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

Benjamin L. Chartock*

Department of Economics,
Bentley University

May 12, 2025

Abstract

What does introducing quality ratings do? Ratings may shift consumers towards higher rated sellers while simultaneously causing congestion. I find evidence of these effects studying the introduction of a universal quality rating disclosure policy for doctors. Using a difference-in-discontinuities design, I show that disclosure causes 54% more patient demand at higher rated doctors yet patients wait 3 days longer for one standard deviation higher quality. Many markets including health care rely on waiting rather than prices to allocate scarce quality, and in such environments, quality disclosure benefits some patients but not others.

*I am grateful to Abby Alpert, Yiqun Chen, Ginger Z. Jin, Jonathan Kolstad, Claudio Lucarelli, Matt Notowidigdo, Mark Pauly, Ellie Prager, Maggie Shi and seminar participants at the NBER Spring Health Care Meeting, The Becker-Friedman Institute at the University of Chicago, the Harvard–MIT–BU Health Economics Seminar, ASSA annual meeting, ASHEcon, the Midwest Health Economics Conference, Indiana University, WVU, Lafayette College, and Lehigh University for valuable seminar comments and discussion. 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. All errors are my own. 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 (Chevalier and Mayzlin 2006), motivating sellers (Kolstad 2013), and ameliorating adverse selection by stimulating competition (Jin and Leslie 2003; Kaye et al. 2024). On the other hand, disclosure can have unintended negative consequences, such as inducing inefficient effort on the part of suppliers (Dranove et al. 2003), causing multitasking problems (Holmstrom and Milgrom 1991; Feng Lu 2012), or exacerbating differences across income gradients (Brown et al. 2023). Although the literature has numerous studies about the effect of quality disclosure on market outcomes (see, e.g., Scanlon et al. 2002; Jin and Sorensen 2006; Chernew et al. 2008; Dafny and Dranove 2008; Dranove and Sfekas 2008; Pope 2009; Bundorf et al. 2009; Schwartz et al. 2021), a major 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, a glut of buyers may arrive seeking 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 at the point-of-sale for many in-network providers regardless of quality. In the absence of a price, wait times may serve as an equilibrating factor to clear the market (Richards-Shubik et al. 2021).

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. 43% of adults aged 50-80 report looking at doctor ratings online according to a 2019 University of Michigan National Poll on Healthy Aging (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 include Amazon for retail products, Yelp for restaurants, or AirBnB for vacation rentals), ratings may be particularly relevant in the market for family medicine and primary care because patients typically have a large choice set of potential providers and their insurance benefits often require 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 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 effects. These three outcomes encompass a range of possible impacts of quality disclosure in equilibrium. I study these effects by building a novel data set comprised of a combination of electronic health records (EHRs) and the universe of online doctor reviews that was collected and only later disclosed by a large, integrated health system in the United States with more than 40 hospitals and nearly 1,500 employed physicians. I also adopt a tailored identification strategy suited to the setting at hand.

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 first 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 disclosure event that I study have a number of attributes that make it an ideal laboratory for exploring these questions about quality 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 (for example, see Appendix Figure A1). In addition, the manner in which ratings are gathered from patients differs from other well-known online sources such that these ratings are likely of higher reputability than other star ratings. Ratings disclosed in this setting 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, other websites allow any person (patient or not) to submit a review of a provider. The random sending of surveys to patients in my setting eliminates much of the selection bias that arises due to which individuals are contributing to online ratings. There is also considerably lower availability of other sources of online quality data regarding medical providers in the health system's region (e.g., from HealthGrades, ZocDoc, and Yelp), which suggests that this quality disclosure represents a major, if not the foremost, source of information about providers. Unlike on other websites, the quality disclosure that I study applies universally to all providers; no provider can opt out of having their rating disclosed nor can they pay to have a more prominent placement on the platform like on ZocDoc.

The unique data source is also an advantage 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 directly observe 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.

I find several main 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 (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 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 specialists), highlighting the role of referrals in consumer choice of specialist providers. 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 largely silent on heterogeneity in ratings effect by age.

In addition to these findings about the demand response to quality disclosure, I provide evidence of effects on market clearing in the absence of prices. Specifically, I examine consequences of disclosure on three dimensions of sorting: (1) examining whether information disclosure shifts patients to physicians who supply greater inputs to health, (2) 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 better physicians. One common criticism about doctor ratings is that stars do not reflect actual provider quality but instead reflect orthogonal concerns such as the freshness of the magazines in the waiting room or quality of the fish tank in the lobby. In contrast to these concerns, I document evidence that the online disclosure sorts patients to providers who more frequently perform medically-recommended inputs to health such as vaccinations, screenings, and behavioral health services. Few, if any, studies find positive correlation between health care ratings (subjective) and medical measures of quality (objective). 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 moves 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, e.g., restaurants, which can raise the price of their entree when they get a higher Yelp score, physicians employed by a health system cannot charge a higher copay if they are a 5-star doctor relative to a 4-star. In this health system, the patient pays the same out-of-pocket price for family medicine irrespective of subjective quality. This parallels insurance network design broadly. 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 to ration the scarce quality. I document that congestion is occurring at high-quality sellers, and that this congestion is affecting both new patients (who wait 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 underscores 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 new 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. Econometrically, this congestion effect biases my main estimates downwards; in the absence of congestion, the demand effects would presumably be even larger.

Taken as a whole, these results paint a multidimensional picture of the economic consequences of quality disclosure as a remedy to markets with information frictions. As markets in health care and beyond increasingly adopt star ratings (such as CMS Hospital Compare) and quality certification becomes a mainstream method to ameliorate market woes caused by imperfect information, market designers will face trade-offs between increasing ease of shopping for experience goods and inducing congestion at high-quality sellers. This tradeoff suggests that quality disclosure kickstarts 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 describes the data and institutional setting and Section 3 presents the empirical strategy. Section 4 presents the results, including potential mechanisms and robustness checks. Section 5 concludes.

2 Institutional Setting and Data

2.1 The Large Midwestern Health System

To acquire data for this study, I partnered with 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 if the health system is in-network for any of the major insurers in the region, the patients would have full access to all health system providers.[1] This uniform insurance coverage importantly shuts down the role of out-of-pocket price in patient choices conditional on the insured choosing to receive care at the system. All patients pay the same price for a family doctor regardless of whom they select.

2.2 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 CG-CAHPS (Clinician and Groups Consumer Assessment of Healthcare Providers and Systems). This is a private–public partnership meant to develop surveys which elicit valid responses about patients perceptions of care. 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?". About 5% of total outpatient visits are followed up with a completed survey. 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.[2]

2.3 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

[1]The majority of the health system's doctors are compensated on a work relative value unit (RVU) schedule.

[2]The full list of survey questions is found in Appendix Section A.3. Details of the scaling transformation performed by the health system and their survey provider are available from me upon request.

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.

2.4 Summary Statistics & Sample Construction

In Figure 1, I document a negative relationship between physician star rating and number of new patients per month. If patients like quality, as conventional economic models predict, the empirical relationship between quality scores and new visit volume should be positive, not negative. The figure shows a binned scatterplot reflecting the conditional expectation function of new visits as a function of a doctor's star rating, with and without controls for month-year and physician type (MD, NP, etc.). A one-tenth star increase, e.g., from 4.7 to 4.8, is associated with 1.6 *fewer* new patients per month. One hypothesis that explains the inverse relationship is that good doctors are somewhat like absorbing states. If they are high quality, it is also likely their panel is full (that they are not accepting new patients) and thus high quality doctors see fewer new patients per month[3]. But the effect that I am interested

[3]A t-test between doctors who are and are not accepting new patients, shown in Appendix Table A12, indicates a 0.043 ($p < 0.05$) difference in star ratings between the full panel doctors (avg. 4.835) and the accepting new patients doctors (avg. 4.792).

in is the causal effect of disclosure on demand, i.e., the marginal effect of a star. My causal design is described in Section 3.

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. 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 the data suggests a typical patient composition. 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. 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.[4] The average provider rating is a 4.78 and the standard deviation is 0.13. A histogram of doctor average ratings is available in Figure 2. Half of providers have a rating that is rounded up, and the other half have a rating that is rounded down. At the time quality disclosure was launched, the average count of reviews used to determine the average score of a provider was 228. As more surveys came in, the average rating count increased to 298. 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 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.

[4] A top regional competitor posts star ratings and has a similar distribution of average provider ratings but does not post star ratings for all providers, perhaps because it does not employ most providers.

2.5 Pre- and Post-Disclosure

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 in an Excel file 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 while 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. According to the health system's disclosure policy, which is common across the United States for hospitals, doctors with fewer than 30 ratings were not displayed until they reached the 30-rating minimum. The health system used a 2-year look-back window to late 2016 so ratings did not have N=1 at launch. 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.

For each provider, I have information on listed specialty from the system website, their professional licensing credential (e.g., MD, RN, PA, etc.), provider gender, and national provider identifier (NPI). These data come from hospital human resources data and the health system website. Using the entire history of individual patient surveys, I reconstruct the mean raw 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.[5] In the next section, I discuss my empirical strategy to identify the causal effect of star ratings.

3 Empirical Strategy

3.1 Baseline Regression Discontinuity

I use regression discontinuity methods to compute the effect of an increased provider rating on demand for new patient visits (Angrist and Lavy 1999; Lee and Lemieux 2010; 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, website ratings are displayed 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 (reviewed in Appendix Figure A2). 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. Doctors may 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.

[5] 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 first series of regressions are based on the classical regression discontinuity estimator:

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

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. as 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), β_2 is the coefficient on the distance to the nearest rounding threshold, and β_3 is the coefficient on the interaction between the running variable and being rounded up, allowing 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.[6] The linear polynomial is preferred because a visual examination of the binned

[6] 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.

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.

3.2 Difference-in-Discontinuities

The second series of estimators I construct are difference-in-discontinuities estimators. 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.[7] 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 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.

[7] 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.

3.3 Difference-in-Differences

Another potential way of conceptualizing measurement of the effects of ratings on demand is through a difference-in-differences framework. Because I have a pre- and post-disclosure period, I implement a standard difference-in-differences regression comparing doctors with rounded up ratings to doctors with rounded down ratings before and after the disclosure. The specification is

$$Y_{it} = \beta_0 + \beta_1 POST_{it} + \beta_2 ABOVE_{it} + \beta_3 POST_{it} \times ABOVE_{it} + \varepsilon_{it} \tag{3}$$

where Y_{it} represents new visits for doctor i in month t, $POST_{it}$ indicates after Nov 2018, and $ABOVE_{it}$ indicates the physician has a rating above the halfway point on the star rating interval measured at one decimal place. A potential drawback to this approach is that there is no natural treatment and control group (all physicians are exposed to the introduction of ratings). In order to exploit the time variation without relying on the continuity assumption of RD, I construct a treatment group based on doctors having an extra star displayed due to website rounding but do not rely on information about how close the physician is to being rounded up or down. This difference-in-differences approach is used as a robustness check for the main analysis, and complements the difference-in-discontinuity design in the previous section, which combines a DiD approach with an RD design.

4 Results

4.1 Information Disclosure and Demand Response

4.1.1 Baseline Regression Discontinuity

In Figure 3, I show 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. Points to the left of the vertical dashed line in the figure 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. 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. 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. 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 2, I provide a regression-based estimate of the causal impact of an increased provider rating on new patient visits. Columns 1 to 6 of Table 2 present various alternative specifications of Equation (1): linear, quadratic, and cubic in the running variable and allowing for vs. not allowing for alternative slopes on each side of the discontinuity. In particular, Column 4 of Table 2 corresponds with the best fit lines in Figure 3, the baseline regression discontinuity graph. Based on the absence of a non-linear relationship between the running variable and the outcome variable in Figure 4, 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. 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. Compared to two related papers in the literature (Kaye et al. (2024) and Brown et al. (2023)), these estimates are larger but comparable.[8]

[8]Specifically, in the first paper, the authors study ZocDoc.com in a setting that involves rounding to the nearest half-star and find an approximately doubling of patient volume across the cutoff from 4.5 to 5 stars. Scaling this effect of 100% from a 0.5 jump down to a tenth-star corresponds to a 20% increase

I next show that the coefficient estimates for the effect of ratings on demand are insensitive to the bandwidth used (Figure 4). This is akin to implementing local linear regression. Point estimates are in red and error bars are in blue. Moving from right to left on this figure represents going from wider to narrower bandwidths. Across all bandwidths, the results are statistically significant. The bias–variance tradeoff is seen in the widening of confidence intervals as the bandwidth shrinks; however, regardless of the bandwidth the effects remain notable. The vertical dashed line represents the mean squared error optimal bandwidth estimator of Calonico et al. (2014) denoted by "CCT". Many consider this to be a conservatively narrow. The fact that the point estimates do not very much across bandwidths lends confidence to my estimates.

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

As a placebo test, I exploit the unique institutional setting in which the health care system collected ratings for more than two years prior to ever disclosing to patients. I plot two separate series in a single graph (Figure 5): 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 5 are striking. Before online information disclosure, 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

in demand, this is well within my confidence intervals. The other paper's estimates also fall within my estimated confidence intervals.

that to the left of the vertical dashed line, the blue dots and red triangles are commingled; in contrast, to the right of the rounding threshold, virtually all of the blue dots lie above the red triangles.

I estimate the causal effects that are shown in Figure 5 by using a difference-in-discontinuities regression and report the results in Appendix Table A1. This regression corresponds to Equation 2. 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. These regressions support the findings of a large demand response to disclosure.

Lastly, Figure 6 shows the discontinuity at each individual rounding threshold. The top panel shows the pre-disclosure data and the bottom panel shows the post-disclosure data. As in Figures 3 & 5, each bin represents the conditional mean in 40 equally sized bins with fitted lines to the left and right of each threshold. Before disclosure, the relationship between new patients and ratings is smooth across the discontinuity; after disclosure, at each threshold there is a stark jump in the number of new patient visits. The sawtooth pattern and jump at each discontinuity is similar to the study of Angrist and Lavy (1999) on classroom size where there were multiple cutoffs.

4.1.3 Difference-in-Differences

The results from the difference-in-differences regression are found in Appendix Table A2. The estimates suggests that the effect of an additional star is 2.786 new patients per month. This estimate is statistically significant and consistent in magnitude with the regression discontinuity estimates. Confidence intervals overlap with both the direct regression discontinuity estimates as well as the difference-in-discontinuity estimates.

4.2 Heterogeneity & Potential Mechanisms

In this section I explore heterogeneity analyses. Although not causal due to unobserved differences across sub-populations, these analyses shed light on potential mechanisms driving demand response to disclosure.

4.2.1 Provider Specialization, the Role of Choice versus Referrals, and Provider Credentials

In Table 3, 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 3 shows a 54% increase in the number of new patient visits per month for family medicine doctors (also reported in Table 2). This effect is large and statistically significant. In contrast, however, in columns 2-5 of Table 3, I do not find statistically significant causal effects on the amount of new patient visits for providers with different specialties (pediatrics, internal medicine, cancer, and OB/GYN).[9] This confirms the prior hypothesis that family medicine providers may be those whose demand is most impacted by rating disclosure. One potential explanation 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.

I also investigate whether provider credentials matter (e.g., MD vs. non-MD) and find the response to quality scores is concentrated among MDs (Appendix Table A7). This is at odds with the notion that ratings would be more important for providers with fewer credentials (i.e., reputation as a substitute for advanced degrees). I find that the types of patients who

[9] I drop a single Internal Medicine doctor with an exceedingly high monthly new patient count.

are seen by MDs are sicker on average: a t-test shows that MD patients have about 25% higher Charlson Comorbidity Scores than non-MD patients, consistent with patients having a stronger demand for highly rated physicians when they have more complex health needs (Appendix Table A8).

4.2.2 Older or Younger Patients? Healthy Patients or Sick Patients?

In Appendix Table A3, 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 as older patients visit new family medicine doctors at a much lower rate.

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. 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 2) 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 Kaye et al. (2024) 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 Appendix Table A4, 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, they are categorized as unhealthy, e.g., a diagnosis of COPD, dementia, or cancer, (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.[10]

Columns 1-3 of Appendix Table A4 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 Appendix Table A4 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

[10]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.

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.

4.3 Sorting

In the previous section, 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.

4.3.1 Inputs to Health

I show a positive relationship between stars and inputs to health in Figure 7. 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 giving antibiotics for viral ear infections). Doctors often complain about having scores posted online. (The most frequent critics are the low-rated providers.) The concern about doctors reallocating effort towards tasks based on alternative performance measures is detailed extensively by Feng Lu (2012).

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 classified by Donabedian as process measures (Dranove, 2011). Outcome measures (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.

Doctor performance on these metrics, such as mammography and colorectal cancer screenings, is measured only for clinically eligible patients (e.g., the mammography denominator is based only on women within the government age guidelines). I compare the propensity of a doctor to undertake recommended medical care to their average star rating. The relationships are plotted in Figure 7; 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.

4.3.2 Is Disclosure Causing Market Expansion or Switching?

In Appendix Figure A3, I assess whether disclosure is causing market expansion, switching, or both. 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). 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.

Appendix Figure A3 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 A6, 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 A5). I view these results 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.

I also find that evidence that patients who switch doctors may use their current providers score as a benchmark. In the pre-disclosure period, switchers move to higher-rated doctors (0.012 stars higher, $p<0.05$), consistent with wanting to switch to better doctors. However, post-disclosure, the increase after a switch is even higher (0.0414, $p<0.01$). Comparing the difference in differences via comparison of means suggests a statistically significant increase in the size of the jump when information was versus wasn't available of 0.029 stars ($p<0.01$). This evidence is consistent with patients using the score of their current provider as a benchmark for switching.

4.3.3 Congestion, Wait Times, and the "Price of a Star": Theory

I evaluate congestion by studying wait times: 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).[11] 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; and when wait time was less than 0 days (coding error). Across all visits (N=436,591), the mean wait time is 13.86 days. New patients have shorter waits on average (mean 9.10 days) whereas established patients mean wait time is longer (mean 14.09 days). About 5% of visits are new visits. Wait times for MDs are longer than non-MDs (16.29 days vs 11.00 days); visits to MDs correspond to 55.6% of visits.

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-

[11] The patient may see a different rating on the website when they book relative to when the appointment took place; to account for this lag, I use the rating on the website from the day of booking rather than the day of the visit.

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 3. These models assess whether patients wait longer to see a provider with a higher rating. The regression is similar to the model estimated earlier, but run at the individual visit level rather than provider-month level, and I also include a diagnosis code fixed effect to de-noise the regression and ease interpretation to wait days across all conditions (by residualizing 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 4, I restrict to narrow bandwidths on both sides of the cutoff of the normalized running variable, and report robust standard errors.

I find four main results about congestion. First, patients are willing to wait longer for an increase in quality rating. Second, both new and existing patients wait longer. Third, patients wait longer for both urgent and non-urgent care. And fourth, the congestion effect builds over time.

Columns (1) and (2) of Table 4 show the effect of a higher rating (β_1 from Eq. 1) on wait days for new patients before and after disclosure, respectively. Before the disclosure, there is no significant effect of a higher rating on days waiting for an appointment. After disclosure, new patients wait 2.453 days longer. I interpret this finding to represent a "shadow price of a star." That is, new patients are willing to wait 28% longer ($\frac{2.453}{8.703}$) 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. If I make the assumption that the effect size scales linearly as ratings increase, my estimate of a willingness-to-wait of 2.453 wait days for a 0.1 star increase represents a 3.12-day willingness-to-wait for a standard deviation increase in star rating (st. dev = 0.13).

As a robustness exercise in the spirit of Imbens and Lemieux (2008) and Eggers and Hainmueller (2009), I test for jumps at non-discontinuity points. This is found in Columns 3-6. I re-assign the cutoff to be 0.025 and -0.025 and find (as would be expected) effects that are statistically indistinguishable from zero. Only at the true threshold and only at disclosure do I observe an effect of ratings on waiting.

Columns 7-10 of Table 4 show that both new and established patients wait longer to see a doctor when their provider's rating is rounded up. Specifically, both new *and* established patients do not wait longer to see physicians before disclosure [Columns (7) and (8)] but do after disclosure [Columns (9) and (10)]. This highlights some of the losers of quality disclosure: patients who were previously seeing high-quality doctors before disclosure, but after disclosure, when new patients were able to observe quality, as well, these established patients needed to wait 0.634 days longer (4.5%, $\frac{0.634}{13.949}$) to see *the exact same provider* whom they were already seeing.

Next, I show that patients are willing to wait approximately the same length of time for a higher quality rating when seeking both urgent and seeking non-urgent care. I am agnostic about whether it is efficient for a patient to wait longer for a checkup because s/he likes the magazines in the lobby, but a tick bite or HIV test after a risky sexual encounter are types of conditions that if treated early (with an antibiotic or post-exposure prophylaxis) can stave off great expenses later. I test this in Columns 11 and 12 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.[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, patients are willing to wait 2.374 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

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

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. This coefficient value is nearly exactly the same as the value found for all care (not just urgent care), suggesting patients are waiting for quality when it might be inefficient.

Another important feature of congestion is that it builds over time. In the first weeks of a disclosure policy, if new demand is reallocated towards higher-rated sellers, it is unlikely that capacity constraints will be binding and also unlikely that congestion will be observed in the data. However, with time, as more patients move towards higher-rated physicians, eventually a mass of patients may build up. Accordingly, I would expect to see a congestion effect grow over time. To test this hypothesis, I split the post-disclosure period into two five-month two halves, early and late disclosure. I run the same regression discontinuity design with wait times on the left hand side for early and late periods and find suggestive evidence of a congestion effect that builds with time. In the early disclosure period, I observe no statistically significant increase in days patients wait due to a higher star rating (Column 13). However, in the later half of the disclosure period, patients do wait longer, about 2.36 days (Column 14). Although not statistically significant, the direction is consistent with patterns expected by this hypothesis.

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 might equilibrium be 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. Importantly, the potential impact of congestion that may occur concurrently with the introduction of a quality rating system may result in biased measurements of the true effect of a quality disclosure on demand. The direction of this bias can be signed downward (assuming consumers have a disutility from waiting); had there been no concurrent congestion, I would have expected the effects on demand to be even larger. The numerous papers that use a discontinuity design to estimate the demand response to star ratings may be systematically underestimating the effect of rating introduction in the absence of accounting for congestion. This econometric justification for potentially downward biased estimates of the impact of quality ratings extends beyond the health care setting to any market where congestion may occur.

4.4 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 vary minimally across some specifications, the direction and magnitude of my estimates holds up under the barrage of traditional regression discontinuity robustness tests. In fact, the first of these robustness tests are seen in Table 2, where I show that the results of the baseline regression discontinuity model are invariant to linear, quadratic, or cubic polynomial functional form, as well as in Figure 4, where I vary the bandwidth from wide to narrow and find the results are similar across bandwidth.

4.4.1 Covariate Balance

In Table 5, 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 five 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, the elapsed years since that provider started working at the health system, and previous month's visits). Appendix Figure A4 shows covariate balance across each of these available predetermined attributes. Physicians with ratings rounded up seem to be no different than physicians with ratings rounded down based on available predetermined observables.

4.4.2 Manipulation, Density Tests, Alternative Sample Definitions, and Weighting

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. 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). Figure 8 shows 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 8 plots this histogram for all the providers in the data, whereas Appendix Figure A3 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.

As an additional robustness check to make sure that the baseline regression results are robust to not dropping the provider-months which doctors cross the rounding threshold in a given month, I plot the regression discontinuity results for the sample where I do not drop these observations (Appendix Figure A6). The results are quantitatively and qualitatively similar to the baseline specification. Finally, 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. Appendix Table A11 shows the estimates from the Rounded Up coefficient of interest for the same six baseline specifications as the cutoff-specific fixed effects model. The estimates are comparable in both magnitude and direction to the baseline model across all specifications.

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 Appendix Table A10. 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).

5 Discussion & Conclusion

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 7 (relationship between star ratings and medical metrics) 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 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.

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. As a back-of-the-envelop exercise 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 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

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.

Brown, Z. Y., Hansman, C., Keener, J., and Veiga, A. F. (2023). Information and disparities in health care quality: Evidence from gp choice in england. Technical report, National Bureau of Economic Research.

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., 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., 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.

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

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.

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.

Kaye, A. P., Luca, M., and Vats, S. (2024). Differential impacts of online ratings in the market for medical services.

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.

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.

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

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.

6 Figures and Tables

Figure 1: Conditional Expectation Function, With and Without Controls

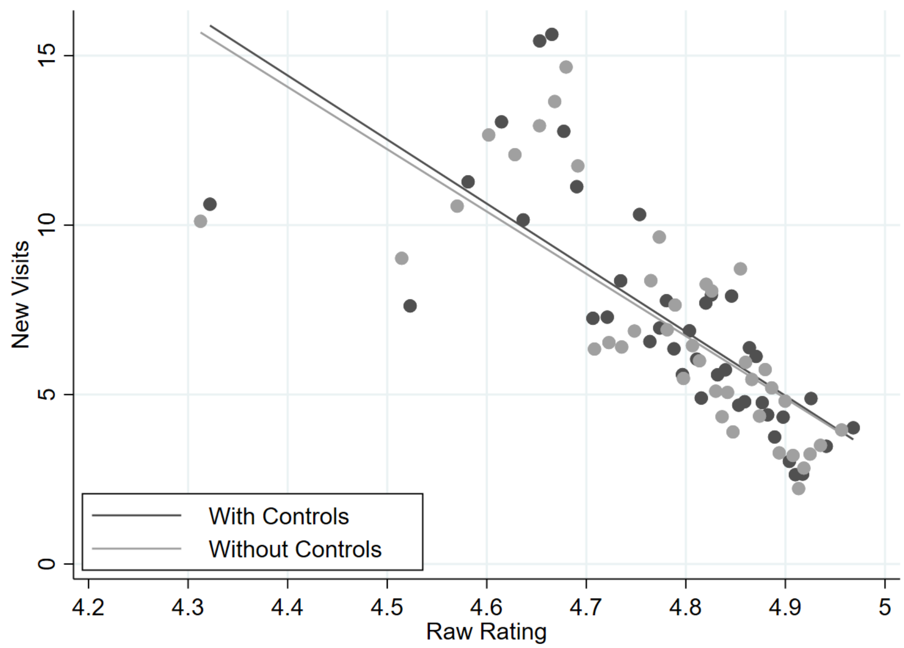


Figure 2: Distribution of Provider Average Ratings

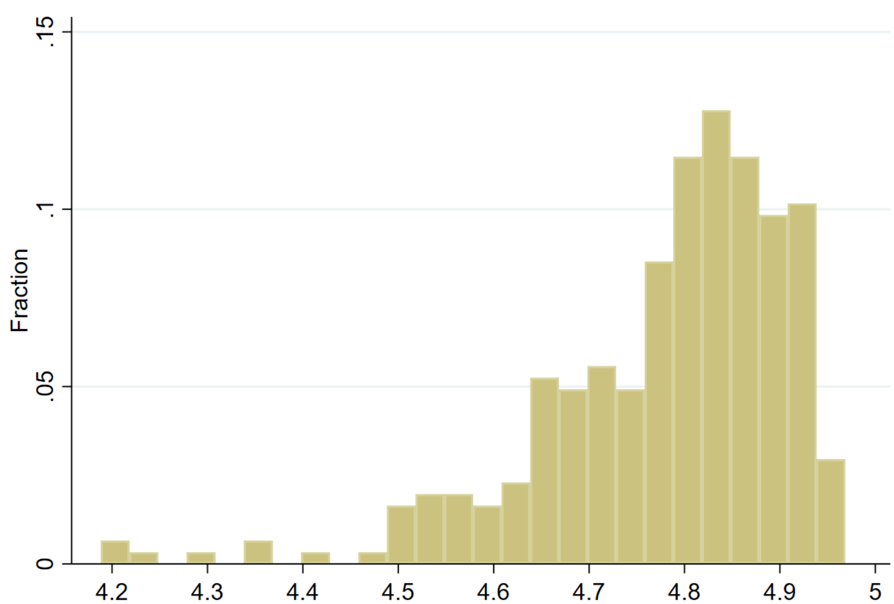


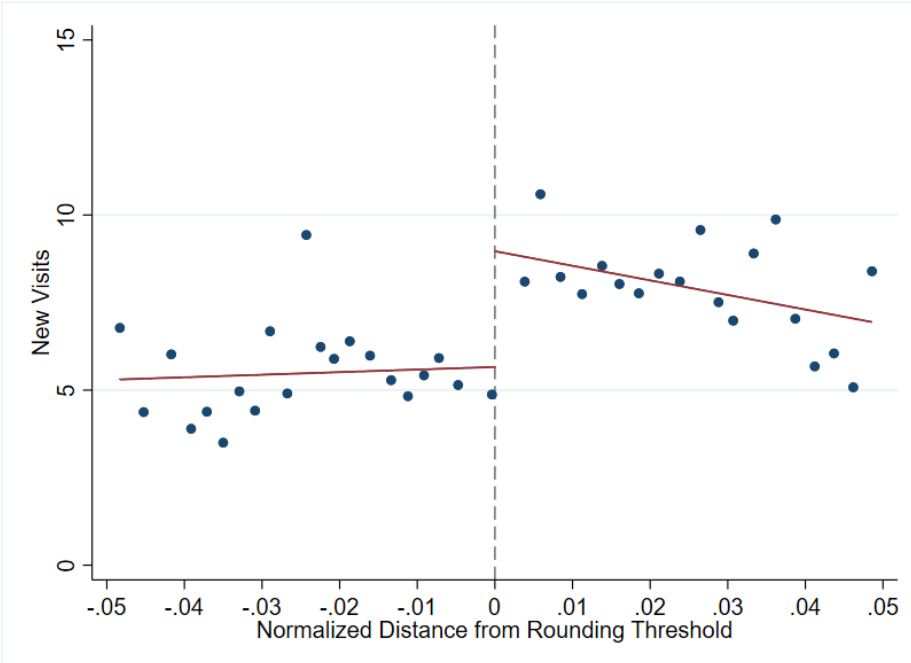
Table 1: Summary Statistics

Patient Level					
	Mean	*Median*	*SD*	*Min*	*Max*
Age	38.76	36.86	24.49	0	112
BMI	27.49	26.98	8.05	13.33	66.22
B.P. (systolic)	118.86	119.46	13.58	84	172
B.P. (diastolic)	72.06	72.00	9.06	46	102
Race = White	0.89				
N (Visits)	12,575,115				
N (Patients)	998,225				
Provider-Month Level					
	Mean	*Median*	*SD*	*Min*	*Max*
Monthly New Visits	7.34	4.00	10.08	0	118
Monthly Visits	178.48	172.00	94.34	1	692
Rating Score (continuous)	4.78	4.82	0.13	4.17	4.97
Rating Count (Dec ’18)	228.55	206.50	127.30	33	665
Rating Count (Aug ’19)	298.28	264.00	171.59	34	905
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.

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.

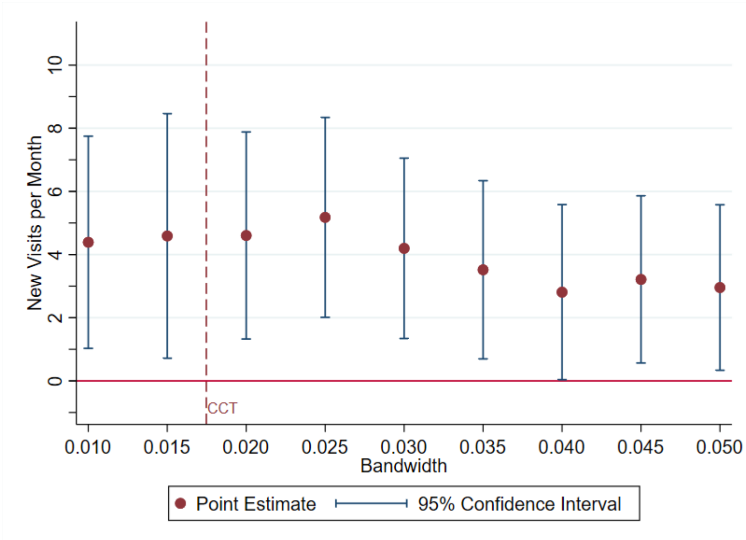
Table 2: Monthly New Visits - Family Medicine

	(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 discontinuity.

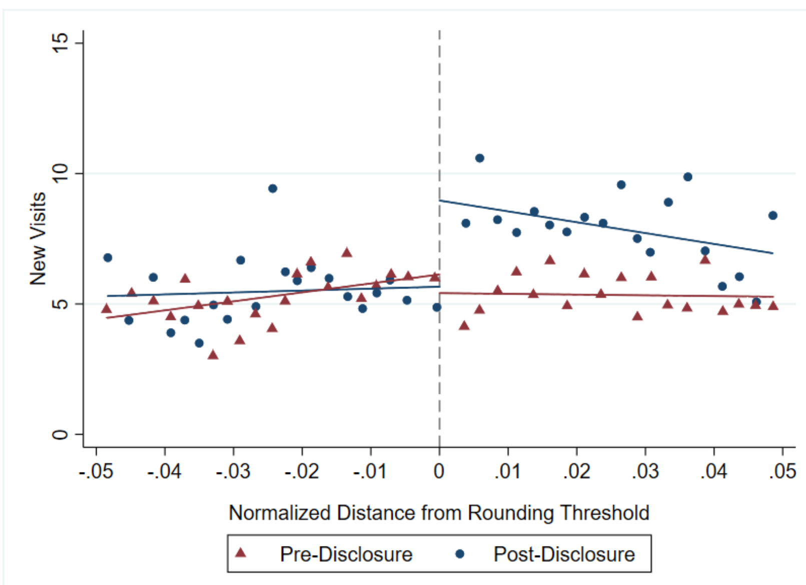
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 4: 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).

Figure 5: 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 6: Demand Response to Quality Disclosure, Difference in Discontinuities

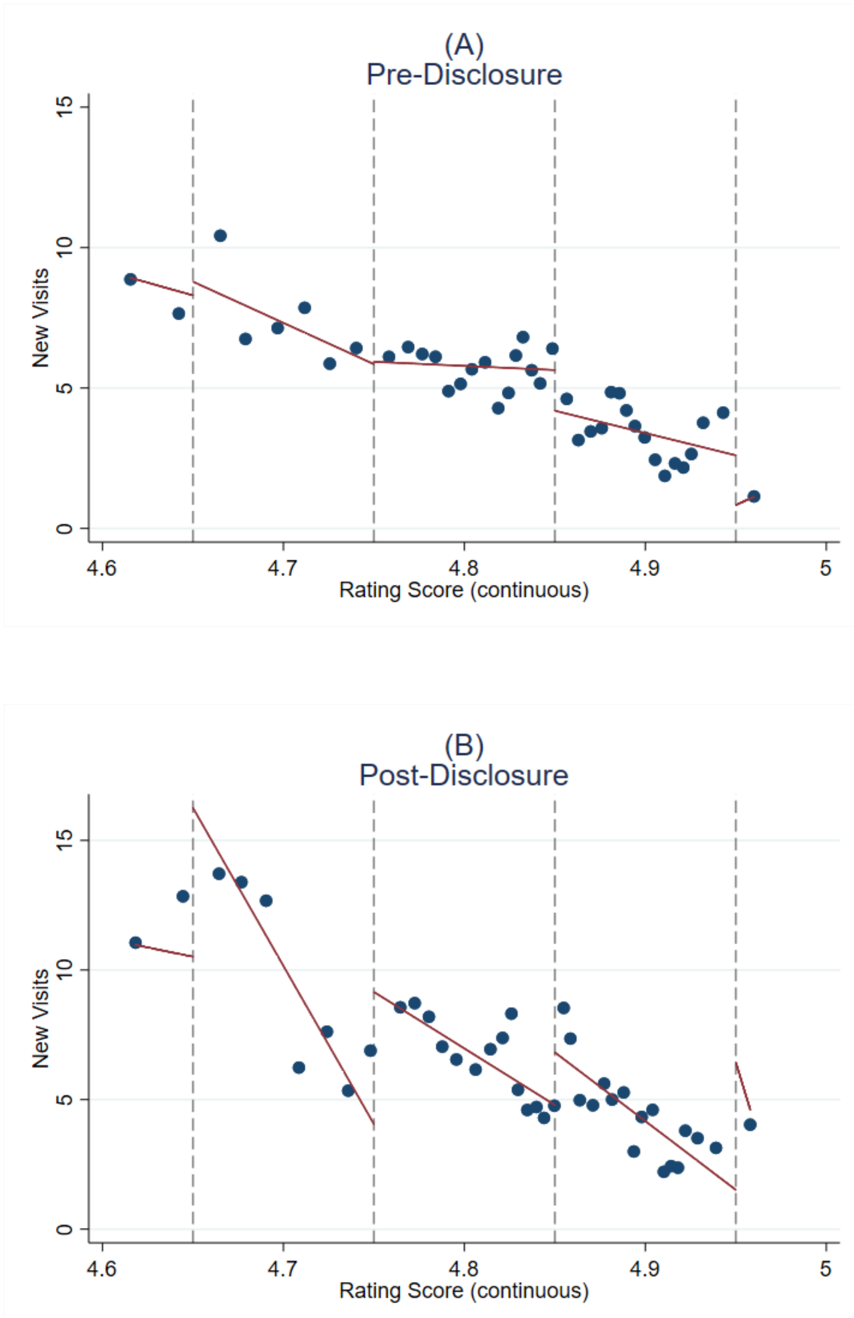
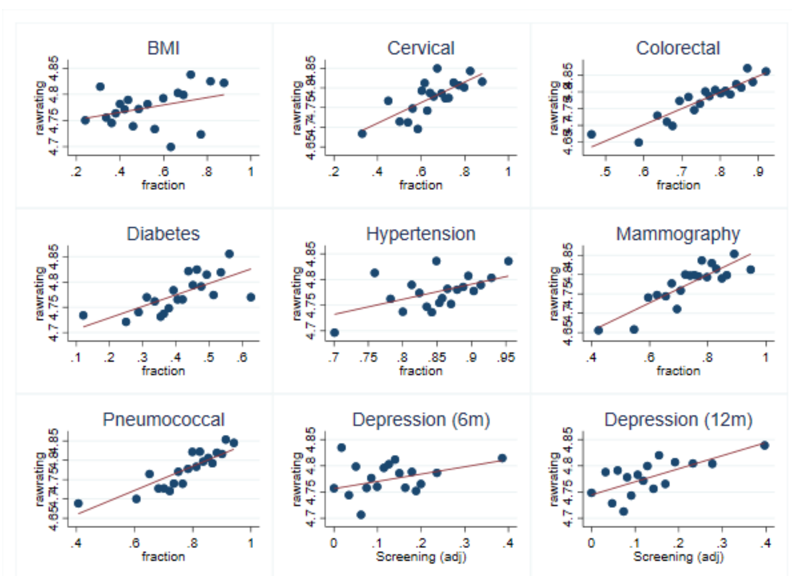


Table 3: Monthly New Visits - By Leading 5 Specialties

	(1) Family Med	(2) Pediatrics	(3) Internal Med	(4) Cancer	(5) OB/GYN
Rounded Up	2.956**	0.0532	-2.062	2.055	-2.086
	(1.332)	(1.394)	(1.441)	(3.219)	(2.231)
Distance to threshold	-26.92	17.80	-66.15	-16.42	-50.78
	(24.86)	(28.06)	(48.22)	(94.48)	(102.6)
Dist × Rounded	-35.84	-94.96*	99.53	-113.9	134.4
	(45.82)	(51.64)	(85.68)	(141.2)	(156.8)
Cutoff FEs	Yes	Yes	Yes	Yes	Yes
Mean below threshold	5.475	4.805	5.287	14.664	14.060
% Change	54.0	1.1	-39.0	14.0	-14.8
R-squared	0.115	0.134	0.220	0.271	0.028
Observations	2730	983	520	657	499

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

Figure 7: Relationship Between Star Ratings and Medical Metrics (Vaccinations, Screenings, and Counseling)



Note: Depression (6m) and Depression (12m) 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.

Table 4: Outcome: Congestion (Wait Days)

	Baseline RD		Placebos				New Vs. Established Patients				Urgent Conditions		Congestion Buildup	
			0.025		-0.025		Before		After					
	Before	After	Before	After	Before	After	New	Estab.	New	Estab.	Before	After	Early	Late
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
RD Est	-0.910	2.453***	2.004	1.881	0.157	-1.427	-0.459	-0.283**	3.320***	0.634***	-0.771	2.374**	-1.551	2.362
	(0.790)	(0.858)	(1.378)	(1.347)	(0.956)	(1.184)	(0.687)	(0.144)	(0.815)	(0.227)	(1.090)	(1.096)	(1.998)	(1.567)
Mean Below Threshold	9.200	8.703	7.518	7.476	7.511	7.157	8.216	12.975	8.072	13.949	3.398	3.598	8.388	8.056
% Change	-9.9	28.2	26.7	25.2	2.1	-19.9	-5.6	-2.2	41.1	4.5	-22.7	66.0	-18.5	29.3
Obs	25643	16760	25643	16760	25643	16760	11198	373217	7251	171094	1124	650	8021	8739

Note: The outcome variable (wait days between appointment booking and occurence) is residualized to adjust for covariates (diagnosis code and nearest rounding threshold). All regressions drop children under 18 and those who visit walk-in clinics. Columns (1)-(6) and (13)-(14) report optimal bandwdith local linear RD estimator (Colonico, et al., 2017). Remaining columns report local linear RD estimator with cutoff interaction and bandwidths ranging from .021 to .025.

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

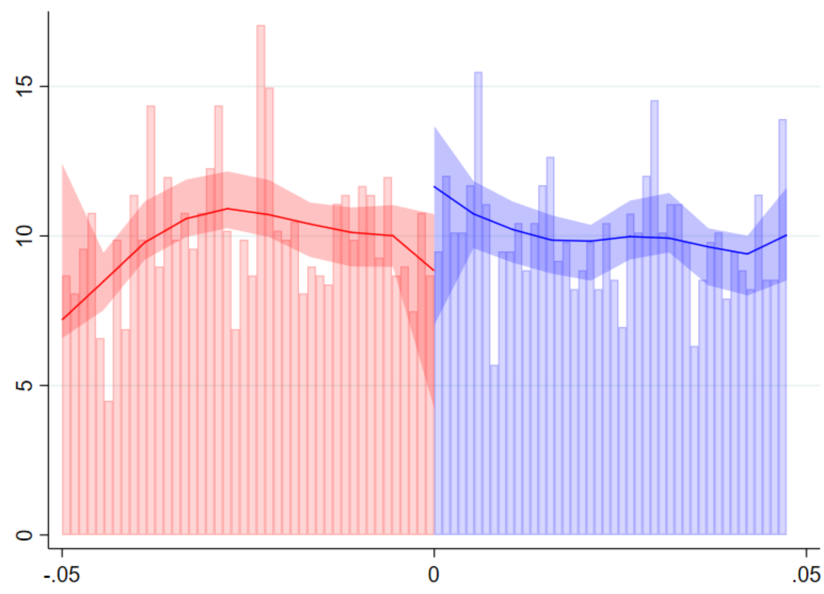
Table 5: Covariate Balancing:

	(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.

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

Figure 8: Manipulation Testing Plot



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

Appendix Materials for:

"Quality Disclosure, Demand, and Congestion"

A Rationing Demand by Wait List: A Theoretical Model

In this Appendix 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.[13] 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.

A.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

[13]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.

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).[14] 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

[14]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.

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 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.

A.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.

A.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 $Q_h = \sum_{k=0}^{\infty}(j_{n-k} - s_{n-k})$.[15] 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:

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

A.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

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

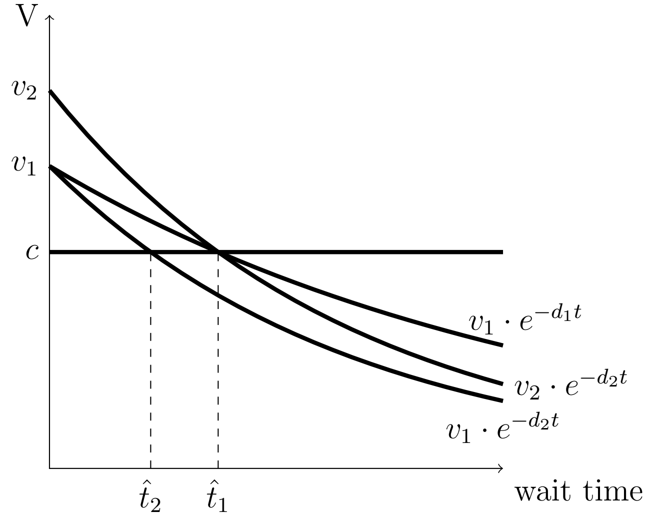
with lower v and higher d.

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 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.[16] 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.

[16]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.

Figure A2: Relationship Between Benefits and Costs of Waiting



Further Analyses

A.5 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). Results available upon request. 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 A9). 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.

A.6 Total Visit by Rounding Threshold

In Appendix Table A13 I show my baseline regression specification, but with total visits and existing visits as well as new visits as the left-hand side variable. I find no effect for total visits or existing visits, as expected. Because medical care is an experience good, a star rating is unlikely to be as informative if a patient has already seen a doctor many times as compared to a first-time visitor. I find no discernable jump in existing patient visits nor in total visits, which are mostly non-new visits.

A.7 Capacity Constraints

I explore whether the congestion effect is concentrated among a small number of doctors. Specifically, I re-run the regression of star rating on wait time (congestion), but leaving out one doctor at a time. The intuition behind this permutation test is to identify if certain doctors who might be capacity constrained are driving the results. I retain the regression coefficients from the "leave one doctor out" regressions and identify the doctors who have particularly high leverage on the regression and when left out lead to low coefficient values, yielding a sample of 7% of family medicine doctors. If this subset of doctors faced a true capacity constraint, and the others did not, one would expect that the more capacity constrained doctors would have longer wait times all else equal and regardless of whether their ratings are rounded up or down. This is precisely what I find when comparing the doctors and their visits to the other doctors. They have longer baseline wait times: for new visits, they have 4.8 days longer wait time (95% C.I. 4.16 to 5.52) and for established visits, they have 3.8 days (95% C.I. 3.48 to 4.04) longer wait times. This suggests that most doctors, the ones who do not drive the congestion results, generally have excess capacity because their wait times are low to begin with, but those who have less capacity, because wait times were already high to start out with, are also the ones most impacted when star ratings are introduced in the post-rating period. This evidence is consistent with wait times as a rationing mechanism and with capacity constraints binding on some doctors, but not others.

A.8 Example of Survey Questions

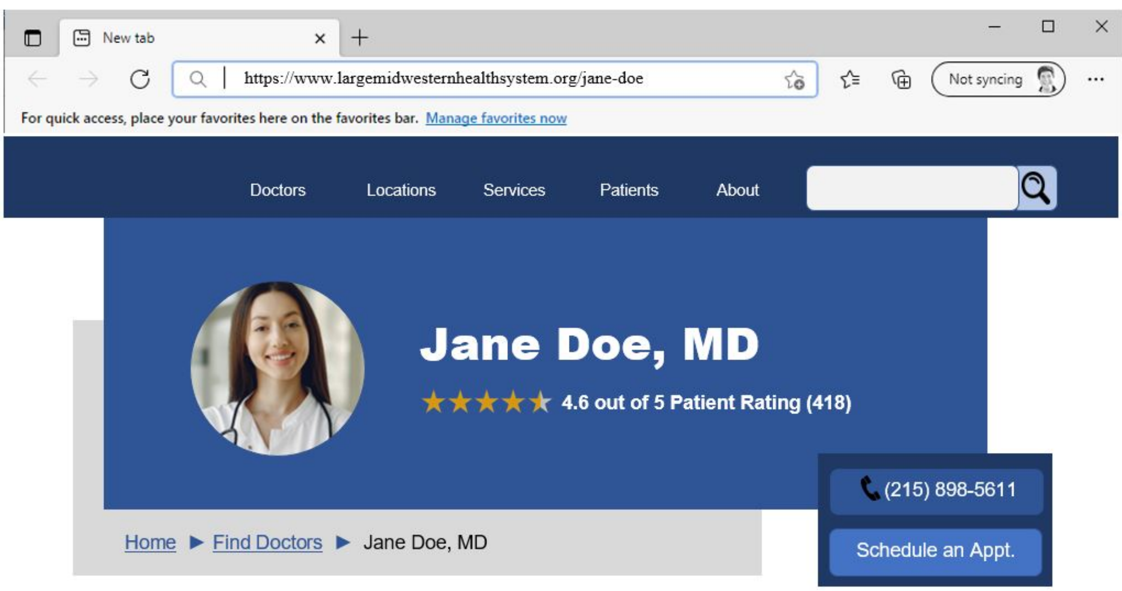
Survey Questions:

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?

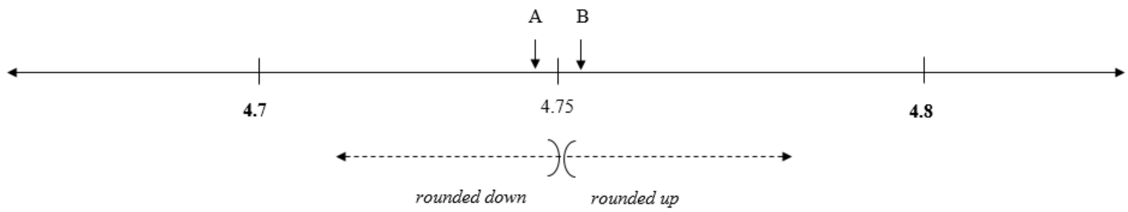
B Appendix Tables & Figures

Figure A1: Sample Physician Rating Webpage



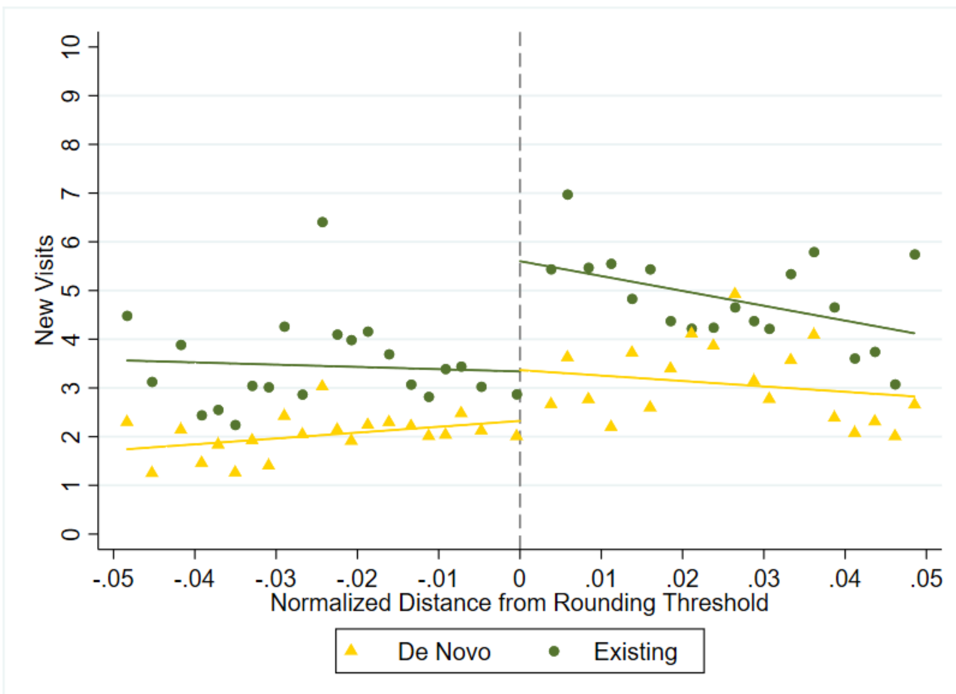
Note that before 11/2018, webpage looked exactly the same except without the star ratings.

Figure A2: 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 A3: 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.

Table A1: 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 A2: Difference-in-Differences Estimates

	(1)
Post X Above	2.786***
	(0.919)
Post	0.111
	(0.479)
Above	-0.405
	(0.723)
R-squared	.01933088
Observations	3562

Note: Standard Errors clustered at the provider level and observations weighted by review count.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: 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 A4: 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 A5: DeNovo = Yes, New Visits - Family Medicine

	(1)	(2)	(3)	(4)	(5)	(6)
	Linear	Quadratic	Cubic	Linear	Quadratic	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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6: 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 A7: Monthly New Visits - By Provider Credentials

	(1) MDs	(2) 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 A8: Charlson Score t-test Results

	MD vs. not MD
Difference	0.137***
	(0.001)
N	7024166

Note: Std. Err in parentheses.
Diff = MD minus non MD

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A9: 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 A10: 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.

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

Figure A4: Covariate Balance on Baseline Regression (Provider–Month Panel)

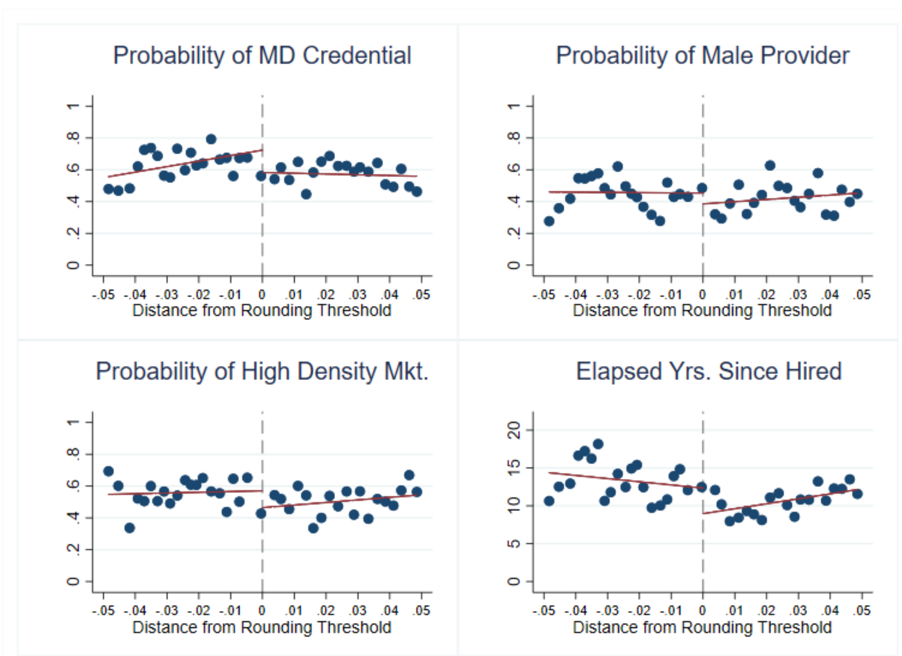
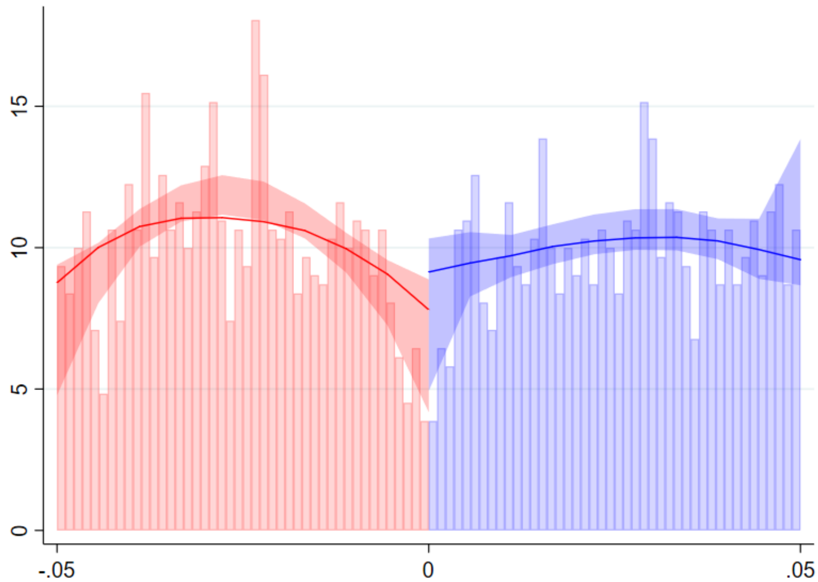
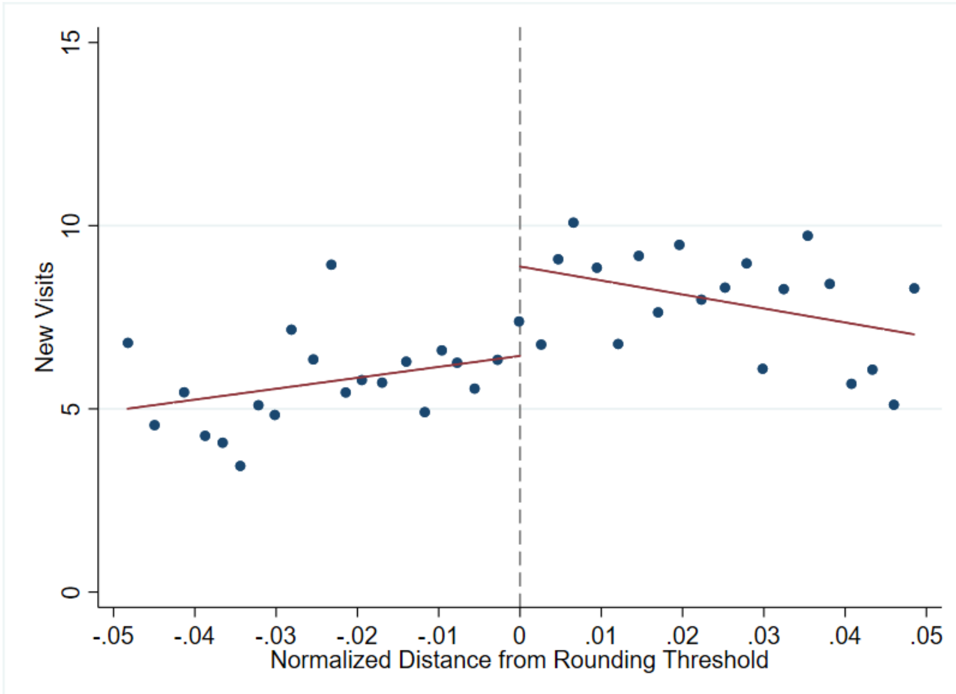


Figure A5: Manipulation Testing Plot



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

Figure A6: 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 A7: Comparison of Using the Full Support of the Data vs. the Restricted Support Used to Plot Figure 6.

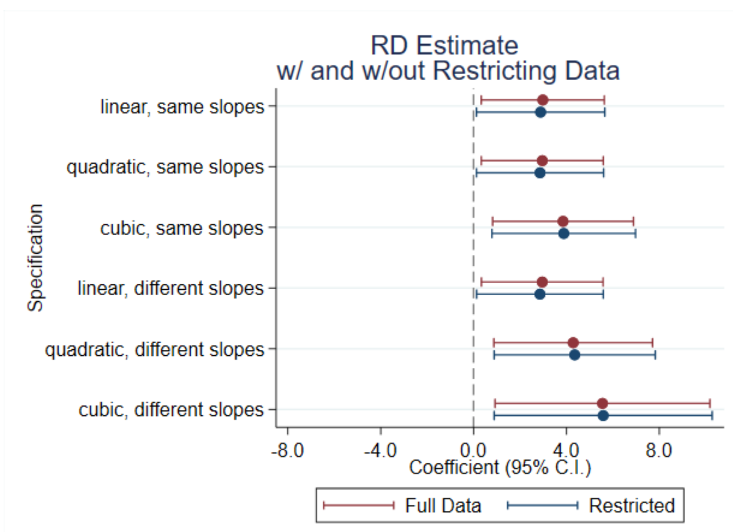


Table A11: Monthly New Visits - Family Medicine

	(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	No	No	No	No	No	No
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.

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

Table A12: t-test on Raw Rating

	Full Panel vs. Accepting New Patients
Difference	0.0434**
	(0.0169)
N	3282

Standard errors in parentheses

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

Table A13: Monthly New Visits - Family Medicine

	Post-Disclosure		
	(1)	(2)	(3)
	Total Visits	Existing Visits	New Visits
Rounded Up	-20.01	-22.96	2.956**
	(23.44)	(22.96)	(1.332)
Functional Form:	Linear	Linear	Linear
Treatment Interaction	Yes	Yes	Yes
Cutoff FEs	Yes	Yes	Yes
Mean Below Threshold	188.913	183.438	5.475
% Change	-10.6	-12.5	54.0
Observations	2730	2730	2730

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

Figure A8: Relationship Between Star Ratings and Medical Metrics, adjusted for Doctor-level Case Mix

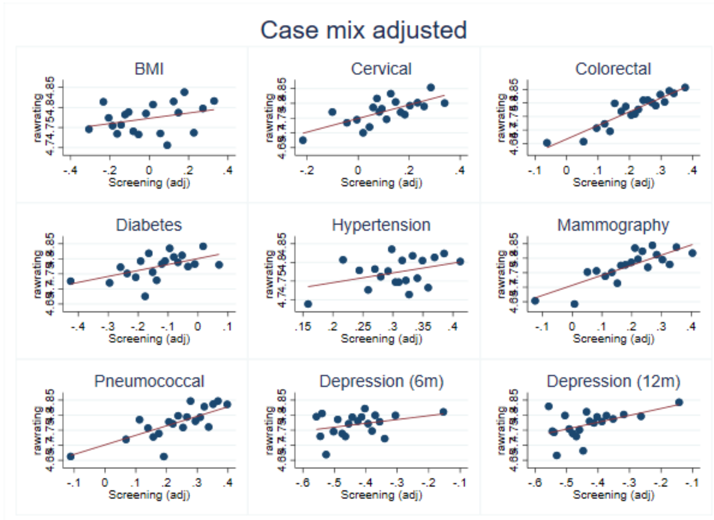


Table A14: Monthly New Visits - Family Medicine

	(1)
	Linear RD
Rounded Up	3.306**
	(1.404)
Distance to threshold	7.368
	(24.73)
Dist × Rounded	-49.00
	(49.61)
Functional Form:	Linear RD
Treatment Interaction	Yes
Cutoff FEs	No
Mean Below Threshold	5.475
% Change	60.4
Observations	2730

Note: Standard Errors clustered at the provider level and observations weighted by review count
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$