(2)	(3)	(4)	(5)	(6)
RD_Estimate	2.006	1.941	0.225	-1.452	-0.906	2.450***
	(1.375)	(1.376)	(0.968)	(1.168)	(0.786)	(0.859)
MSE-Optimal Bandwidth	0.007	0.007	0.006	0.008	0.018	0.019
Mean Below Threshold	7.499	7.319	7.426	6.890	9.203	8.703
% Change	26.8	26.5	3.0	-21.1	-9.8	28.1
Observations	25643	16760	25643	16760	25643	16760

Note: Regressions denoted `Pre` corresponds to before quality disclosure and `Post` corresponds to after disclosure. This table reports the regression discontinuity estimate from optimal bandwidth local linear regression using the `rdrobust` package in Stata (Colonico, et. al. 2017). Left hand side variable RD_Estimate is a residualized wait time in days for a new patient visits for adults not going to walk-in clinics. The outcome is residualized prior to estimation with an OLS regression with cuttoff-specific and presenting diagnosis specific fixed effects (e.g., Lee, 2010) and standard errors are HC0 robust.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Monthly New Visits - Family Medicine

	(1)	(2)	(3)	(4)	(5)	(6)
Rounded Up	2.192**	2.163**	2.445**	2.157**	2.563*	2.861*
	(1.104)	(1.093)	(1.212)	(1.089)	(1.346)	(1.647)
Functional Form:	Linear	Quad.	Cubic	Linear	Quad.	Cubic
Treatment Interaction	No	No	No	Yes	Yes	Yes
Cutoff FEs	Yes	Yes	Yes	Yes	Yes	Yes
Mean Below Threshold	5.725	5.725	5.725	5.725	5.725	5.725
% Change						
Observations	2941	2941	2941	2941	2941	2941

Note: Standard Errors clustered at the provider level and observations weighted by review count. Treatment Interaction refers to an indicator permitting different slopes on each side of the discontinuty. Sample does not exclude providers who display more than 1 rating/month.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6: Covariate Balancing:

	(1) MD Credential	(2) Male Provider	(3) High Density	(4) Elapsed Tenure
Rounded Up	-0.134	-0.0577	-0.0930	-3.319
	(0.104)	(0.121)	(0.117)	(2.078)
Functional Form:	Linear	Linear	Linear	Linear
Treatment Interaction	Yes	Yes	Yes	Yes
Cutoff FEs	Yes	Yes	Yes	Yes
Mean Below Threshold	0.636	0.456	0.558	13.377
% Change	-21.1	-12.6	-16.7	-24.8
Observations	2730	2637	2575	2730

Note: Standard Errors clustered at the provider level and observations weighted by review count. Treatment Interaction refers to an indicator permitting different slopes on each side of the discontinuty.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A7: Monthly New Visits - Family Medicine

	(1)	(2)	(3)	(4)	(5)	(6)
Rounded Up	3.333**	3.306**	3.180**	3.306**	3.349*	4.982**
	(1.410)	(1.406)	(1.611)	(1.404)	(1.823)	(2.514)
Functional Form:	Linear	Quad.	Cubic	Linear	Quad.	Cubic
Treatment Interaction	No	No	No	Yes	Yes	Yes
Cutoff FEs	Yes	Yes	Yes	Yes	Yes	Yes
Mean Below Threshold	5.475	5.475	5.475	5.475	5.475	5.475
% Change	60.9	60.4	58.1	60.4	61.2	91.0
Observations	2730	2730	2730	2730	2730	2730

Note: Standard Errors clustered at the provider level and observations weighted by review count. Treatment Interaction refers to an indicator permitting different slopes on each side of the discontinuty.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.