Differential Impacts of Online Ratings in the Market for Medical Services*

Aaron P. Kaye† Michael Luca‡ Sonal Vats§

May 18, 2024

Updated frequently, please click here for the latest version

Abstract

This project focuses on ZocDoc.com – a unique website that integrates physician profiles, patient reviews, and appointment scheduling onto a single platform. We collected data from the website every day for over a year to construct a novel dataset consisting of profiles, reviews, and ratings for primary care physicians in eight metropolitan divisions. We infer bookings from daily records of appointment availability. ZocDoc displays ratings on a scale of one to five stars, with overall average ratings rounded to the nearest half-star. We use a regression discontinuity design to identify the causal impact of reviews on patients' choice of physician. Our preliminary results suggest that patients care quite a bit about quality. However, due to physicians' capacity constraints and the level of demand, 4, 4.5, and 5-star doctors find most of their offered appointments are booked. The main distinction is timing, with lower rating physicians' appointments booked once the appointments with higher-rated physicians become scarce. We find approximately a doubling in patient volume across the cutoff from 4.5 to 5 stars. We conclude by evaluating the differential impact of ratings, finding that the effects are higher for women physicians and physicians with more reviews. We find a small but insignificant difference for hospital affiliate physicians.

JEL: D12, I11, L15
Keywords: online reputation, healthcare, platform design, two-sided markets, e-commerce, consumer search

*We are grateful to Zach Y. Brown, Francesco Decarolis, Randall P. Ellis, Ching-to Albert Ma, Sarah Miller, Markus Mobius, and Johannes Schmieder. We want to thank Marlies Bar and Jessica Van for helpful feedback as discussants. We also thank conference participants at ASHEcon, iHEA, IRDES Workshop on Applied Health Economics and Policy Evaluation, and seminar participants at Boston University and University of Michigan for helpful comments and suggestions. This paper replaces our previous paper, "Digitizing Doctor Demand: The Impact of online Reviews on Doctor Choice."

†Aaron P. Kaye: University of Michigan (corresponding author), Email:apkaye@umich.edu
‡Michael Luca: Harvard Business School
§Sonal Vats: Sacred Heart University

1 Introduction

Credence goods are a type of good with qualities that are difficult or impossible to fully judge by a consumer even after purchase and consumption (Darby and Karni, 1973). Medical services are a prominent example of credence goods. In the market for physicians services, consumers face both ex-ante and ex-post uncertainty regarding the quality of care.

In the market for medical services, consumers face a challenge in both selecting a physician and also judging the quality of the services received from that physician. In the past, consumers have relied heavily on social learning to resolve these asymmetries. For example, consumers may ask their peers to recommend a physician. In fact, the National Institute for Aging tells patients to "ask people you trust" for physician recommendations.

1.1 The Role of Online Ratings

Online ratings are an increasingly important driver of economic activity and consumer decision-making. The three industries where online ratings are most viewed are restaurants, hotels, and healthcare.[1] Through websites like ZocDoc, a unique website that integrates physician profiles, patient reviews, and appointment scheduling, online reputation, is a potentially important source of information for consumers about physicians. The large-scale distribution of information from numerous other consumers, could help resolve information asymmetries among a much broader peer group than was previously been possible.

On the provider side, these subjective reviews could be a source of feedback to physicians they would not receive otherwise. In health care markets, there is strong evidence that public disclosure of quality data has been effective in better matching patients with products and providers. Studies find that consumers tend to prefer higher-quality providers. Dafny and Dranove (2008) and Jin and Sorensen (2006) find that the publication of report cards boosted the market share of insurance plans that received higher scores. Another example is Bundorf et al. (2009), who find that fertility clinics with high birth rates gained market share after the US Center for Disease Control and Prevention began publishing success rates in 1997. Using a fixed effects framework, Wang et al. (2011) find that surgeons who received poor ratings on Pennsylvania CABG report cards experience a decrease in patient volume. Similarly, Cutler et al. (2004), find that lower ranking hospitals in New York lost market share, especially among less severely ill patients. However, disclosure can also harm

[1] https://www.brightlocal.com/research/local-consumer-review-survey-2020/

consumers if sellers can superficially boost their performance through window dressing. For example, Dranove et al. (2003) find that cardiac surgery report cards in New York and Pennsylvania led to selection by providers, which suggests a serious downside risk on quality reporting in health care. Werner et al. (2009) and Feng Lu (2012) find similar evidence with the Nursing Home Quality Initiative.

It is unclear whether consumer review websites should significantly affect markets for credence goods. Consumer review websites help fill the void left by the absence of any government or nonprofit agency assuming the role of information provider on primary care physician quality. Consumer reviews can also be a complement or substitute for existing information—education, board certification, and malpractice claims—on physicians, some of which may not be easily available or understood by a lay person. Alternatively, a consumer writing a review cannot fully evaluate the treatment or service received, since they are unfamiliar with the intricacies of the medical knowledge possessed by the primary care physician. Further, patient-created reviews can be difficult to interpret—they reflect the views of a non-representative sample of patients and are subjective.

Notwithstanding these challenges, an emerging body of economics literature studies the impact of online ratings on demand for medical services. McCarthy et al. (2022) combine Yelp reviews with claims data to show that patients are willing to travel further to receive care from hospitals with higher Yelp ratings. Brown et al. (2023) study demand for General Practitioner (GP) offices in England and show that patient demand from low-income neighborhoods responds sharply to summary star ratings. Collectively, these studies underscore that online ratings could be a driving force in healthcare decision-making.

There is also reason to suspect that the returns to online reputation are different for different physicians. In a field study, Chan (2023), finds that signals of doctor quality reduce 90% of the racial gaps in willingness to pay for doctors. Brown et al. (2023) find that the impact of ratings could be mediated by private information. Other work shows evidence of the differential impact from other forms of reputation, for example Sarsons (Sarsons) evaluates how patient deaths impact referrals to surgeons and finds that female surgeons experience a larger drop in referrals after a patient death.

Zocdoc presents an ideal context to study the impact of online reviews as it has the following notable features: 1) A discontinuous rating system; 2) Consumers face little variation in prices across providers in their insurance network; And 3) A closed-loop review system. Zocdoc displays ratings on a scale of one to five stars, with overall average ratings, rounded to the nearest half-star. We take advantage of the fact that Zocdoc rounds ratings to the nearest "half-star." As a result, two

physicians with nearly identical ratings can straddle the cutoff to display 4.5 versus 5-stars. These may be viewed as very different by consumers, even if the underlying quality is quite similar. We use this natural experiment to estimate the causal impact of ratings on booking volumes using a regression discontinuity design. Further, we explore the differential impacts of rating by repeating the analysis for economically interesting subgroups of physicians.

2 Background on ZocDoc.com

Launched in 2007, Zocdoc is an online medical care search, and scheduling service. Freely available to patients, the website enables patients to search for physicians by insurance, location, specialty, procedure, hospital affiliation, gender, and languages spoken. Based on the selection criteria ZocDoc provides patients with a list of physicians, patients can view open slots in physicians' schedules and make an appointment online. According to ZocDoc, most of the appointments happen in 24-72-hour window.[2] Zocdoc appointment service was initially limited to dentists in Manhattan, as of 2013, Zocdoc claims to serves 40 percent of the U.S. population across more than 1,800 cities. More than 2.5 million patients use Zocdoc to find doctors every month.[3]

A patient looking for physicians on Zocdoc can use the website's search feature, in Figure 2.1, to search for physicians based on specialty, location, and insurance. For example, Figure 2.2 shows a snapshot of the list of physicians in Boston, with no restriction on the type of insurance they accept. This search results page presents the patient with an ordered list of physicians, displaying displaying the physician's photograph, practice address, rounded average rating, main specialty, medical degree, hospital affiliation, and all open slots in their appointment schedule for the current week. Clicking on any physician name takes the patient to the physician's profile page which displays additional information about the physician's education, specialty, languages spoken, types of insurance accepted, and also displays the detailed ratings and text reviews left by any patient. Figure 2.3 display the physician profile and the individual patient ratings as they appear on Zocdoc's physician profile page.

[2] 12 Facts about Zocdoc Users
[3] Zocdoc Announces Patients Have Booked More Than 1,000 Different Procedures Through Its Free Service

Figure 2.1: Search Page

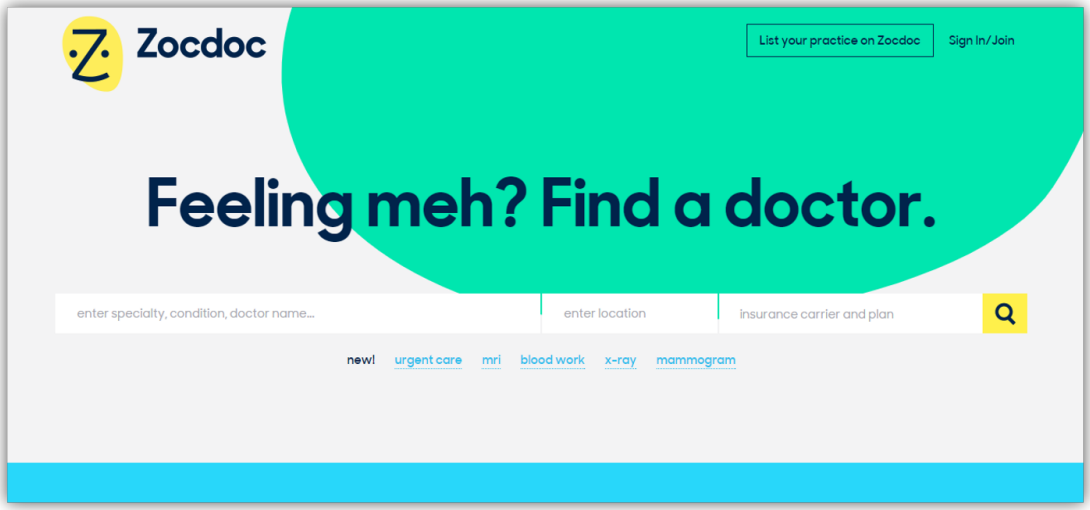


Figure 2.2: Example Zocdoc Search Results

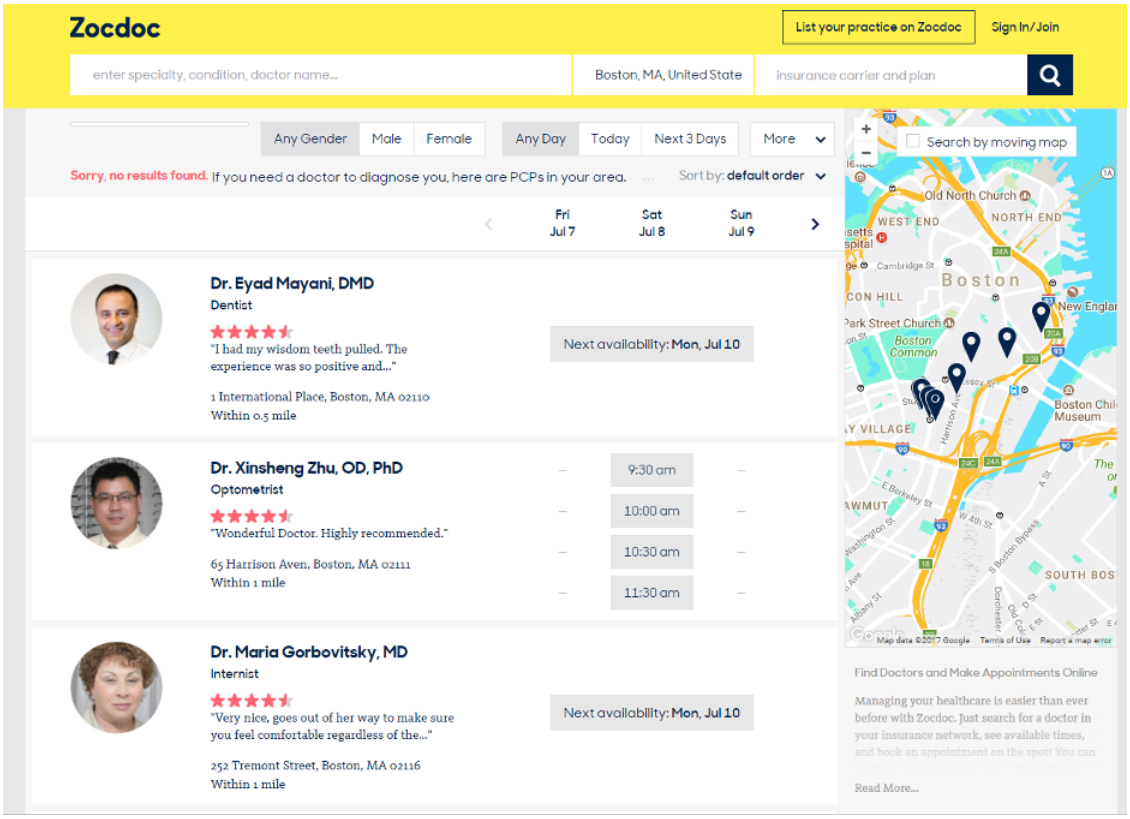
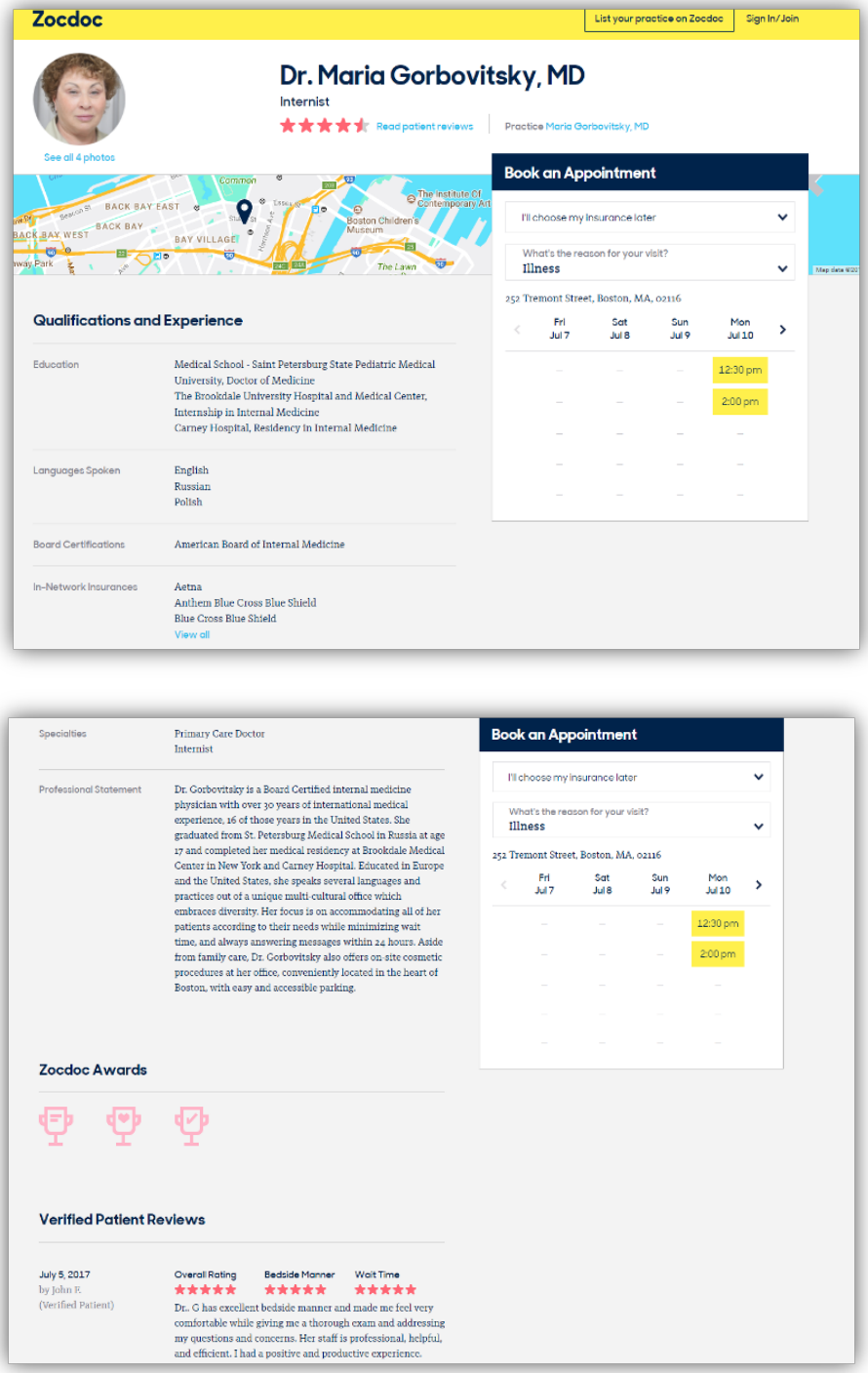
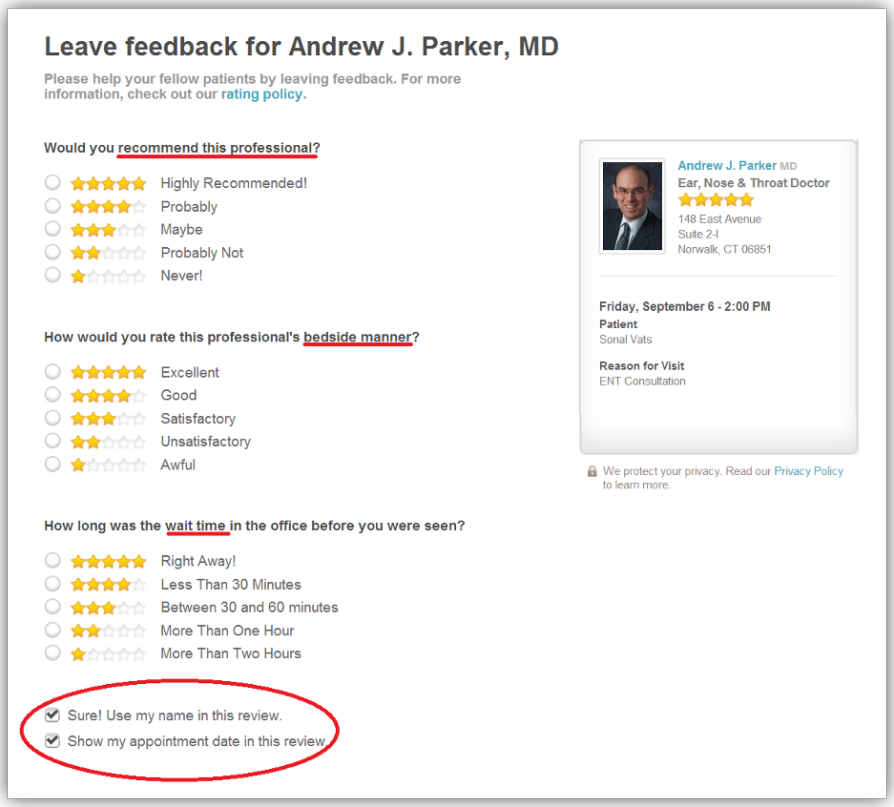


Figure 2.3: Example Profile Page



After an appointment, Zocdoc emails a thank-you note, encouraging patients to review and rate (from 1-5 stars) their physicians for bedside manner, wait time and overall impression, as illustrated in Figure 2.4. The patient can then rate the physician they visited and can also enter a text review. Once a review is written, anyone (with or without an account) can access the website and read the review. Patients will come across reviews within the context of the search for a physician. This allows the patients, looking for physicians on Zocdoc, to compare and assess them on common quality characteristics. Since each verified patient is encouraged to leave a review, it may not be that patients who have had extreme experiences, and who are proactive, are the only ones to leave reviews.

Figure 2.4: Post Appointment Feedback Prompt



During the time of our study, Zocdoc was free to patients and followed a subscription model for providers. Physicians can choose to subscribe by paying a monthly subscription of $250. For each subscribing physician the website has a profile page with 'verified' credentials, and patient–submitted reviews. Subscribing physicians could benefit by attracting new patients, and by filling the last-minute cancellations and postponements (10-20 percent of total appointments) by Zocdoc patients. However more recently, Zocdoc proposed pricing changes have been a source of concern for physicians.[4]

3 Data Construction

We construct a novel dataset using Zocdoc. These data include physician professional information, patient generated reviews, and appointment schedules. We focus on primary care physicians in eight metropolitan divisions where Zocdoc has a significant presence, with New York City being by far the largest. Our data includes over 5.8 million offered appointments from over two thousand primary care physicians, with 94% of profiles having at least eight reviews. Table 3 has the number of appointments and physicians per metropolitan division:

Table 3.1: Offered Appointments and PCPs by Metro Division

Metro Division	**Appts.**	**PCPs**
Boston, MA	86,512	117
Cambridge-Newton-Framingham, MA	71,159	68
Chicago-Naperville-Arlington Heights, IL	795,000	331
Fort Lauderdale-Pompano Beach-Deerfield, FL	184,185	80
New York-Jersey City-White Plains, NY-NJ	3,629,392	1,291
San Francisco-Redwood City-South San Francisco, CA	69,384	31
Silver Spring-Frederick-Rockville, MD	232,236	82
Washington-Arlington-Alexandria, DC-VA-MD	774,756	305

The data collection process works in three steps. In the first step, it uses Zocdoc's search engine, to compile a list of Primary Care Physicians and query specific search ranks. The second step collects profile information for each physician in the list. The third step collects all of each physician's reviews and available appointments for the next 35 days. The first two steps were repeated monthly. The third step was repeated daily at 2am. This exercise was repeated for a period of over one year, starting February 22, 2016 and running through April 17, 2017, and ended after updates to Zocdocs's website starting in March 2017 make our data less reliable.

[4] CNBC: Zocdoc Price Surge has Doctors Fretting

For the first step, each search would yield a maximum of 50 results, so it was necessary to repeat each search with narrow filters in order to construct a comprehensive list of physicians in the area. In this step the search was repeated for each zip code, physician type, appointment type, and physician gender. Along with a list of physicians, we also collected a query specific search rank. The query specific search rank tells us the page rank of each physician for a given search. The first step is repeated once a month to search for new physicians.

For steps 2 and 3 we collect the following information each month:

Each month we also collect physician profile pictures. We then use Microsoft's Face API to collect additional demographic information.

3.1 Platform Updates

Our collected data spans from 2016 to 2017, a timeframe that proves particularly fitting for our study. The platform has since introduced numerous updates, adding complexity and making a similar study considerably more challenging for researchers today. Notably, it now displays an average rating rounded to the nearest hundredth, a departure from its previous practice of rounding off the overall rating to the nearest half-star, the platform design feature on which our natural experiment depends. The platform's redesigned landing page and profile pages provide users with an array of additional information and navigation options, which, while enhancing the user experience, obscures the simplicity of user interactions that our study aims to examine.

4 Empirical Strategy

Zocdoc employs a system that prominently displays on the landing page a physician's average overall rating rounded to the nearest half-star, with detailed rating information relegated to the profile pages. For example, as illustrated in Figure 4.1 a physician with a 4.74 average rating will be rounded down to 4.5 stars, while a physician with 4.75 stars will be rounded up to 5 stars. This allows us to examine observations with nearly identical underlying average ratings but a half star difference in the rating displayed to consumers. We leverage this rating system to analyze the impact of ratings on patient volumes using a regression discontinuity design. We supplement this with comprehensive profile data to evaluate the differential impacts of ratings by gender, number of ratings, and hospital affiliation.

Table 3.2: Data Collection of Physician Profiles, Ratings, Reviews, Practice Information, and Availability

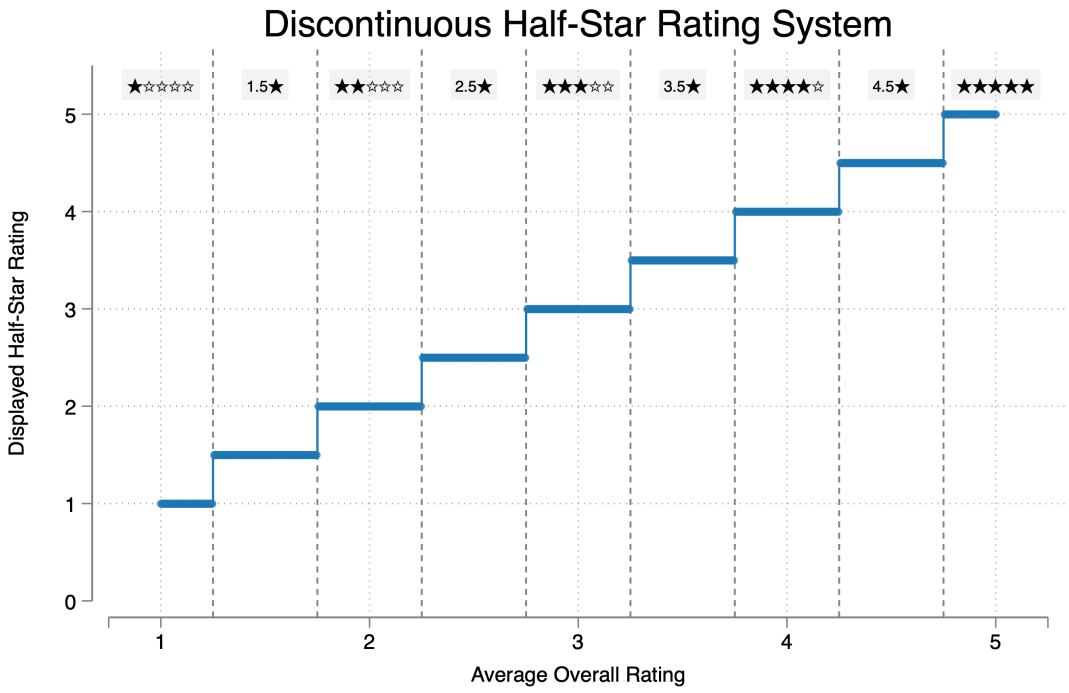
Category	Variable	Visibility
Physician Profile Information (Collected Monthly)		
	Profile URL	html
	Doctor Name	Landing Page
	Title	Landing Page
	Specialty: Primary Care, Internist	Landing Page
	Badge - Rapid registration	Profile Page
	Badge - See You Again	Profile Page
	Badge - Speedy Response	Profile Page
	Badge - Scheduling Hero	Profile Page
	Practice Name	Profile Page
	Specialties	Profile Page
	Education: Medical School and Residency	Profile Page
	Hospital Affiliations	Profile Page
	Languages	Profile Page
	Board Certifications	Profile Page
	Awards and Publications	Profile Page
	In Network Insurances	Profile Page
	Doctor Code	html
	Professional Statement	Profile Page
Ratings and Reviews (Collected Daily)		
	Rounded Rating Overall Rating (Half-Star)	Landing Page
	Review level - Patient Name	Profile Page
	Review level - Date	Profile Page
	Review level - Overall Rating (1-5)	Profile Page
	Review level - Wait Time Rating (1-5)	Profile Page
	Review level - Bedside Manner Rating (1-5)	Profile Page
	Review Text	Profile Page
Practice Information (Collected Daily)		
	Address	Profile Page
	Coordinates: Latitude and Longitude	html
	Location ID	html
Schedule and Appointments (Collected Daily)		
	Available Illness Appointments	Profile Page*
	Available New Patient Appointments (next 35 days)	Profile Page*
	Appointments Start Times	Profile Page*
	Appointment Locations	Profile Page*
Search Information (Recorded Monthly)		
	Gender	Search Filter
	Zip Code	Search Filter
	Page Rank	Landing Page

* indicates 3-5 days of availability on the landing page, additional availability on profile and scheduling pages

4.1 Discontinuous Rating System

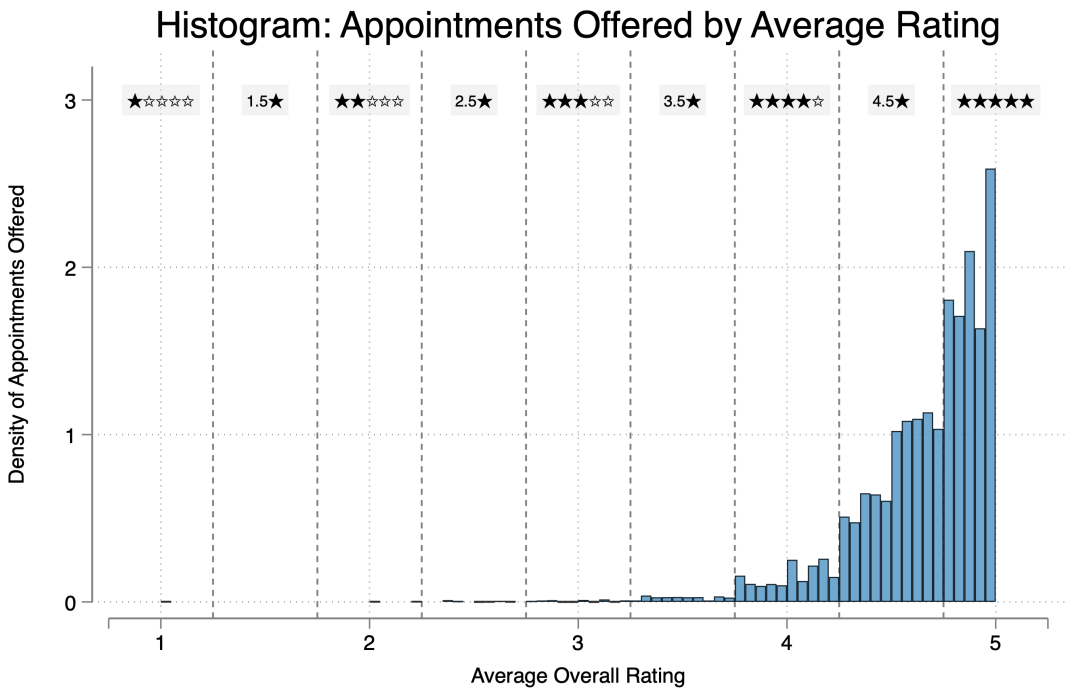
We specifically use a regression discontinuity design with multiple cumulative cutoffs, this exploits the fact that the rating system displays the rounded half-star rating on the landing page. We use the multiple cumulative cutoff approach since it has the interpretation that there is a different cutoff for each half star and that the treatment is different at each cutoff. For example, at the 4.25 cutoff, the control is 4-stars and the treatment is 4.5-stars. At the 4.75 cutoff, the control is 4.5-stars and the treatment is 5-stars.

Figure 4.1: Rounding of Overall Ratings



For our preliminary analysis, we mainly focus on the 4.75 cutoff since that has the most observations in our data. Figure 4.2 plots the distribution of ratings by appointments offered in our primary sample. The data are left skewed with most appointments belonging to physicians with at least a 4.5 star rating.

Figure 4.2: Appointments by Average Rating



The cutoffs faced by a physician on a given day are a deterministic function of the physician's average overall rating. The treatments at each cutoff are different in some respects. We formalize the notation for the average rating, X_j, the displayed star rating, s_j, and the relevant cutoff values, c_i.

- average rating: $X_i \in [1, 5]$
- displayed star rating: $s_i \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$
- cutoff values: $c_i \in \{1.25, 1.75, 2.25, 2.75, 3.25, 3.75, 4.25, 4.75\}$

Next, we can write τ_i, the local causal treatment effect of an increase in a half-star. With the treatment is local to the average rating X_i exactly equal to cutoff c_i:

$$\tau_i = E[Y_j(s_i) - Y_j(s_{i-0.5})|X_j = c_i] = \lim_{x \downarrow c_i} E[Y|X_j = x] - \lim_{x \uparrow c_i} E[Y|X_j = x]$$

Cattaneo et al. (2020) point out that, in situations with multiple cumulative cutoffs, we can use each observation to estimate two different treatment effects. For example, a physician with an average rating of $X_i = 4.2$ could be used to estimate treatment effects for $\tau_{i=4}$ and $\tau_{i=4.5}$. In this case, the physician with an average rating of $X_i = 4.2$ would be considered treated when estimating $\tau_{i=4}$ and control when estimating $\tau_{i=4.5}$.

We estimate each τ_i cutoff-by-cutoff with local polynomial methods, and asymmetric data-driven bandwidth selection using the `rdrobust` package. For our primary specification, we use a first-order local polynomial with a triangular kernel function to construct the point estimator τ_i. We use a second-degree local polynomial for bias correction. To account for the panel structure of our data, we estimate cluster robust nearest neighbor variances.

4.2 Primary Outcomes

Our primary outcome of interest is weekly patient volumes. We identify offered appointments as booked if they disappear from schedules up to three days before the date of the appointment. As a pilot we chose three days to avoid measurement error as many physicians choose not to list appointment same day or two days before the appointment date. These estimates should inform both: 1) how consumers use ratings to select primary care physicians, and 2) the gains to physician for having a better online reputation. For this draft we use the inverse hyperbolic sign (IHS) of the count of weekly bookings, denoted in equation 4.1.

$$\sinh^{-1}(y) = \ln(y + \sqrt{y^2 + 1}) \tag{4.1}$$

The IHS transformation has desirable properties similar to natural log, but its domain includes zero, which account for approximately 16% of our observations. For this draft we present the raw parameter estimates and loosely interpret them as percent difference. Like the non-linear transformation natural log, parameters alone do not have a clear interpretation. Norton (2022) provides details on the methods to transform parameter estimates into marginal effects.

4.3 Covariates

For the primary specifications, we include market-week fixed effects. The market-week fixed effects account for the fact that booking patterns could be different in different MSA's and that our data cover a period of rapid growth of the platform. Additionally, we include controls for number of reviews at the start of the week, appointment type (illness, new patient visit, or cross-listed), number of locations, appointment length, appointments offered, and hospital affiliation.

4.4 Sample Selection

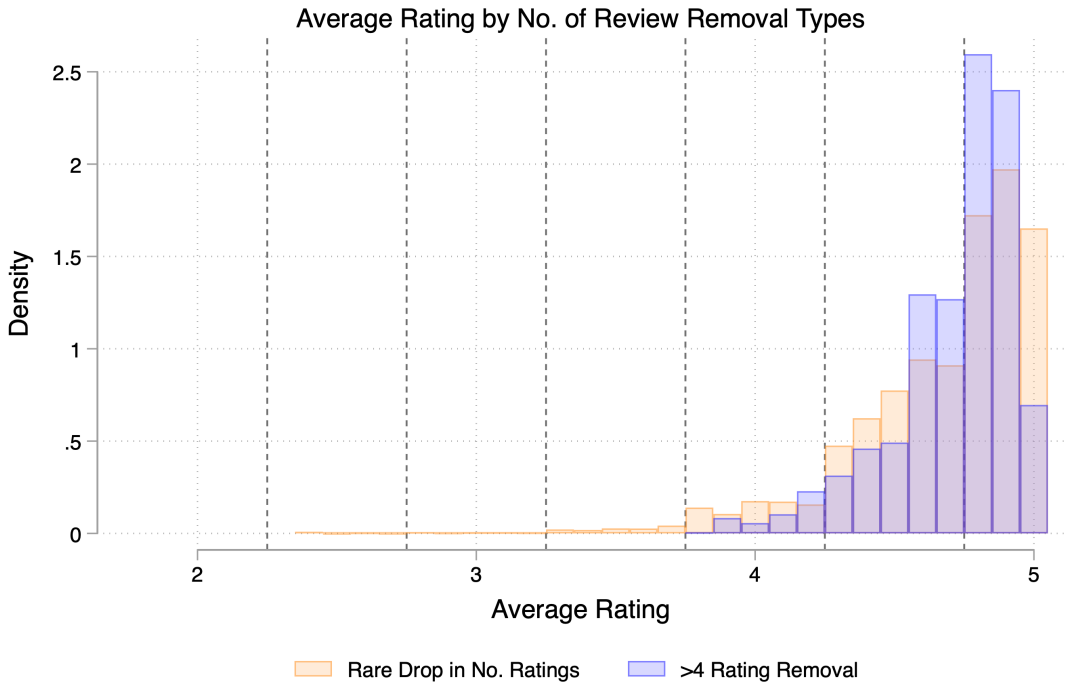
For our primary specification, the observations are on the physician week level. We limit our sample to appointments offered on weekdays between 8am and 6pm local time. We keep physician weeks where at least one appointment is available three weeks in advance. Our identification depends on cross-sectional variation in ratings near the cutoff values. Including observations with a large change in rating or were rating frequently change could be problematic for a number of reasons. For example, this type of within individual variation, a change in rating, depends on receiving new ratings, which requires new bookings, which is our outcome measure. Additionally, it is possible that the true underlying treatment effect of ratings accumulates over time. To account for this issue, we limit our observations to physician-weeks where the physician is at a "stable" half-star rating, meaning this is their half-star rating for at least 90% of included weeks.[5] We further limit our specification to exclude physician-weeks where the physician has fewer than eight ratings to avoid numerical issues in the running variable.

We make one other notable sample restrictions. We remove observations from physicians where there is evidence on possible deleted reviews or possible data collection issues related to number

[5]Rating changes are rare, in our sample ∼\$82% of observations are at a "stable" rating, and ∼\$77% of observations are associated with physicians that who never change half-star rating during our period.

of ratings. If reviews are never deleted, then for a given physician the number of reviews should be monotonically increasing. We remove observations for physicians with more than four weeks where the number of ratings decreased from the previous week. Figure 4.3 shows the distribution of physicians by average rating and review removal type. The figure below documents the distribution of observations flagged for deleted reviews.[6]

Figure 4.3: Appointments by Average Rating and Review Removal Type



4.5 Differential Impacts

In our analysis, the regression discontinuity estimate τ_j is a weighted average treatment effect of potentially heterogeneous treatment effects. In our case, this heterogeneity is not only expected but is of primary interest. We investigate the differential impact of star rating by partitioning our data into economically interesting subgroups and estimating treatment effects separately by cutoff and subgroup. Our current analysis focuses on physician gender, number of ratings, and hospital affiliation.

[6] See **Appendix Deleted Reviews** for supplementary analysis on deleted reviews.

5 Results

In this section we first present data visualizations relating the appointment level likelihood of booking and patient volumes. Then we present estimates for patient volumes and remaining vacant appointments. We end with the differential impact of ratings. For this version we primarily focus on the 5-star cutoff.

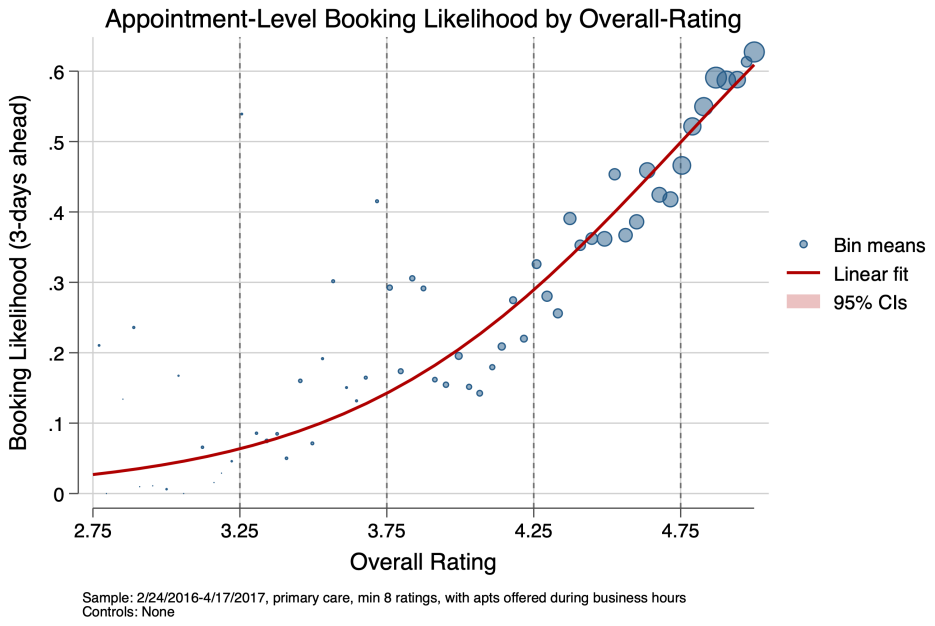
5.1 Naïve Approach Visualizations

These results show how likely a given appointment is to be booked by average rating, and then examine the CDFs of booked appointments over the number of days in advance that they are booked. Then we zoom into the 4–5-star physicians and present patient volumes by overall rating and star-rating.

5.1.1 Appointment Level Booking Likelihood

On the appointment level, as illustrated in Figure 5.1 we see a clear correlation between booking likelihood and average rating both across and within half-star cutoffs.

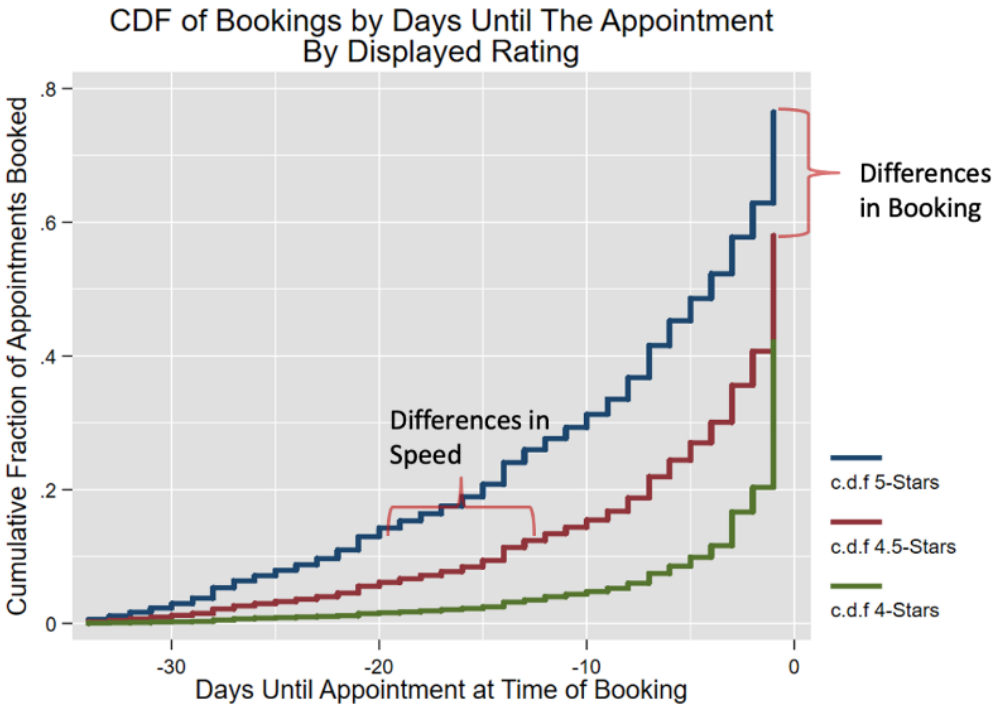
Figure 5.1: Booking Likelihood by Overall Rating



5.1.2 Cumulative Booking Likelihoods by Days Before the Appointment Date

Here we explore, by overall rating, when appointments are booked and how likely appointments are to be booked. Figure 5.2 presents a the fraction of appointments that are booked by time of booking. The horizontal axis is day of booking minus the day of the appointment, telling us the number of days before an appointment. The furthest in advance that an appointment could be offered is 35 days in advance. The CDF runs from 35 days before an appointment (left) to the day before an appointment (right).

Figure 5.2: Comparison of Booking likelihood by Days in Advance of Appointment

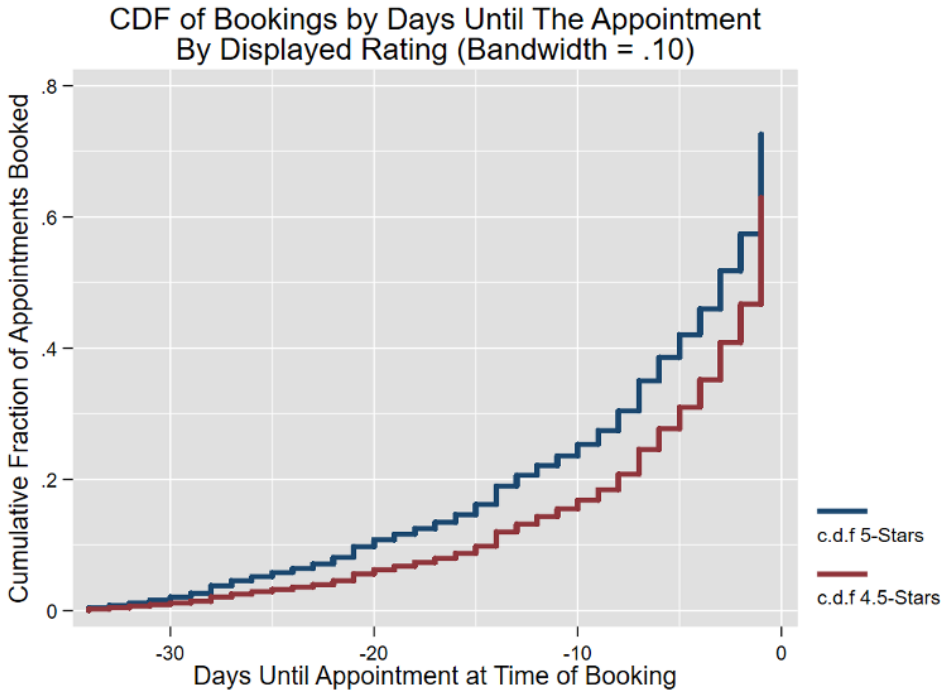


Two useful ways to compare the CDFs are differences vertically and horizontally. The vertical difference tells us the difference in percent of appointments booked at a given number of days in advance. The horizontal differences can be interpreted as the difference in how many days in advance the same percent of appointments were booked. The CDF shows notable results. First, we see that higher rating doctors fill a higher percent of their appointments and do so further in advance. Second, we see that a high percent of appointments are booked the days immediately before the appointment date. About one third of 4.5-star bookings are done the day before, and over half the 4-star bookings happen the day prior. As mentioned above, this could be partially attributed to

physicians removing vacancies from their schedules just before the appointment date.

Next, in Figure 5.3, we use a pilot bandwidth of .1 to compare these CDFs of physicians just above and just below the 4.75 threshold to have five stars. Here we see that the CDFs are closer together, but there are still differences in the percent of appointments booked and how far in advanced the appointments are booked. Here we see that the CDFs are closer together, but there are still differences in the percent of appointments booked and how far in advanced the appointments are booked.

Figure 5.3: Comparison of Booking likelihood by Days in Advance of Appointment near 5-Star Cutoff

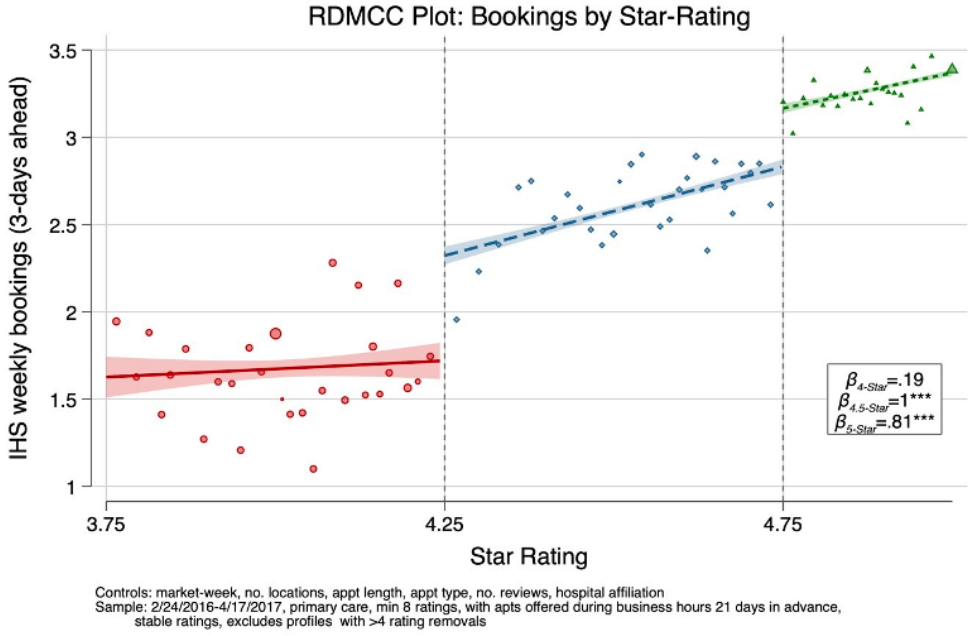


We next aim to quantify how much of these differences can causally be attributed to having 4.5 stars displayed versus having 5-stars displayed. However, one challenge highlighted by these figures is that the difference in booking likelihoods differs by time in advance of appointments. One possibility is that physicians have fixed capacity, so lower-rated physicians could receive more bookings closer to the time of appointments once appointments with 5-star doctors become scarce. One might underestimate consumer responsiveness to ratings if only looking at bookings just before an appointment. To address this concern, we repeat our analysis for different times leading up to appointments.

5.1.3 Booking Volume

While the previous figures focus on appointment level differences booking rates, we next consider booking volumes. Figure 5.4 plots weekly booking volumes by overall rating and displayed half-star. The horizontal axis is the average rating, and the vertical is the IHS of bookings aggregated to the week level. The jump at each cutoff presents suggests a jump in patient volume at the 4.5 and 5-star cutoffs. This figure is only suggestive, as it still takes a naïve, linear regressions approach instead of the data-driven regression discontinuity methods we present next.

Figure 5.4: Regression Discontinuity with Multiple Cumulative Cutoffs for Booking Volumes (4-5 Stars)



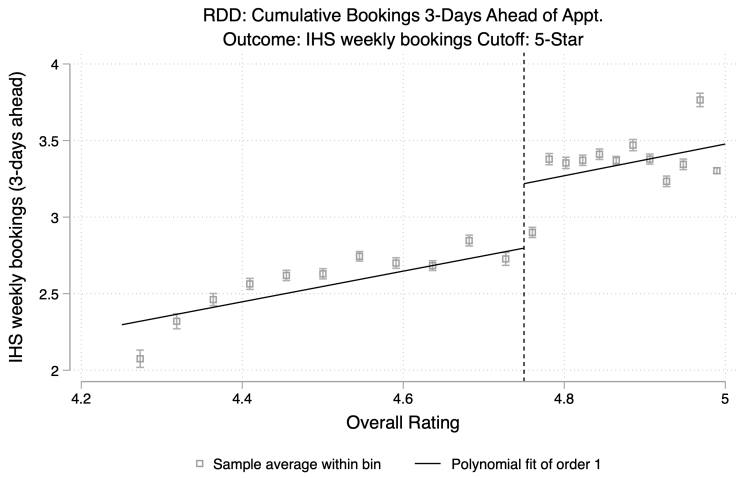
5.2 Regression Discontinuity Results: Impact of 5-Stars on Booking Volume

Figure 5.5 presents our RD results for the local average treatment effect of a five-star versus four-and-a-half star rating. Figure 5.5 presents the results for the weekly booking volume measured three days before the target appointment date. The left part of the figure displays the Regression Discontinuity plot, using the average overall rating as the running variable and booking volume as the variable on the vertical axis. On the right, the table shows the RD estimates calculated using conventional, bias-corrected, and robust methods.

We primarily rely on the robust result for our estimate. Our analysis identifies a substantial treatment effect of .761, which roughly approximates to a doubling of booking volumes. This finding

underscores the influence of star ratings on patients' booking decisions.

Figure 5.5: RDD Results Cumulative Bookings Three-Days Aheah of Appt.



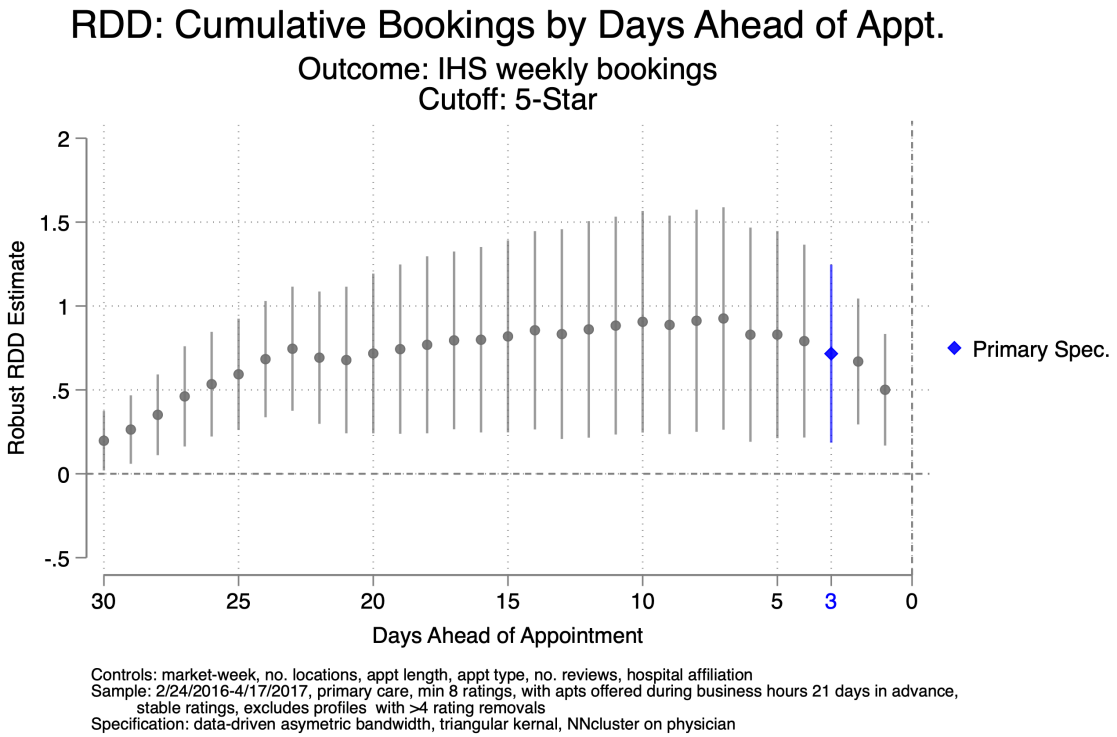
RD Method	(1) IHS(Weekly Bookings)
Conventional	0.686***
	(0.228)
Bias-corrected	0.716***
	(0.228)
Robust	0.716***
	(0.271)
Observations	54,263
Controls	✓

Controls: market-week, no. locations, appt length, appt type, no. reviews, hospital affiliation. **Sample:** 2/24/2016-4/17/2017, primary care, min 8 ratings, with appts offered during business hours 21 days in advance, stable ratings, excludes profiles with >4 rating removals. **Specification:** data-driven asymmetric bandwidth, triangular kernel, standard errors use NN clutters by physician

Next, we assess the treatment effect across various booking windows. We repeat our analysis for different cutoffs of days ahead of the appointments, considering the cumulative booking volume leading up to these dates. As depicted in Figure 5.6, the leftmost estimate represents our regression discontinuity (RD) estimate for cumulative bookings made at least 30 days ahead of the appointment dates and the rightmost estimate corresponds to cumulative bookings made up to one day before the appointment.

Our findings suggest that the treatment effect builds up until about a week before the appointment dates, then diminishes in the final week. One plausible explanation for this trend could be capacity constraints. Consumers show a clear preference for 5-star physicians, but as these top-rated physicians reach capacity and demand outstrips availability, 4.5-star physicians begin to receive bookings.

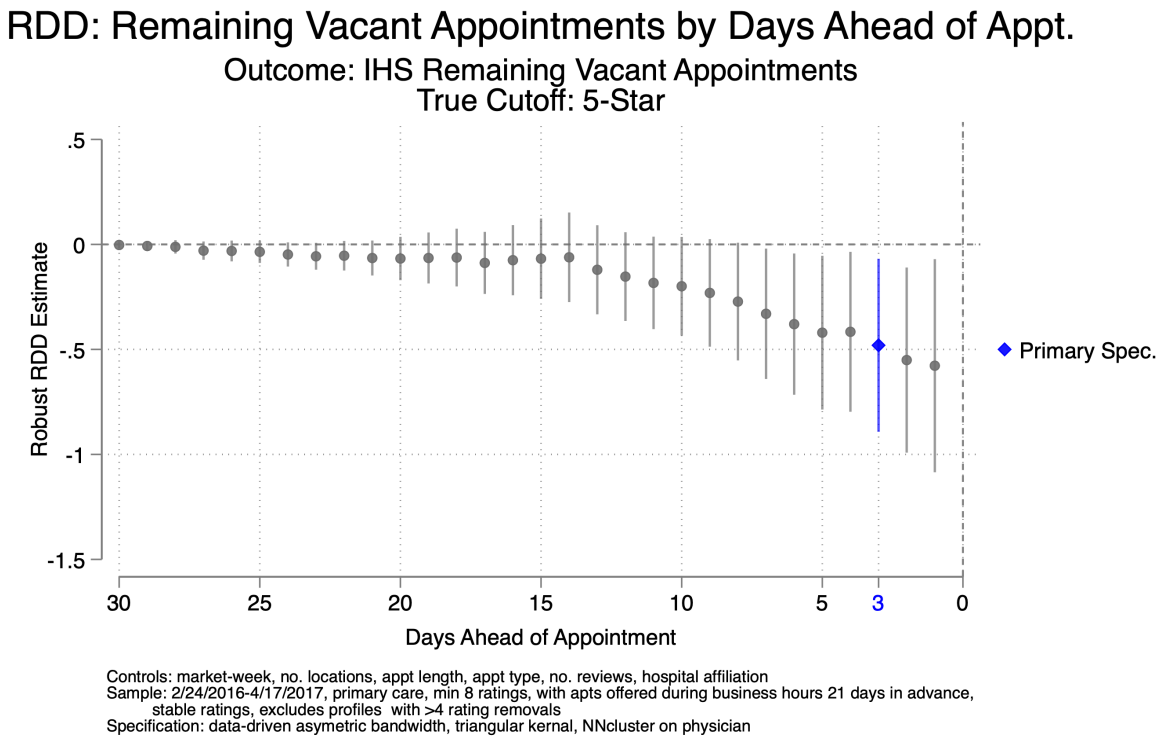
Figure 5.6: Booking Volume RD Estimates by Days in Advance of Appointment



5.2.1 Alternative Measure: Impact of 5-Stars on Vacant Appointment Volume

As an alternative measure of demand, we also analyze remaining vacant appointments. As physicians are capacity-constrained, measuring remaining capacity could also be a reasonable outcome. At the 4.75 cutoff, we find an effect where, compared to 4.5-star physicians, physicians with five stars have approximately 40% fewer vacant appointments three days before the appointment date. Figure 5.9 shows RD estimates from 30 days before the appointment dates to the day before the appointments.

Figure 5.7: Vacant Appointment Volume RD Estimates by Days in Advance of Appointment



5.3 Differential Impacts

Next, we evaluate the differential impact of ratings by partitioning our data into subgroups and repeating our analysis within each. In this section, we maintain our comparison between five-star and four-and-a-half-star ratings, focusing on weekly booking volumes measured three days ahead of the appointment dates. We concentrate on three comparisons of interest: physician gender, number of ratings, and hospital affiliation.

Our findings reveal that treatment effects are most pronounced for women physicians and those with a high number of ratings. There is a slight, but insignificant difference based on hospital affiliation, with a higher treatment effect noted for hospital affiliated physicians.

It is important to note that these results come with a caveat. While Regression Discontinuity Design estimates can have a causal interpretation under the right conditions, the differential impact might not necessarily hold this interpretation. Differences in parameter estimates could arise from heterogeneous treatment effects on other features, which might correlate with gender, number of ratings, or hospital affiliation.

5.3.1 Differential Impacts by Physician Gender

Figure 5.8 includes the RD plots of cumulative bookings three days ahead of appointments for women and men physicians separately, using the 5-star rating as the cutoff. The plots suggest an increase in booking volumes at the 5-star threshold for both genders. The figure also shows that women physicians have higher booking volumes relative to similarly rated men on both sides of the 5-star cutoff.

Figure 5.9 reports the RD estimates for booking volumes at the 5-star cutoff, separated by gender, and using three different estimation methods: conventional, bias-corrected, and robust. For reference, the figure also includes the pooled "base" estimates. The results indicate large significant effects on patient volumes for all subgroups and methods, though some results for men are significant at the 10% but not 5% level. Converting these point estimates to percentage changes, men see approximately a 106% higher booking volume, while women see a larger, approximately 276% higher booking volume.

Figure 5.8: Regression Discontinuity Plot of Booking Volume by Physician Gender

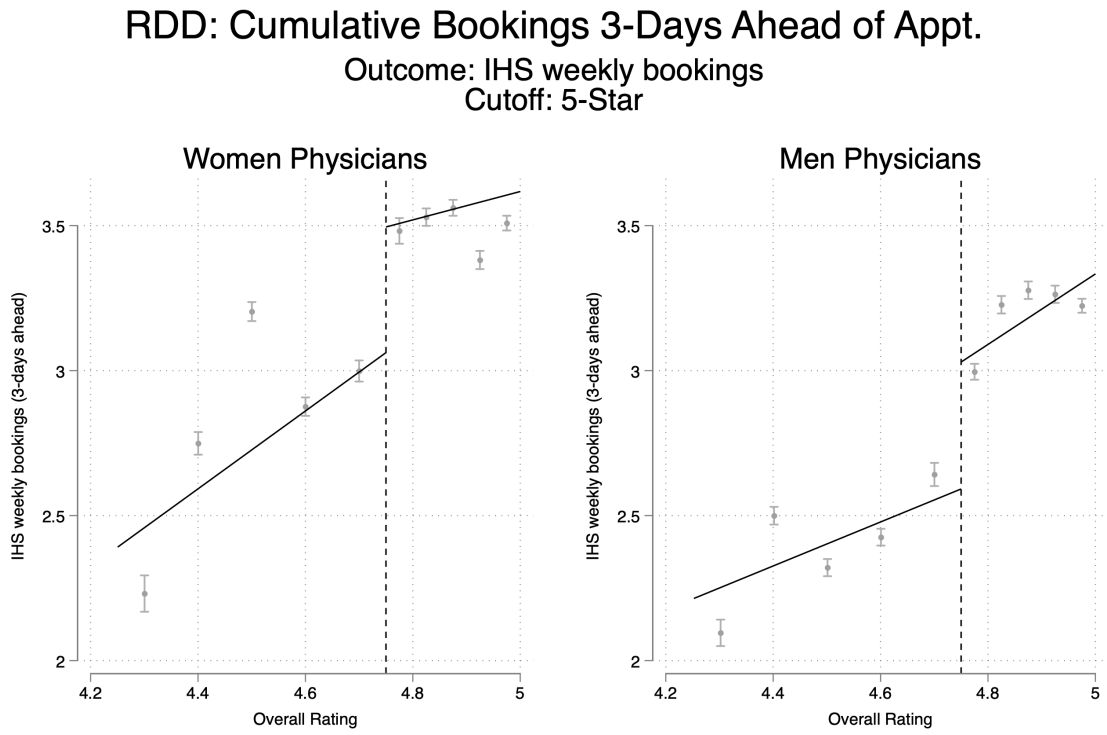
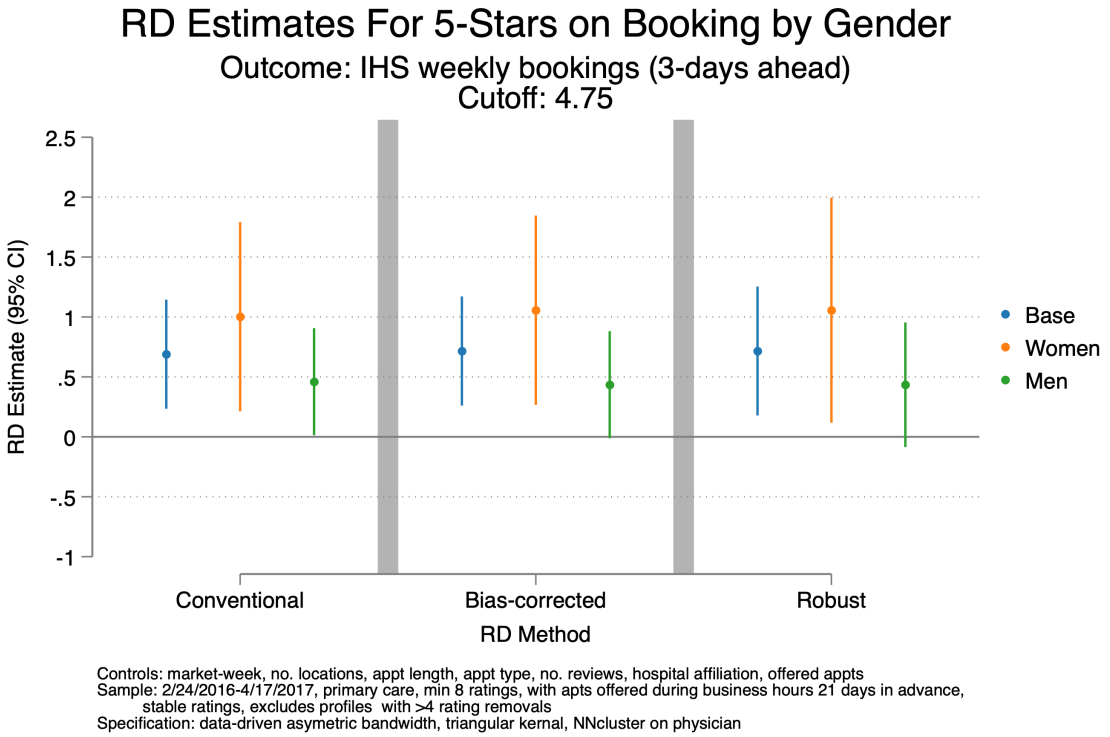


Figure 5.9: Regression Discontinuity Coefficients of Booking Volume by Physician Gender



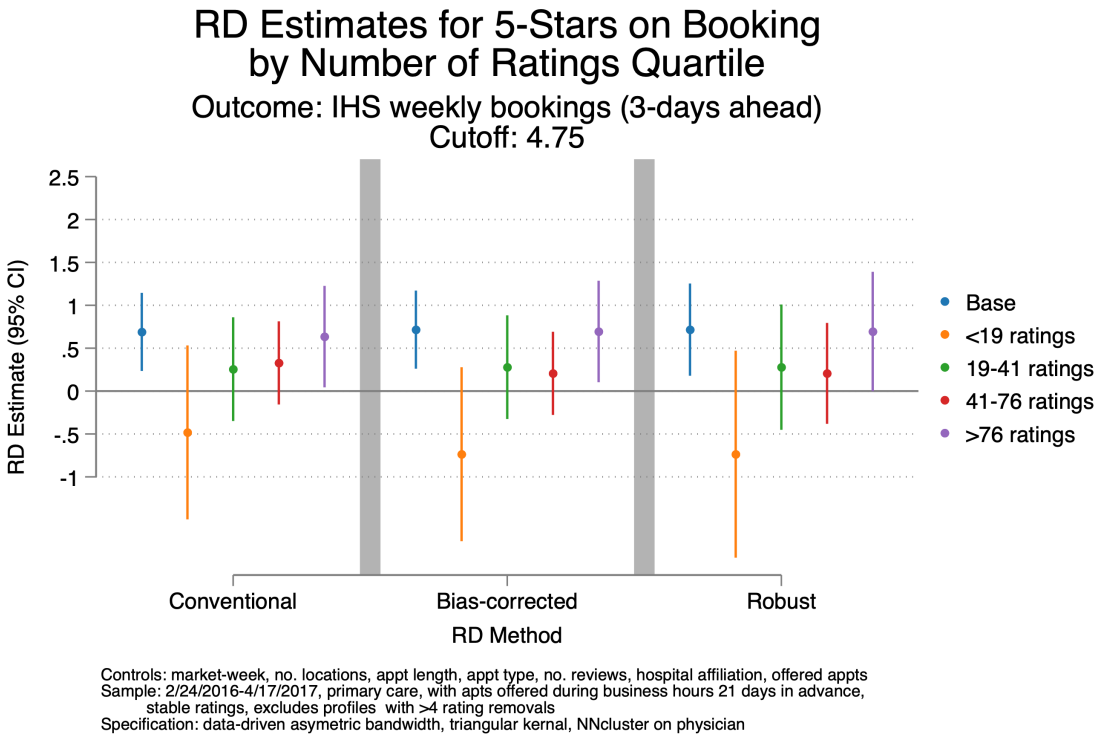
Controls: market-week, no. locations, appt length, appt type, no. reviews, hospital affiliation, offered appts
Sample: 2/24/2016-4/17/2017, primary care, min 8 ratings, with apts offered during business hours 21 days in advance, stable ratings, excludes profiles with >4 rating removals
Specification: data-driven asymetric bandwidth, triangular kernal, NNcluster on physician

5.3.2 Differential Impacts by Number of Reviews

Next, we report the differential impact of rating by number of reviews. We included the number of ratings as a control in our primary specification to account for the fact that the number of ratings on its own could impact booking volume directly as a signal or indicate tenure on the platform or popularity. However, this does not rule out the possibility that the impact of ratings is different by number of reviews. For example, Luca (2011) finds different effects of ratings by number of ratings in restaurant markets. Luca (2011) suggests this could be due to Bayesian learning, where each review is a signal of quality, and more positive reviews, at the same rating, provide a stronger signal of quality. This results in a greater impact of average rating when there are more reviews.

Figure 5.10 reports the coefficient plot for cumulative bookings three days ahead of appointments, separated by quartile of number of ratings. The smallest quartile includes physicians with less than 19 ratings, and the largest includes physicians with more than 76 ratings. We find that the effect is greatest and most significant for physicians in the highest quartile ($>$ 76 ratings). The results are not significant for the other quartiles. One of the possible drivers of this result is that consumers interpret ratings in a manner similar to Bayesian learning, whereby average rating has a greater impact when there are more ratings.

Figure 5.10: Regression Discontinuity Coefficients of Booking Volume by Number of Ratings



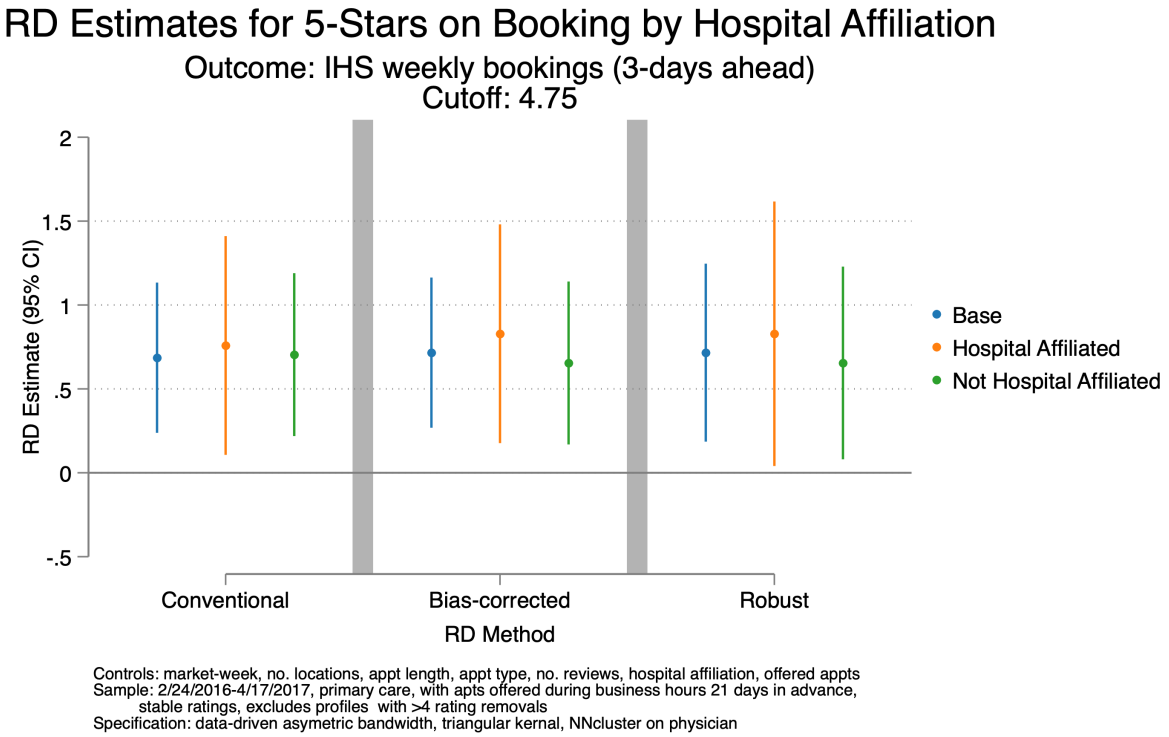
Controls: market-week, no. locations, appt length, appt type, no. reviews, hospital affiliation, offered appts
Sample: 2/24/2016-4/17/2017, primary care, with apts offered during business hours 21 days in advance,
stable ratings, excludes profiles with >4 rating removals
Specification: data-driven asymetric bandwidth, triangular kernal, NNcluster on physician

5.3.3 Differential Impacts by Hospital Affiliation

Next, we investigate the differential impact of ratings by hospital affiliation. Hospital affiliation could influence patient booking behavior as it may signal higher quality of care or access to better resources. We included hospital affiliation as a control in our primary specification to account for its potential direct impact on booking volumes. However, the impact of ratings may vary between hospital-affiliated and non-affiliated physicians. For example, patients might perceive ratings differently based on the perceived quality or reputation of the hospital affiliation.

Figure 5.11 reports the RD estimates for booking volumes at the 5-star cutoff, separated by hospital affiliation and using three different estimation methods: conventional, bias-corrected, and robust. For reference, the figure also includes the pooled "base" estimates. The results indicate significant effects on patient volumes for both hospital-affiliated and non-affiliated physicians across all methods. Although the magnitude of the effect appears to be slightly larger for hospital-affiliated physicians, there is no significant difference between hospital-affiliated and non-affiliated physicians. The results suggest that online reputation is a driver of booking volumes for both groups.

Figure 5.11: Regression Discontinuity Coefficients of Booking Volume by Hospital Affiliation



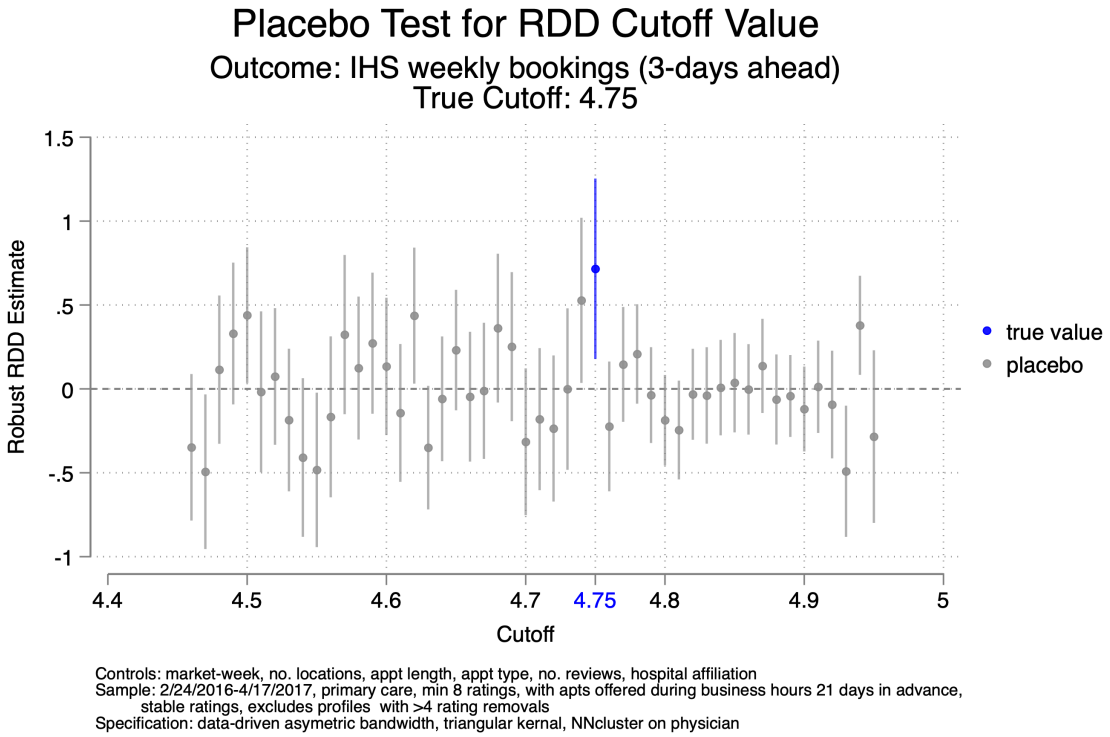
6 Robustness

In this section we investigate the robustness our primary specification using a placebo test and by investigating the distribution observations near the 4.75 cutoff.

6.1 Placebo Tests of Overall Rating at Alternative Cutoffs

As a placebo test, we repeat the RD analysis for specifying the discontinuity at alternative points other than the true cutoff. We find the greatest treatment effect estimate at the true cutoff of 4.75. There are however some points, for example 4.5, that would give false positive results.

Figure 6.1: Placebo Test of RDD Cutoff Value



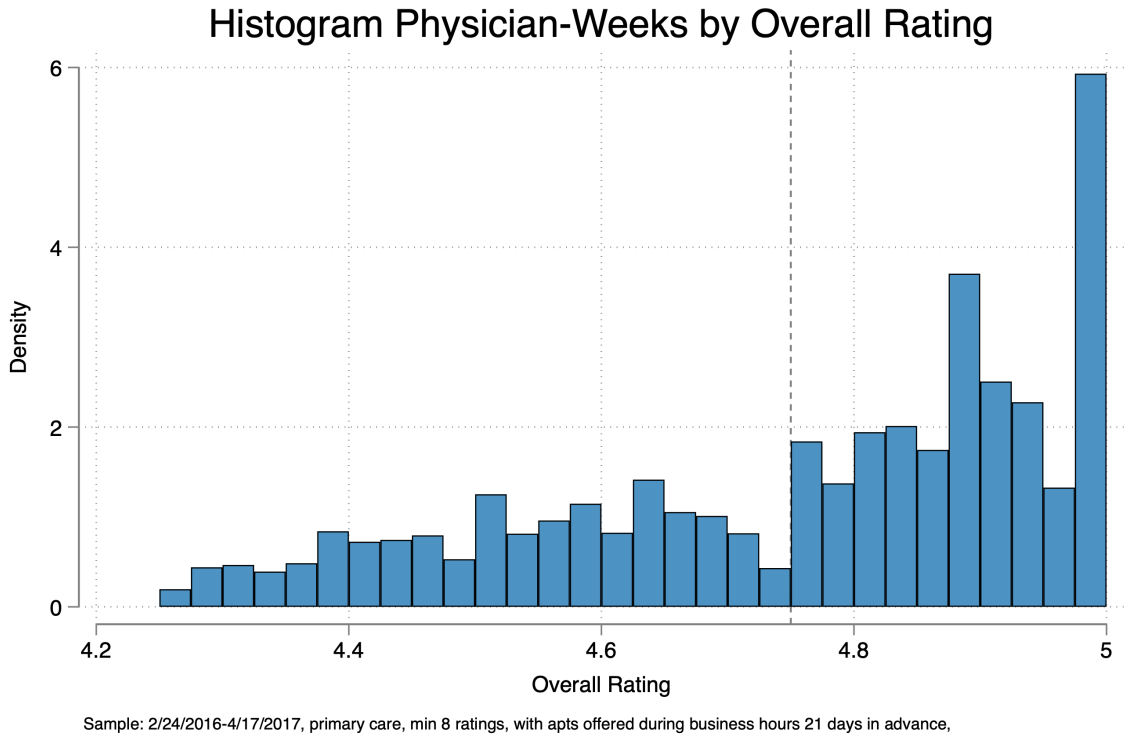
6.2 Rating Manipulation

A common concern when implementing RD methodology is gaming. In this setting, a physician that is near the overall rating discontinuity and knows that there is a discontinuous rating system may game the system to get their overall rating above the discontinuity. It is worth noting here that Zocdoc's closed loop rating limits the possibility of fake reviews. Only patients who have been

verified to have visited the physician after booking an appointment through Zocdoc are encouraged to leave feedback. However, a physician may still encourage favored patients to submit positive reviews, or dispute negative reviews with the platform. As documented in the appendix, we do in fact find evidence of deleted reviews.

If gaming is driving the results then one would expect overall ratings to be clustered just above discontinuities. Figure 6.2 shows the density of physician-weeks by overall rating. We do see some clustering just above the discontinuity. The next step is to test the possibility that gaming is biasing the results. This would involve implementing the density test from McCrary (2008).

Figure 6.2: Placebo Test of RDD Cutoff Value



7 Discussion

We find positive and significant effects of ratings on patient volume, and different impacts depending on gender and number of ratings. A number of mechanisms could drive these results. For example, patients could have correlated preferences for physician gender and for ratings, and these preferences could also be correlated with frequency of going to the doctor. Fink et al. (2020) find that patients have same gender preferences for both men and women physicians. Zocdoc reports that two in three

of their users on the patient side are female.[7]

We find that ratings have the greatest impact for physicians with many rating. These results on the differential impact of overall rating by number of reviews are consistent with Bayesian Learning, were the influence of average rating increases with the number of signals, in this case ratings.

7.1 Platform Mechanics

It is important to also consider how Zocdoc's recommendation system might mediate the impact of ratings on patient volumes. In a search model, consumer choices would hinge on both their preferences and search costs, the latter depending on positions on the page. If the recommendation system ranks physicians based, at least in part, on average or rounded ratings, the effect of these ratings on patient volumes would depend not only on consumer preferences for ratings but also on how these ratings affect search costs through the recommendation system.

Moreover, depending on the recommendation system's training, the system's sensitivity to ratings could vary for different searches. This factor introduces another layer of complexity when assessing the relationship between ratings and patient volumes.

7.2 Next Steps

This subsection briefly discusses the next steps of this paper along four dimensions: data and analysis, robustness, and framing.

7.2.1 Additional Data and Analysis

The current analysis focuses on the impact of the 5-star rating threshold on booking volumes. A valuable next step would be to expand this investigation to other rating cutoffs, such as 4-star or 4.5-star thresholds. This analysis would provide a more comprehensive understanding of how different levels of ratings affect patient behavior. By examining multiple cutoffs, we can identify whether the estimated effects are unique to the 5-star threshold or if they generalize across various rating levels. This can also help to understand if there is a nonlinear relationship between ratings and booking volumes. However, as noted in Section 3, most physicians have high ratings. Unlike other online reputation platforms such as Yelp, where it is typical to see a normal distribution of average ratings, on Zocdoc, ratings tend to be very high. There is not much data below the 4.5-star cutoff, so one might worry about selection issues when evaluating differences in low ratings.

[7] 12 Facts about Zocdoc Users

Beyond star ratings, the text of reviews can offer a deeper understanding of patient perceptions and the aspects of care that influence their decisions. Natural language processing (NLP) methods, including sentiment analysis and topic modeling, could reveal common themes and sentiments expressed in the reviews. The text of the reviews is noteworthy in three ways: 1) the direct impact of the text of reviews on booking volumes, 2) how the text of reviews changes the impact of ratings, and 3) what the text of the reviews can tell us about the information contained in ratings. Summarizing the differences in the sentiment of reviews by rating could inform us how much online ratings convey information about actual quality.

This paper evaluates the differential impact of ratings by gender, number of ratings, and hospital affiliation. As next steps, we plan to include additional comparisons by economically interesting groups and groups whose results would inform us about the underlying mechanisms driving the impact of ratings on booking volume. For instance, examining the differential impacts of ratings by apparent demographics inferred from profile photos could reveal biases or preferences related to demographics. Comparing subgroups based on specialty or education might uncover mechanisms driving the observed effects. Additionally, repeating our analysis for different booking windows, as we do in section X, could inform us about some of the mechanisms driving the large difference in booking and treatment effects by subgroups.

7.2.2 Dynamic Incentives and Physician Behavior

We find that the ratings, and by extension the rating system, influence which physicians consumers choose. Our results, some of which show more than doubling booking volumes, suggest a considerable incentive to receive good ratings. While potentially outside this project's scope, these findings open questions about how physicians might adjust behavior in response to the rating system. For example, do physicians take actions to improve ratings? If so, are these welfare-improving actions, such as improving quality of care? Do they include actions like deleting or contesting bad ratings?

7.2.3 Additional Robustness Checks

The paper would benefit from expanding the robustness checks and including those standard in the regression discontinuity design literature. A sensitivity analysis would repeat our analysis with different sampling rules and RD settings. A balance test would check for differences in physician characteristics across the cutoff. Additionally, a formal test for bunching around the rating thresholds, such as the McCrary (2008) test, might indicate strategic behavior by physicians to improve ratings

to get just above a respective cutoff. Given the apparent mass point just above the five-star cutoff, one approach is donut regression discontinuity, where we would drop all observations near the cutoff. These robustness checks will help ensure the reliability and robustness of the conclusions drawn from the analysis.

7.2.4 Framing

There is also room to improve this paper to better frame it at the intersection of digital markets generally and healthcare specifically. In many settings, consumers use online ratings to choose experience goods such as hotels, restaurants, and movies, or retail settings with ex-ante, but not ex-post, uncertainty about product quality. Physician services are a credence good with ex-ante and ex-post uncertainty about product quality. On one hand, one might expect ratings to be less important in consumer decision-making since online ratings might be less informative about quality, as those leaving reviews are not fully informed about the quality of care they receive. On the other hand, consumers might appear even more sensitive to ratings since they have even less ex-ante information about medical services. Our planned analysis that looks at the impact of ratings in the presence of other quality information, as well as analyzing the text of reviews, could improve the paper in this direction.

8 Conclusion

Using data collected from Zocdoc, this paper analyzes a unique data set containing physicians' appointment schedules, professional information, and reviews from verified patients. Zocdoc displays ratings on a scale of one to five stars, with overall average ratings rounded to nearest half star. Since ratings are rounded to the nearest half star, we use a regression discontinuity framework to identify the causal impact of patient reviews on patients' choice of physician. On the appointment level, our results indicate that a half star improvement in displayed rating means that appointments are more likely to be booked and also booked further in advance. On the physician-week level we find that 5-star physicians have higher patient volume, and that these results are most pronounce for women physicians, and physicians with many ratings. We test the robustness of our model using alternative specifications and placebo tests.

References

Brown, Z., C. Hansman, J. Keener, and A. Veiga (2023, March). Information and Disparities in Health Care Quality: Evidence from GP Choice in England. Technical Report w31033, National Bureau of Economic Research, Cambridge, MA.

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

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare (2020, December). Analysis of regression-discontinuity designs with multiple cutoffs or multiple scores. *The Stata Journal: Promoting communications on statistics and Stata 20*(4), 866–891.

Chan, A. (2023). Discrimination Against Doctors: A Field Experiment.

Cutler, D. M., R. S. Huckman, and M. B. Landrum (2004, April). The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery. *American Economic Review 94*(2), 342–346.

Dafny, L. and D. Dranove (2008, September). 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.

Darby, M. R. and E. Karni (1973). Free Competition and the Optimal Amount of Fraud. *The Journal of Law & Economics 16*(1), 67–88. Publisher: [University of Chicago Press, Booth School of Business, University of Chicago, University of Chicago Law School].

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

Feng Lu, S. (2012). Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes. *Journal of Economics & Management Strategy 21*(3), 673–705. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1530-9134.2012.00341.x.

Fink, M., K. Klein, K. Sayers, J. Valentino, C. Leonardi, A. Bronstone, P. M. Wiseman, and V. Dasa (2020, January). Objective Data Reveals Gender Preferences for Patients' Primary Care Physician. *Journal of Primary Care & Community Health 11*, 2150132720967221. Publisher: SAGE Publications Inc.

Jin, G. Z. and A. T. Sorensen (2006, March). Information and consumer choice: The value of publicized health plan ratings. *Journal of Health Economics 25*(2), 248–275.

Luca, M. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.Com. *SSRN Electronic Journal*.

McCarthy, I., K. Sanbower, and L. S. Aragon (2022). Online Reviews and Hospital Choices.

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

Norton, E. C. (2022, September). The inverse hyperbolic sine transformation and retransformed marginal effects. *The Stata Journal: Promoting communications on statistics and Stata 22*(3), 702–712.

Sarsons, H. Interpreting Signals in the Labor Market: Evidence from Medical Referrals.

Wang, J., J. Hockenberry, S.-Y. Chou, and M. Yang (2011, March). Do bad report cards have consequences? Impacts of publicly reported provider quality information on the CABG market in Pennsylvania. *Journal of Health Economics 30*(2), 392–407.

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

Appendix

A Additional Figures

Figure A.1: Vacant Appointment Volume RD Estimates by Days in Advance of Appointment

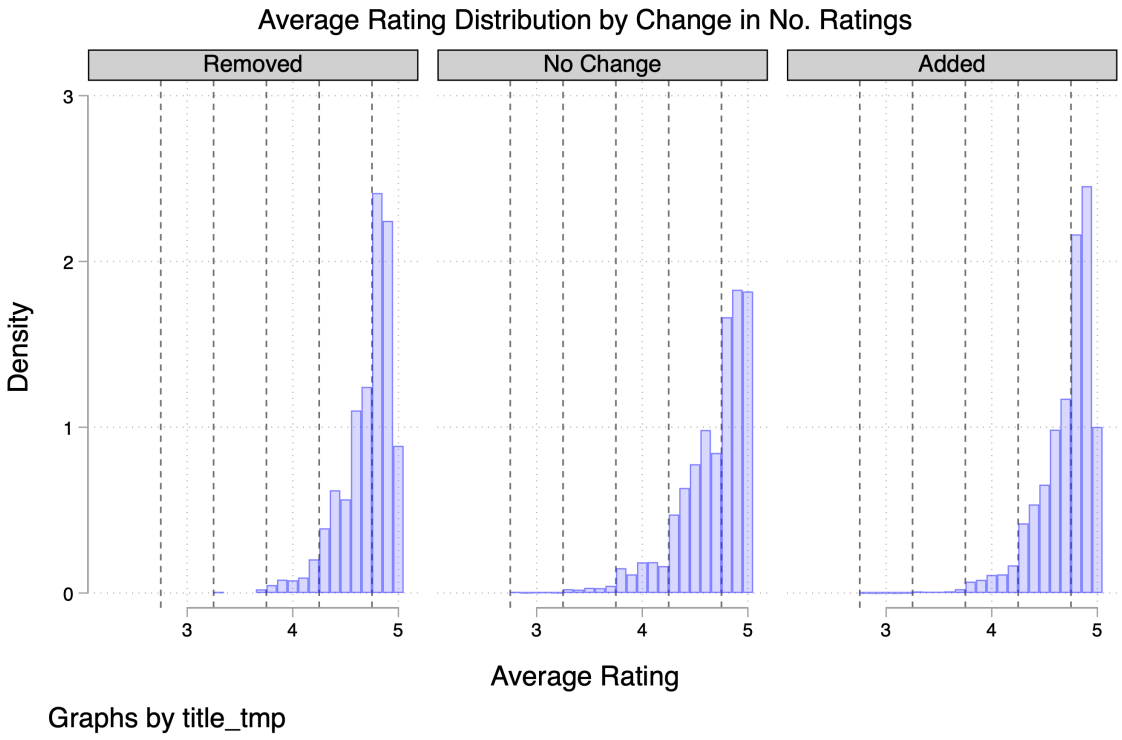


Figure A.2: Vacant Appointment Volume RD Estimates by Days in Advance of Appointment

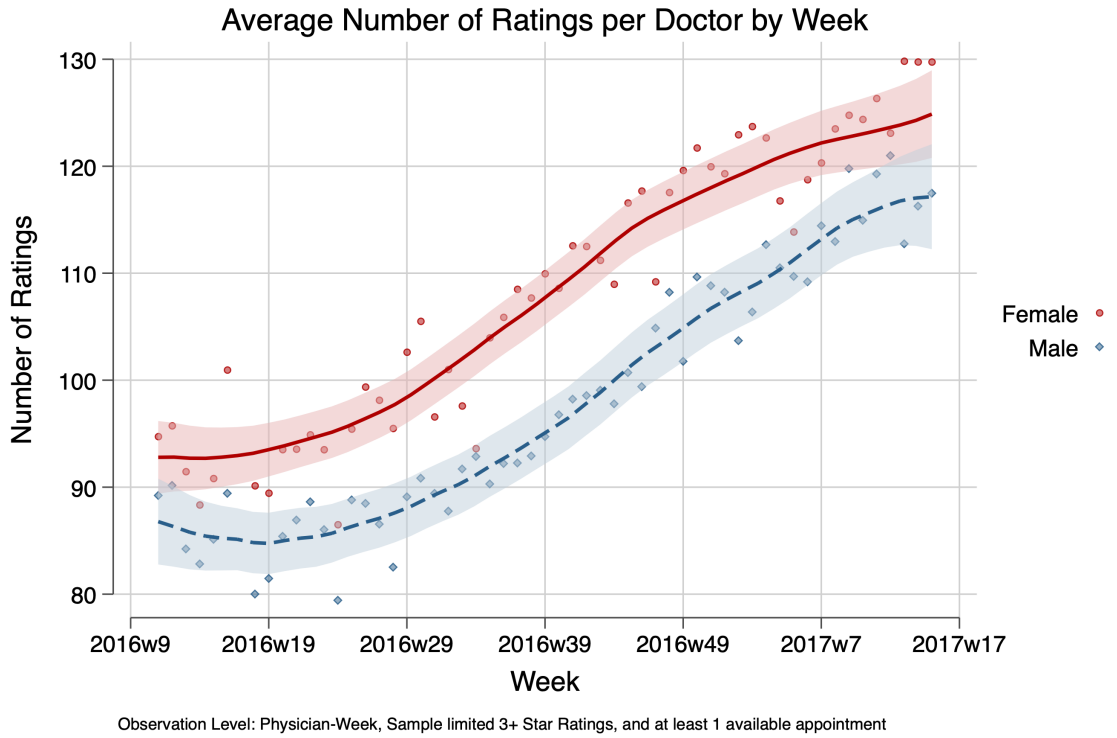


Figure A.3: Vacant Appointment Volume RD Estimates by Days in Advance of Appointment

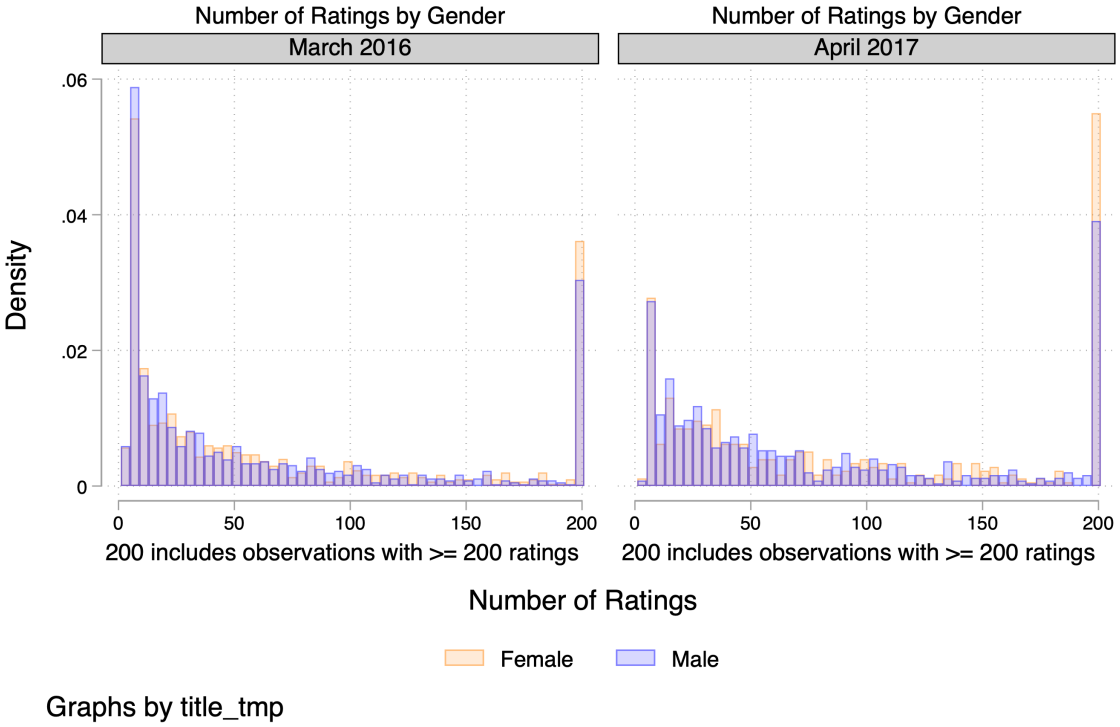
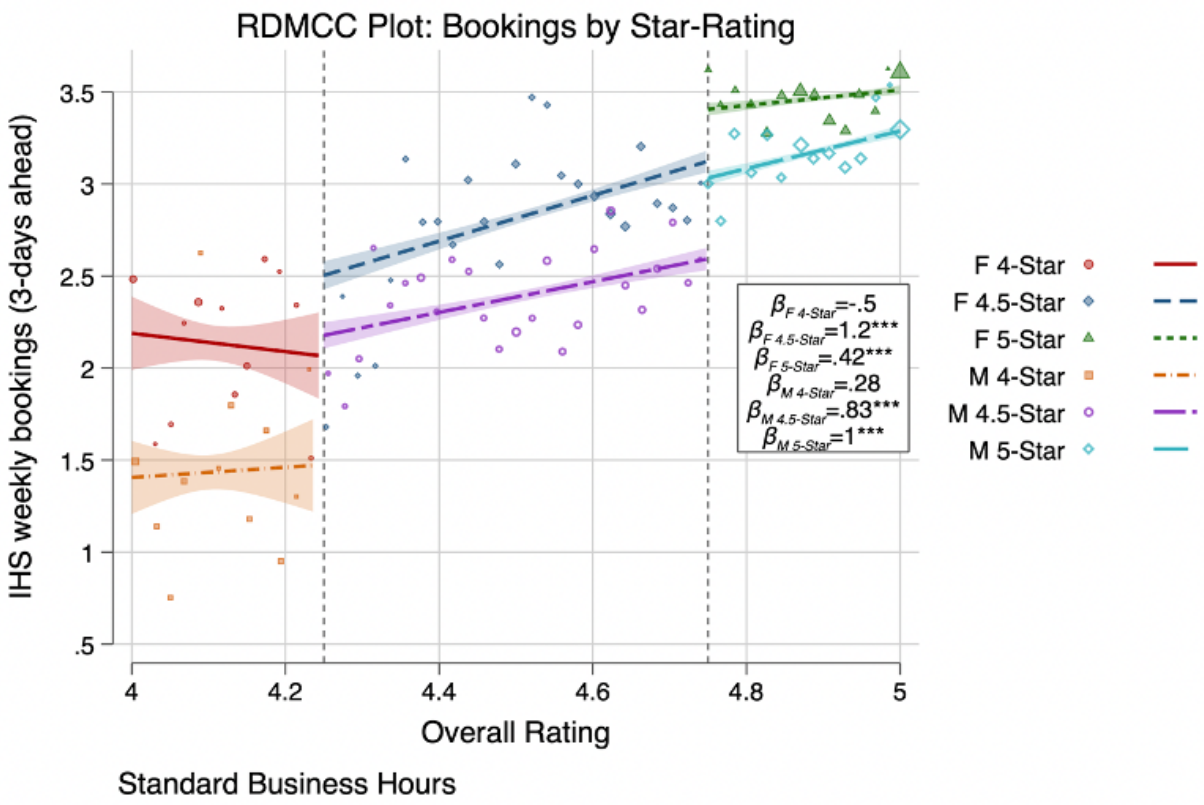


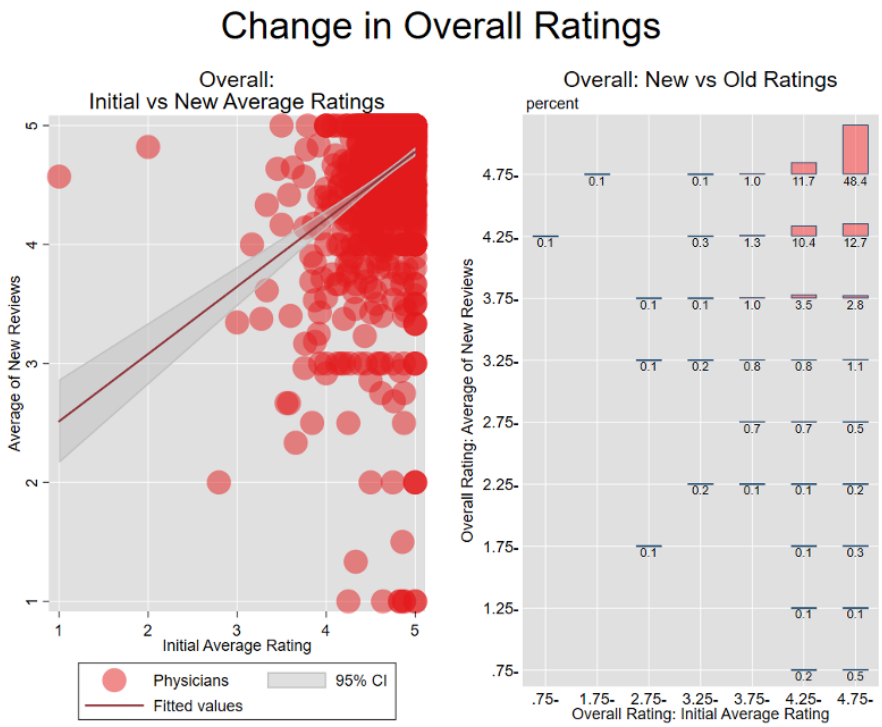
Figure A.4: Vacant Appointment Volume RD Estimates by Days in Advance of Appointment



B Changes in Ratings

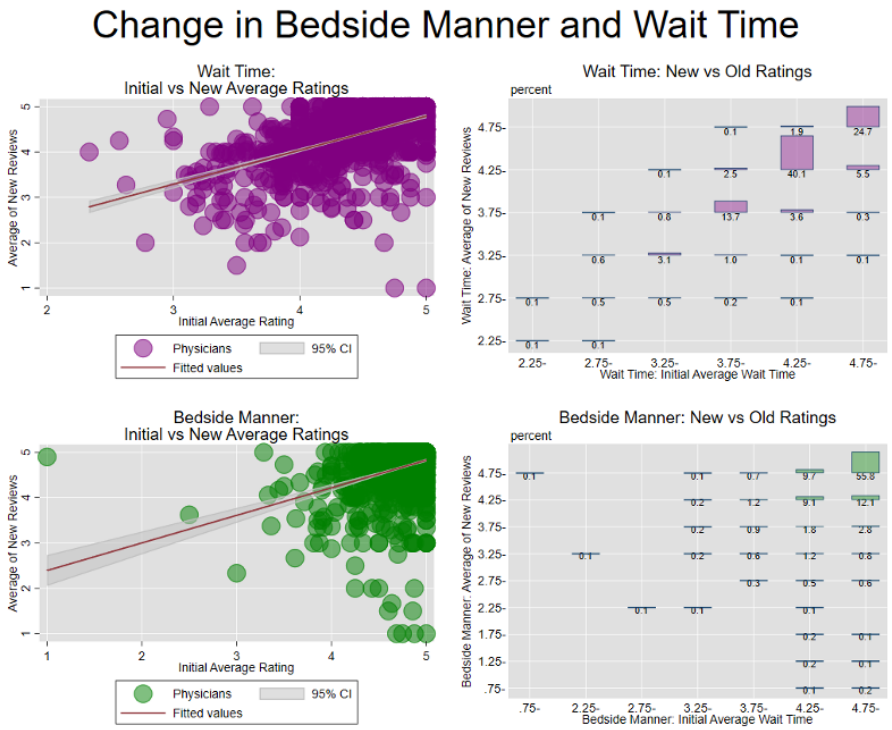
Next, we consider if this platform is causing physicians to improve or change their performance. If patients respond to these ratings, then poorly performing physicians have incentive to change their behavior and the physicians receiving good reviews have incentive to maintain their performance. Additionally, these reviews could be providing feedback the physicians would not otherwise receive. To answer these questions, we look at changes in the ratings over the period of the study. These results are based on a selected a random sample of 1,314 physicians that were persistent in the entire sample. We compute their initial average ratings, which are their overall, bedside manner, and wait time ratings as of February 26, 2016. Next, we compute the averages of all of their new ratings as of February 25, 2017. Figure B.1 & Figure B.2 illustrate the averages of new ratings (y axis) by initial average rating (x-axis). On the left, there are scatter plots with linear fits of average rating. The right has a two-way tabplot of these same data at the half star level. As an example, on the right plot in Figure B.1 , we see that 12.7 percent of physicians had initial overall ratings in the 5-star category and received new ratings that were on average in the 4.5-star category.[8]

Figure B.1: Overall Rating Transitions



[8]We round to the nearest half star, so 5-stars would be [4.75, 5], 4.5 star is [4.25, 4.75) and so on.

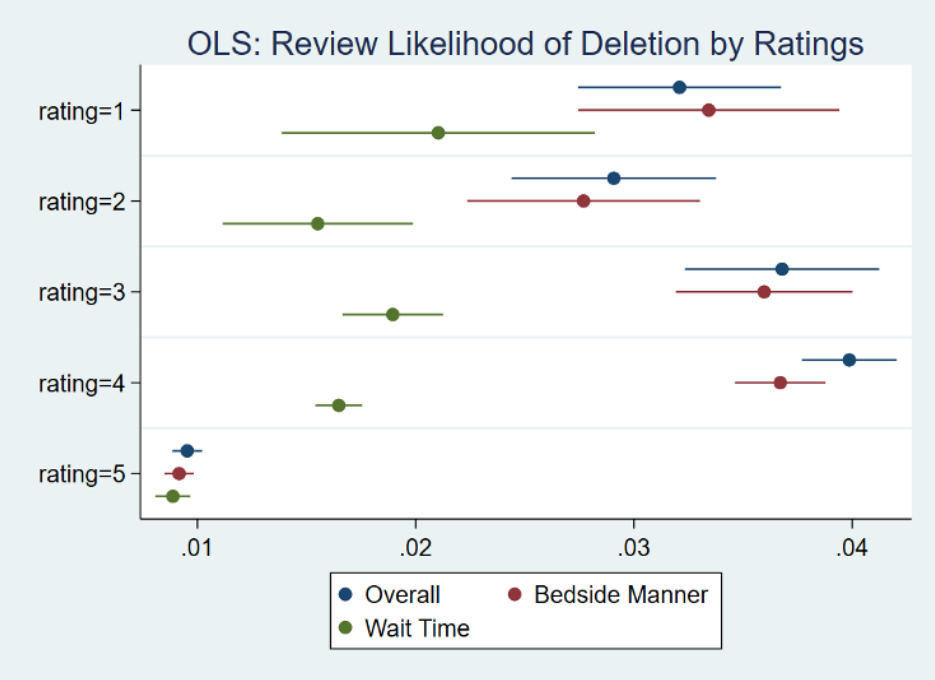
Figure B.2: Bedside Manner and Wait Time Rating Transitions



C Deleted Reviews

Here we look at the likelihood of a review being deleted based on the ratings in that review. The results suggest that for overall, bedside manner, and wait time, any review with ratings below 5 stars is more likely to be deleted than 5-star reviews.

Figure C.1: Coefficient Plot - Linear Probability of Review Deletion by Rating



We plan to look into this further, in particular, the interaction terms are interesting. Intuitively one doctor might have an incentive to delete a 4-star review, while another might want to keep it, depending on their average rating, and the bedside manner and wait time ratings in the given reviews.