Can We Trust Online Physician Ratings?

Evidence from Cardiac Surgeons in Florida

Susan F. Lu

Krannert School of Management, Purdue University, West Lafayette, IN 47907

Huaxia Rui

Simon Business School, University of Rochester, Rochester, NY 14627

Forthcoming in Management Science

Abstract

Despite heated debate about the pros and cons of online physician ratings, little systematic work has examined the correlation between physicians' online ratings and their actual medical quality. Using the ratings of cardiac surgeons at RateMDs and the patient outcomes of coronary artery bypass graft surgeries in the 2013 Florida Hospital Inpatient Discharge Data, we investigate whether online ratings are informative about physicians' medical quality. To account for potentially non-random matchings of patients of different severity levels to surgeons of different rating categories, we focus on patients who arrived through the emergency department and explicitly consider how observed and unobserved patient health conditions jointly affect the matching arrangements and surgical outcomes. Both reduced form and two-stage estimation results show that, compared with surgeons rated four stars or higher, or those without rating information, lower rated surgeons are associated with significantly higher in-hospital mortality rates. Our findings suggest that online physician ratings could be a valuable information source for patients to learn about physician quality, at least for cardiac surgeons, a specialty for which treatment outcomes are relatively observable to patients and their family members.

JEL: L15, I1

Key words: word of mouth, physician ratings, medical quality, sound of silence

"Nearly two thirds of the 2,137 people polled in the Internet-based survey said that physician ratings were at least somewhat important in choosing a doctor."

— NBC News[1]

"They told me that 'patients aren't smart enough to figure out whether I'm a good doctor.'"

— The New York Times[2]

1. Introduction

Recently, online physician ratings have sparked heated debate.[3] Advocates argue that these ratings reflect physicians' less observed quality and help consumers search for physicians more efficiently, while critics worry that patients' lack of medical knowledge, limited numbers of reviews, and possible reporting bias may inhibit this unique type of online word of mouth (WOM) from disseminating reliable information. Patients and the public are left to wonder: Can we trust online physician ratings as an informative source of physicians' medical quality?

Unlike online ratings of many other goods and services, such as movies, books, and restaurants, when it comes to online physician ratings, the stakes are high, for both the physicians and the patients. The experience of mandatorily disclosing quality information and bringing more transparency to the public demonstrates some of the pitfalls of providing publicly available health care ratings. In the last decade, the Centers for Medicare & Medicaid Services (CMS) initiated several report card policies that require health facilities such as hospitals, nursing homes, and home care agents to release their quality information to the public. Unfortunately, such initiatives directed to individual physicians failed to gain traction. In a well-known initiative beginning in 1990, New York State has been publishing physician coronary artery bypass graft (CABG) surgery mortality rates. As a result, physicians who care for more difficult or severely ill patients are penalized and have strong incentives to decline to treat such patients (Dranove et al. 2003). As of 2013, all but 10 of the 50 U.S. states have failed to find ways to provide their residents with meaningful information on physician quality.[4] Efforts on the mandatory disclosure of physician quality have progressed slowly.[5]

The emergence of online physician ratings introduces a completely different type of information channel to the landscape of physician quality disclosure. Online WOM offers *subjective* quality measures, which arguably do not suffer from measurement selection and might be better than an *objective* measure

[1] "Online doctor ratings important to patients," www.nbcnews.com.

[2] See http://www.nytimes.com/2012/03/10/your-money/why-the-web-lacks-authoritative-reviews-of-doctors.html?pagewanted=all&_r=0.

[3] CBS News, http://www.cbsnews.com/news/can-you-trust-online-doctor-rankings/.

[4] See the report "50 State Report Card on Physician Quality Transparency" released by the Health Care Incentives Improvement Institute (http://www.hci3.org/sites/default/files/files/IssueBrief-Dec2013.pdf).

[5] In 2014, the CMS started reporting the quality of care ratings for physician group practices; the ratings for individual physicians are not available as of August, 2014. (http://www.medicare.gov/physiciancompare/staticpages/data/aboutthedata.html).

in certain aspects.[6] Nevertheless, as some critics worry, these online physician ratings may introduce noise and disturb the status quo because they can be misleading and harmful.[7]

In this study, we aim to address a fundamental question about online physician ratings: Do they inform us about the medical quality of physicians? Although this question engenders considerable debate among policy makers, physicians, and consumers, the literature examining this issue is relatively sparse for two reasons. First, it is not easy to find a good performance measure that reflects physicians' medical quality.[8] Second and more importantly, the assignment of patients with different health conditions to physicians with different ratings is unlikely to be random, which poses a challenge to researchers in uncovering the true association between physicians' online ratings and their medical performance. The non-random assignment of patients of different severity levels to physicians of different ratings may come from either the patient side or the provider side. On the patient side, severely ill patients may be more likely to seek those physicians with great ratings. On the provider side, physicians who are capable may attract patients, which allows them to cherry-pick patients. Alternatively, hospital administrators may assign severe patients to skilled physicians to improve overall hospital performance. Without carefully addressing these selection issues, the estimation of the association between physicians' online ratings and their medical quality would be biased.

To meet the first challenge, we select one specialty of physicians as the study subjects: surgeons who conduct CABG surgeries. Coronary artery disease is one of the leading causes of death in the United States (Serruys et al, 2009). Statistically, the non-trivial chance of death allows us to use in-hospital mortality rates to measure cardiac surgeon performance, a widely accepted practice in the health economics and management literature (e.g., Huckman and Pisano 2006, Serruys et al. 2009, KC and Terwiesch 2011, Clark and Huckman 2012).

To address potential selections, we focus on CABG surgeries on patients who arrived through the emergency department (ED). Because patients sent to the ED for CABG are typically in urgent conditions, the chance of them or their families checking surgeons' online ratings and making selection accordingly is small. Hence, investigating the association between surgeons' online ratings and their medical performance using ED patients suffers from little patient selection bias. Besides, the internal scheduling of surgeons is pre-determined and patients could not plan their heart attack in advance. Based on these institutional facts, we start from a linear probability model (LPM) assuming that cardiac surgeons are randomly assigned to patients in the ED.

[6] It is easy to game with objective measures. For example, Dranove et al. (2003) shows that physicians may shun away sick AMI patients after the report card released the mortality rate to the public. Lu (2012, 2016) shows that nursing homes may take a "teaching-to-the-test" strategy to game with the released objective measures.

[7] See http://www.usatoday.com/story/news/nation/2014/02/18/online-doctor-ratings/5582257/.

[8] Non-medical performance measures on which physicians are rated include interpersonal skills, effectiveness in persuading patients to stop smoking, and giving flu shots.

Nevertheless, since most of the ED patients who need CABG could not undergo such a procedure immediately but have to wait for a few days in the hospitals till their health conditions permits the operation,[9] it naturally raises the concerns on the possibility of provider selections during this window between patients' boarding in hospitals and undergoing CABG. As a major robustness check, we relax this random assignment assumption and consider the possibilities that surgeons of different skill levels might be matched to patients in different health conditions. To do this, we model how both observed and unobserved patient health conditions jointly affect the patient–surgeon assignment and surgical outcome. In the surgeon performance model, we specify the probability of death right after cardiac surgery as a function of patient, surgeon, and hospital characteristics and surgeon online ratings. We then introduce a discrete choice model for patient–surgeon matching whose error term shares the same unobserved patient health factor as the error term in the performance model. Finally, the two models are jointly estimated by full information maximum likelihood (FMLE).

Using historical physician rating data from RateMDs and 2013 Florida Hospital Inpatient Discharge Data, we find that cardiac surgeons with below-four-star overall quality ratings are associated with higher mortality rates compared to higher rated surgeons and those without rating information. Moreover, we surprisingly find that surgeons without ratings perform no worse than those rated at least four stars. These results remain qualitatively the same to various robustness checks.

Our work makes contributions along three dimensions. First, we contribute to the online consumer word of mouth (WOM) literature by investigating an important and timely question of whether consumers can trust online ratings to guide their search for one type of credence goods, physician services. To the best of our knowledge, whether online reviews actually contain performance information for professional (credence) services is hitherto unexplored. Second, our work is valuable to consumers, physicians and hospital administrators by directly associating patient clinical outcomes with physician online ratings. Our results show that low ratings do signal low physician quality, which highlight the important role that physician online ratings could play when consumers search physicians for surgical treatments. Last but not least, our work reveals that surgeons without ratings perform better than those with low ratings and no worse than those with high ratings, which is surprising, and seems to be inconsistent with the literature finding regarding the "sound of silence". One possible explanation of this counter-intuitive finding is that quality signals from five star reviews might have been diluted because of the non-uniformity in the probability of being rated for physicians of different skill levels and also the possible existence of positive fraud reviews. This finding calls for change of consumer norms used for online physician search, especially when a large number of physicians in a specialty are unrated.

[9] http://patient.info/doctor/acute-myocardial-infarction-management

2. Literature Review and Hypotheses Development

2.1 Literature Review

Although Internet-based consumer ratings of physicians have gained much attention and generated considerable debate about whether patients can trust them, the literature examining whether and to what extent online physician ratings reflect physicians' medical performance is relatively sparse (Emmert et al. 2013a).

Our work contributes to a growing body of papers on online physician reviews. In a seminal work, Gao et al. (2012) analyze the physician ratings at RateMDs.com and find that online reviews are generally quite positive for those physicians with reviews. There are positive correlations between rating scores and observable physician characteristics such as physician experience, board certification, education, and the absence of malpractice claims. Segal et al. (2012) and Luca and Vats (2013) show that physicians with high ratings are associated with high demand, suggesting that physician ratings have become an important source of reputation for physicians. In a systematic review, Emmert et al. (2013b) summarize frequently asked questions about physician rating websites, such as who have been rated, the number of reviews, and differences in ratings related to socioeconomic status.[10]

Most relevant to the present work is a small set of papers that test the relationship between online physician reviews and their quality. Greaves et al. (2012) document the weak association between Internet-based patients' ratings and various measures of clinic quality in primary care (e.g., the proportion of patients with diabetes receiving flu vaccinations) without considering any selection issues. Gao et al. (2011) find that the association between online reviews and perceived physician quality is the weakest for five-star physicians. Our study considers different types of selection issues and use clinical medical quality measure to further our understanding of whether a physician's online rating is informative of the physician's actual medical performance.

More broadly, our work contributes to the online consumer WOM literature. This literature comprises two major branches of study. One branch focuses on the *production* side of online reviews by studying questions such as how a review was generated, what consumer characteristics lead to a good or bad review, and how to identify fraud reviews (i.e., Dellarocas and Wood 2008, Dai et al. 2013; Luca and Zervas, 2016; Godes and Silva 2012). For example, by comparing online and offline WOM, Lovett et al. (2013) find that, whereas the social and functional drivers are the most important for online WOM, the emotional driver is the most important for offline WOM.[11] The other branch emphasizes the *consumption*

[10] A total of 24 articles are reviewed, including those of Emmert and Meier (2013), Emmert et al. (2013a), and Ellimoottil et al. (2013).

[11] Consumers spread WOM for three fundamental purposes: (1) The function driver refers to the motive to provide and supply information; (2) the social driver refers to the motive to send a social signal to the environment; and (3) the emotional driver refers the motive to share positive or negative feelings to express or ease emotional arousal (Berger and Milkman, 2012). See http://www.kellerfay.com/what-drives-online-vs-offline-word-of-mouth-major-differences-revealed-in-new-academic-study/.

side of online reviews by studying questions such as how the generated ratings affect consumer behavior. For example, a recent 2013 Nielsen report shows that WOM is not only the most trusted source of information, but also the most likely to stimulate consumers to action.[12] This stream of literature has tested the impact of WOM under various settings, including electronics reviews at sites such as Amazon, eBay, and Taobao (Chevalier and Mayzlin 2006, Cabral and Hortacsu 2010, Mudambi and Schuff 2010), movies and music reviews at Yahoo!, Rotten Tomatoes, and Twitter (i.e. Liu 2006, Duan et al. 2008, Dhar and Chang 2009, Chintagunta et al. 2010, Rui et al. 2013), restaurant food reviews at Yelp (Luca 2011), and hotel reviews at different travel agency websites (Vermeulen and Seegers 2009). Most of the settings studied involve search or experience goods, whose quality consumers have sufficient knowledge to judge. Whether online reviews actually contain quality information for professional (credence) services is underexplored. In this study, we aim to understand whether, given potential shortcomings in the review production process, consumers can trust online ratings to guide their search for physician services.

2.2 Hypothesis Development

The main objective of this study is to assess whether online physician ratings are informative of physician medical quality. Although there is reason to believe that online physician rating systems will become commonplace, given recent debates on this topic, it is far from certain that these ratings convey reliable information about physicians' quality. In this section, we summarize current theories and develop testable hypotheses.

The existing WOM literature endorses online WOM as a useful and efficient channel for disseminating quality information to consumers (e.g., Yang et al. 2012). Although patients and their families are unlikely to have the medical training to directly judge the medical quality of a physician, it is still possible for them to gain some quality signal through their own experience. Typically, a patient and his or her family members have a reasonable understanding of the patient's health conditions before and after receiving medical services. From the improvement in health conditions and by comparing that of other patients suffering from similar diseases, the patient and his or her family members can infer some quality signal about the physician. By definition, a high-quality physician is more likely than a low-quality physician to successfully help a patient improve his or her health conditions. Hence, it is reasonable to expect that patients treated by high-quality physicians are more likely to be satisfied than patients treated by low-quality physicians. As long as satisfied patients (or their family members) are more likely to contribute good ratings and less likely to contribute bad ratings than less satisfied patients (or their family members) do, then good rating should indicate high quality.[13]

[12] See http://www.idiro.com/2013/09/nielsen-report-finds-that-word-of-mouth-is-the-most-trusted-source-again/.

[13] This argument can be formally presented and rigorously analyzed with a mathematical model, which is available in the Online Appendix.

To take the CABG surgery as an example, family members can observe at least whether their beloved survived after the surgery in hospital or not, and they typically also have some knowledge about the health conditions of the patient. Based on their perceived improvement or deterioration, they can infer how well the surgery was performed and how good the surgeon is, although not completely accurate. If online reviews voluntarily contributed by web users carry sufficient credibility and accuracy, we expect that online physician reviews, as one particular type of online WOM applications, should be a valuable source in delivering credible and useful information to consumers just like the online reviews of many other products despite the different mechanisms through which quality information is generated. Therefore, we would expect the following hypothesis to hold.

Hypothesis A (Validity Hypothesis): *All else being equal, patients treated by physicians with low ratings have worse medical outcomes than patients treated by physicians with relatively high ratings.*

Alternatively, physicians' online reviews have unique features that may inhibit online WOM from disseminating reliable information. First, as an example of credence goods, the quality of care provided by physicians is difficult for consumers to observe ex ante and to verify ex post. Unlike experience goods, consumers typically do not have sufficient medical knowledge to evaluate a physician's medical quality (Arrow 1963).[14] A patient with minimal medical training tends to infer a physician's quality from the physician's interpersonal skills rather than from the physician's true medical quality (Dranove 2008).

Second, individuals' characteristics and their related demographics could influence whether they will post a review on a website and how they will rate a physician. For example, Dellarocas and Wood (2008) document that individuals with positive opinions are more likely to post reviews. As a result, online ratings may be associated with intrinsic biases (Gao et al. 2011). These concerns cast doubt on the validity of online physician reviews. If the bias due to consumer behavior in posting a review is strong enough or the signal-to-noise ratio in the ratings is sufficiently low, we would expect the following alternative hypothesis to hold.

Hypothesis B (Null Hypothesis): *All else being equal, patients' medical outcomes are uncorrelated with physicians' online ratings.*

It should be noted that, unlike what some critics argue, the limited number of reviews a physician receives is not a sufficient condition that conceptually lead to the null hypothesis. Rather, the limited number of reviews makes it less accurate, in a statistical sense, to infer the medical quality of a given physician who is rated by patients. In other words, when considering the small number of reviews of rated

[14] It usually takes eight to ten years of post-graduate medical training to become qualified to practice medicine.

physicians, one should not confuse its implication at the individual level with its implication at the population level.

One unique feature of online physician ratings is the large proportion of physicians who have not received ratings yet. Therefore, it is important to compare the performance of these physicians with those with ratings. The theory of "sound of silence" suggests that lack of ratings for a product or service signals poor consumer satisfaction. This theory is derived from a *bidirectional* feedback mechanism such as eBay (Dellarocas and Wood, 2008). To avoid being retaliated and rated poorly by the seller parties, buyers tend to keep silent on those platforms after receiving a poor product or service. Clearly, in the context of online physician ratings, the feedback mechanism is *unidirectional* because patients are not rated by physicians. Because physician services are credence goods provided locally, online physician platforms evolve much more slowly than those rating platforms for other goods. We suspect that lack of ratings may simply reflect consumers' lack of willingness to post a review rather than signal low quality. Hence, the performance of physicians without ratings remains unclear ex ante, especially given the large population of this group of physicians.

3. Institutional Knowledge

3.1 RateMDs

RateMDs[15], launched in 2004, is one of the earliest physician review websites in the United States and records the largest number of user-submitted reviews with narratives (Lagu et al. 2010). According to Gao et al. (2012), as of January 31, 2010, there were 368,559 physician ratings at RateMDs, covering about 16% of all practicing U.S. physicians. The likelihood of being rated varies widely across specialties and is consistent across geographic regions: 32.4% of obstetrician/gynecologists, 24.6% of medical specialists, 20.0% of surgeons, and 16.3% of primary care physicians had received a rating.

Users voluntarily contribute all the reviews and the rating information is publicly available and free to use. RateMDs acknowledges the possibility of duplicate or false ratings and has rules and procedures to minimize the effect of such manipulation. For example, the RateMDs systems are set up to remove multiple entries coming from the same computer, and if their system detects multiple ratings coming from the same source, it may require new raters to login before rating for some period of time. If a doctor thinks a rating should be removed, the doctor can "flag" the rating to initiate an investigation by RateMDs.[16] The site also posts the names of doctors who require waivers to a so-called "Wall of Shame,"

[15] http://www.ratemds.com

[16] For more details, please visit http://www.ratemds.com/about/faq.

which highlight doctors detected by RateMDs.com who make prospective patients sign “gag contracts” before they are accepted as patients.[17]

Several other websites also offer physician reviews, among which three websites specialize in online physician ratings (HealthGrades, Vitals, ZocDoc) and six review a broad array of businesses and services, including physicians (Avvo, Citysearch, InsiderPages, Yahoo!Local, Google Maps, and Yelp) (Segal et al. 2012). We focus on RateMDs.com for the following reasons. First, this website, started in 2004, was the first site specializing in physician reviews in the United States and has accumulated the largest number of user-submitted reviews with narratives so far (Gao et al. 2012). Second and most important to our study, this specialized website records historical patient reviews and their corresponding ratings for each physician, which allows us to recover historical ratings before a patient underwent surgery. HealthGrades only reports the most updated overall quality ratings; hence, we cannot trace historical ratings. Similar issues exist for Vitals.[18] Third, unlike RateMDs, where ratings for cardiac surgeons are available, rating information for cardiac surgeons is rare on Zocdoc, perhaps because Zocdoc specializes in online scheduling while such scheduling for CABG surgery is not practical.

Figure A1 plots the number of reviews as well as the number of rated physicians per year from 2004 to 2013 for all physicians in Florida. In 2004, there were 227 reviews and 205 physicians were rated by web users. Over the years, the numbers have increased dramatically. In 2013, there were 72,654 reviews and 21,799 rated physicians. The average number of ratings per physician was 3.3 in 2013. We limit our sample period to 2013, the most recent year for which inpatient data are available, so that we can obtain a sufficient number of reviews and rated physicians. We also compared the ratings of cardiac surgeons across multiple websites and confirmed that online physician ratings posted at RateMDs are representative.[19]

3.2 Selection of CABG Surgeons

Coronary artery disease is one of the leading causes of death in the United States (Serruys et al. 2009).[20] Patients with severe blockages or multiple narrowing of the coronary arteries are normally treated by CABG, a risky and invasive surgical procedure.

We study CABG surgeons for the following reasons. First, CABG is a common procedure and surgeons performing CABG are very likely to be rated. According to the 2007 National Hospital Discharge Survey, over 405,000 CABG surgeries were performed, accounting for about 5% of total U.S. health expenditures. Second, unsuccessful CABG is directly linked to death, which is an unusual patient

[17] See http://www.msnbc.msn.com/id/34794632/ns/health-health_care/.

[18] At Vitals, the posting dates of ratings not accompanied by a text review are not available to website visitors. This drawback and the fact that there are very few text reviews on Vitals.com make it not only infeasible but also unreliable to recover historical ratings at right before a patient undergoes surgery.

[19] The comparison table is available from the authors upon request.

[20] Source: National Center for Health Statistics (http://www.cdc.gov/nchs/fastats/lcod.htm).

outcome compared to other non–life-threatening procedures. Such a feature allows us to use in-hospital mortality to measure the performance of CABG surgery, which has been widely accepted in the health economics and management literature (e.g., Huckman and Pisano 2006, Serruys et al. 2009, KC and Terwiesch 2011, Clark and Huckman 2012) and in physician report cards in New York (Dranove et al. 2003). Third, a surgeon's skill is crucial to the success of a CABG procedure. However, a surgeon's medical skills are difficult for the public to observe ex ante and to verify ex post. Given the key role played by the surgeon, it is important for the public to know if the online physician ratings reflect surgeons' medical quality.

Patients usually undergo CABG through two channels: emergency and non-emergency department (henceforth, ED and non-ED). Patients and their selected surgeons can schedule a surgery appointment with a local hospital prior to a non-ED surgery. Clearly, the matching between a patient and a surgeon is two-sided in the sense that the matching outcome is affected by the preferences of both sides. On the contrary, in some life-threatening cases, patients hit by a sudden heart attack or heart failure must undergo CABG surgery after being sent to the hospital ED. Unlike the non-ED situation where patients can select their own surgeons, patients may have to accept a surgeon scheduled by hospital administrators in an emergency case. According to the Florida Hospital Inpatient Discharge Data, 37.3% of CABG patients were transferred from hospital EDs in 2013 and the remaining were non-ED surgeries. The in-hospital mortality rate was 2.6% for the emergency cases and 1.5% for the non-ED cases. In this study, we focus on the cases of ED arrival.

4. Data

This study incorporates two main datasets: the online rating information from RateMDs and the Florida Hospital Inpatient Discharge Data. RateMDs provides the physician's name, address, phone, graduation year, and online ratings over time, while the Florida Hospital Inpatient Discharge Data records the license number of each operating physician, as well as basic patient characteristics including payer types, race, gender, and health outcomes.[21] We use Florida's physician license verification data, which provides the physician's license number, name, and address, to merge the two main datasets together.[22] In addition, we supplement the main datasets with hospital information from the American Hospital Association (AHA) Annual Survey and Hospital Compare and market demographics from the Area Resource File.

[21] We found that there are some unexpected recording errors about attending physicians and operating physicians in the raw Florida data during the data cleaning process. For example, some physicians who are coded as operating physician for a corresponding CABG surgery are in fact anesthetists, not cardiac surgeons. We excluded all the cases whose physician team does not have a doctor specialized in cardiac diseases in the final sample.

[22] Florida license verification information is available at http://ww2.doh.state.fl.us/IRM00PRAES/PRASLIST.ASP.

Many physicians across different specialties are rated online. In this study, we focus on one type of physicians: surgeons who perform CABGs.[23] Our full sample covers all reviews from 2004 to 2013 on 246 surgeons who performed at least one CABG in the entire year of 2013 in Florida (including those without online ratings) and 3,819 patients who lived in Florida and underwent CABG surgery after being transferred from a hospital ED in 2013. Table 1 provides the definitions, means, and standard deviations of the variables used in the ED sample, our primary interest.

4.1 Online Ratings

The physician ratings at RateMDs cover four dimensions: *helpfulness*, *knowledge*, *staff,* and *punctuality*.[24] Consumers rate physicians in each of the dimensions on a scale of one (lowest) to five (highest). An overall physician quality measure is automatically generated based on the average of the *helpfulness* and *knowledge* scores.

We download all historical reviews and conduct simple text mining based on the reviews for the cardiac surgeons in the sample. Figure A2 shows the word cloud from the text of the online reviews. The most frequently used word is s*urgeries*. We also list some examples of the text reviews in Table A1. It seems that some users specify *CABG* in their reviews while most users simply mention *surgeries*. We acknowledge that these reviews may not focus on the specific skills required for CABG. Nevertheless, they may be reasonable proxies for surgical skills perceived by patients.

Using the historical data, we recover the mean ratings up to the quarter before a patient underwent CABG surgery. We define *high-rating* surgeons as those with overall ratings equal to or above four stars and *low-rating* surgeons as those with overall ratings below four stars.[25] Surgeons without reviews are grouped together as *no-rating* surgeons. This categorical approach offers several advantages for this study. First, it helps capture surgeons without ratings in the estimation model, whereas a continuous measure of ratings cannot. Second, the dichotomous model of ratings mitigates the under-representation of negative opinions because reporting bias may change cardinal information of ratings, but have little effect on ordinal information for rated products.[26] This simple approach is also adopted by YouTube to adjust for rating bias. Third, from the modeling perspective, keeping the number of surgeon categories small makes the two-stage estimation computationally tractable.

Panel A of Table 2 reports the rating information for the surgeons who are involved in ED cases across categories. The unit of observation is the surgeon-quarter. It seems that high-rating surgeons and

[23] CABG surgery information is available at http://en.wikipedia.org/wiki/Coronary_artery_bypass_surgery.

[24] RateMDs does not provide an explanation of these four dimensions to the public. Reviewers interpret them based on the meanings of the four words.

[25]The four-star cutoff is determined by a spline model analysis reported in Table A2. The results show that ratings are negatively correlated with adjusted mortality rate in the segment between 1 and 4 and there is no statistically significance for that correlation in the segment between 4 and 5.

[26] See http://youtube-global.blogspot.com/2009/09/five-stars-dominate-ratings.html.

low-rating surgeons have similar patient volume, with an average 6.3 and 6.4 CABG surgeries, respectively, for ED patients per surgeon per quarter. These numbers are statistically indifferent.[27] We also notice that those surgeons with low ratings have a lower patient volume in the total number of CABG surgeries (ED and non-ED) per surgeon per quarter than those with high ratings. This seems to be consistent with the view in the literature that demand and quality are positively correlated (Segal et al. 2012, Luca and Vats 2013).

4.2 Performance Measure: In-Hospital Mortality

Our estimation model captures the performance quality of each surgeon using the dependent variable of in-hospital mortality. This is a critical quality measure that is widely accepted in the health economics and management literature (e.g. Huckman and Pisano 2006, Serruys et al. 2009). The Florida Hospital Inpatient Discharge Data record the surgical outcome for each patient who undergoes CABG surgery. We define the in-hospital mortality as equal to one if the patient died before being discharged from the hospital and zero otherwise. As shown in Panel B of Table 2, the average mortality rate was 2.6% in 2013. Without considering any risk adjustments or possible selection bias, we find low-rating surgeons have the highest mortality rates (6.3%) among all categories of surgeons.

Panel B of Table 2 also shows surgeon characteristics across the surgeon categories. These characteristics include the number of years since graduation (experience), the absence of malpractice claims, and attendance at elite schools (education).[28] These measures may be proxies for physician quality (Gao et al. 2012) and are used to explain surgical outcomes. Overall, only one surgeon attended elite school in our sample and only two surgeons had malpractice claims in our sample. The number of years since graduation varies little across the surgeon categories.

4.3 Patient and Hospital Characteristics

The Florida Hospital Inpatient Discharge Data provide patient characteristics such as age, gender, race, payer type, diagnosis codes, and zip code. We use zip code information to infer each patient's income level and later to calculate the patient's travel distance to the hospital. Using ICD-9 diagnosis codes, we extract patient risk factors, such as heart valve and chronic kidney diseases.[29]

The hospital characteristics, obtained from the 2011 AHA data, include a hospital's scale, location, and the availability of a cardiac intensive care unit (CICU). We also obtain hospital rankings from Hospital Compare.[30] Table A3 shows the correlation matrix for main independent variables.

[27] The coefficient of high-rating is -0.118 with p value 0.900. The coefficient of no-ratings is -0.301 with p value 0.725.

[28] The elite school is defined as the top 10 medical schools ranked by US News. Besides, all surgeons in our sample are board certified. The boards include American Board of Surgery, American Board of Thoracic Surgery, American Board of Vascular Medicine and Society of Cardiovascular and Intervention. Some surgeons are board certified and specialized in cardiac diseases according to RateMDs, but we are unable to get their detailed board information.

[29] For the list of patient risk factors, please refer to Table 1 where we provide the definition and summary statistics of the key variables used in our models.

[30] See http://www.medicare.gov/hospitalcompare/search.html.

4.4 The ED Arrivals: A Quasi-random Assignment

We choose the cases of ED arrivals to alleviate the concerns about the potential selections (Hurwitz et al, 2014). The rationales are two-fold. On the patient side, according to the U.S. Emergency Medical Treatment and Active Labor Act,[31] an ambulance usually sends a patient suffering from heart attack to the nearest hospital that is equipped with required treatments for both medical and legal reasons. Patients with heart attack are typically in urgent conditions, and are unlikely to check online physician reviews immediately and select a particular surgeon in a receiving hospital. On the provider side, the assignment of a surgeon to an ED patient is primarily a function of the internal scheduling of physicians (e.g., day and shift) and the time at which the patient had a heart attack. Because the internal scheduling of physicians is pre-determined and a patient cannot plan heart attack in advance, the match between surgeons and patients should be largely exogenous. Based on these institutional facts, we assume that ED patients treated by surgeons of different rating categories have similar levels of severity.

To empirically examine this identification assumption, we first conduct pairwise t tests on all patient characteristics by surgeon categories. The results in Table A4 show that the majority of the patient characteristics are similar across surgeon categories while a few variables are statistically different in the pairwise comparisons. One possible explanation is that patients with certain characteristics and surgeons with different ratings are not distributed evenly across different hospitals (and regions).

To further alleviate the concern of selection, we follow a standard approach used by KC and Terwiesch (2011) in testing selective admission based on health conditions observable to researchers. We first regress mortality over a set of *observed* patient preoperative risk factors and a constant term. Then we predict a preoperative mortality rate for each individual patient based on the patient's own health information. Finally, we regress this predicted variable over the surgeon's rating types and a set of hospital fixed effects. If patients treated by surgeons of different ratings are significantly different in terms of their risk factors, we would expect significant correlation between rating status and the predicted preoperative mortality rate.

Table 3 reports the results. We use the continuous ratings (excluding surgeons without ratings) as the explanatory variable in Column (1) and (3), and use the categorical ratings as the explanatory variable in Column (2) and (4). In Column (3) and (4), we additionally include socioeconomic status variables (e.g., insurance types, income in the neighborhood) in predicting preoperative risk, which are excluded in Column (1) and (2). Using the results in Column (4) as an example, we find the coefficients of *high-rating* and *no-rating* to be small and insignificant, suggesting that there is no significant difference in terms of patient preoperative risk between high-rating surgeons and low-rating surgeons, and between no-

[31] https://www.cms.gov/Regulations-and-Guidance/Legislation/EMTALA/

rating surgeons and low-rating surgeons. The F test result further suggests that there is no significant difference in patient preoperative risk between high-rating surgeons and no-rating surgeons either.

One caveat is that this test is performed on those comorbidity variables observable to researchers. Even though there is no evidence that surgeons of different ratings select patients with different levels of preoperative risk, we have to be cautious that we cannot entirely rule out the possibility of some degree of selection based on unobserved risk factors. Therefore, we rely on the quasi-random setup in our main analysis, but check the robustness of our findings by estimating a two-stage selection model.

5. The Reduced Form Approach

5.1 Surgeon Performance Model

We start from a linear probability model (LPM) assuming that neither patients nor hospitals select cardiac surgeons in the cases of ED arrival. Surgeons of different rating statuses are randomly assigned to patients who arrive at a hospital ED. Below is the baseline performance equation describing the association between physician ratings and patient outcomes:

$$Y_{ijhq} = I[\beta_0 + \beta_1 \cdot R_{jq-1} + \beta_2 \cdot V_{iq} + \beta_3 \cdot W_{jq} + \beta_4 \cdot H_h + \beta_q + \varepsilon_{ijhq} \geq 0] \qquad (1)$$

where Y_{ijhq} refers to the health outcome of patient i who received CABG surgery from surgeon j in hospital h in quarter q and equals one if patient i died in the hospital after surgery before being discharged and zero otherwise. R_{jq-1} refers to the online ratings of surgeon j in the previous quarter, which is classified into three categories: no ratings (N), low ratings (M), and high ratings (S). Hence, our key explanatory variable R belongs to the set $\{N, M, S\} \equiv$ A. We treat the category of low ratings (M) as the omitted category in the regression and denote the set of other categories as $A_0 \equiv \{N, S\}$ for convenience. The term V_{iq} is a vector of observable patient characteristics, including age, gender, race, co-morbidities, CABG types and distance to a hospital; W_{jq} is a vector of observable surgeon characteristics, such as experience, the square of experience, CABG volume, the square of CABG volume and whether the surgeon had malpractice claims;[32] and H_h stands for hospital characteristics such as the hospital ranking, size, and whether a CICU is available at the hospital. We denote these patient, surgeon, and hospital characteristics and the constant terms as $X_i = [1, R, V, W, H]$. We also include the quarter dummies β_q to capture the seasonality effect on patient outcomes. Because of potential serial correlations for patients treated by the same surgeon, we cluster the standard errors by surgeons.

[32] We have information about board certification, attending elite school and crime records. There are little variations in these physician characteristics. Hence, we did not include them in the estimation.

5.2 Main Results

Table 4 reports the results about the correlation between surgeon rating status and patient mortality. To make sure that the results are not driven by multicollinearity, we gradually add in different sets of control variables. Column (1) includes both patient and hospital characteristics. Column (2) additionally adds surgeon characteristics to examine whether rating reflects any performance information after we control the observable physician characteristics. Because the probability we are modeling (i.e., mortality rate) is below 10%, the log odds ratio is a highly nonlinear function of probability in this interval. Hence, a logit model is a good alternative to a LPM. Column (3) and (4) report the estimation results from the logit model with surgeon characteristics excluded in Column (3) and included in Column (4). In Column (5), we excluded patients with a history of heart attack to alleviate the concern that those patients may choose to live near a specific hospital, which might violate the random assignment assumption in the ED setting. Overall, all the key results are qualitatively the same across these specifications.[33]

Taking the results in Column (2) as an example, the coefficient of *high-rating* is negative and significant at ten percent significance level (p=0.052), suggesting that low-rating surgeons are associated with higher mortality rates than high-rating surgeons. Translating the magnitude of the coefficient, we find that being treated by a high-rating surgeon can increase the survival chance of a patient in an emergency by 3.5 percentage point on average compared with being treated by a low-rating surgeon. When we change the LPM analysis to a logit model, the coefficient of *high-rating* in Column (4) remains to be negative and significant at five percent significance level (p=0.028). The difference in mortality rate for a median patient between being treated by a high-rating surgeon and a low-rating one is 7.7 percentage point.

To gain an understanding of the size effect of the findings, we present here some findings from the medical literature about the mortality rate of CABG. Hannan et al (1994) shows that the risk-adjusted in-hospital mortality decreases by 1.7 percentage point or 41% from 4.2% in 1989 to 2.5% in 1992 after the release of physician CABG report cards in New York State. On the physician level, Hartz et al., (1997) evaluated the performance of "best" CABG surgeons, defined by the book "The Best Doctors in America" and articles from city magazines, using data from three states: New York, Pennsylvania, and Wisconsin. They found that the observed mortality rate is 3.4% for "best" surgeons and 4.4% for other surgeons who are not listed by the book or the magazines, given that they have similar average predicted mortality rate (3.5%). Therefore, without risk adjustment, the difference between "best" surgeons and possibly low-quality surgeons is around 1 percentage point. Considering the mortality rate for CABG patients in the

[33] In a robustness check, we replace these hospital characteristics with hospital fixed effects. The results are available upon request.

medical literature, our results suggest that being treated by a surgeon with a high rating rather than one with a low rating, a patient's survival chance would see a significant increase which we believe is sizeable enough for the patient and family members to take the ratings into serious consideration.

Interestingly, we find that patients treated by surgeons without ratings also have a lower mortality rate than those treated by low-rating surgeons, all else being equal. The coefficient of *no-rating* is negative and significant at five percentage significance level. Further, we compare the performance between *high-rating* and *no-rating* surgeons. The F-test shows that there is no statistical significance between these two categories of surgeons (p=0.95).

We also find other interesting results in the surgeon performance model. First, patients who are old or female are more likely to die after cardiac surgeries, all else being equal. Second, some patient risk factors, such as chronic kidney disease or with history of CABG, significantly increase the mortality rate. Third, surgeons' observable characteristics, such as experience, CABG volume, and their quadratic terms, are insignificantly correlated with mortality rates. In addition, a patient's travel time to a hospital does not affect the patient's mortality rate after controlling for the patient's risk factors.

5.3 Positive Fake Reviews

Although RateMDs takes some effective steps to filter out fake reviews,[34] one may still be concerned that the possible existence of fake reviews could lead to inaccurate understanding of the informational value of those authentic reviews (Luca and Zervas, 2016). In particular, positive fake reviews are more difficult to be detected than negative reviews because physicians will have incentive to appeal fake negative reviews but no incentive to report fake positive reviews. Furthermore, five-star reviews may also be the result of physicians' explicit or implicit "lobbying" of their patients to contribute ratings. Hence, being rated five-star may also reflect a physician's effort of self-promotion, which probably distorts the information about physician quality. Therefore, the informational value of five-star reviews might be diluted by those positive fraud reviews.

To gain some insights on this issue, we conduct a robustness check by making an extreme assumption that all reviews with an overall rating of five stars are not credible. We first recalculate the ratings after deleting all reviews with five star overall ratings. Note that the overall rating of each review is generated based on the average of *helpfulness* and *knowledge* scores. As a result, those five-star surgeons in the original sample are automatically dropped and a high-rating physician in the new sample has averaged overall ratings of at least 4-star and strictly below 5-star. For example, if a surgeon has two reviews, one with 3.5-star overall rating and one with 5-star overall rating, the averaged overall rating of this surgeon in the new sample would be 3.5-star, instead of 4.25-star as in the original sample. Therefore,

[34] One referee actually identified a possible negative fake review on RateMDs. Later, we found that the fake review was removed from the RateMDs' website.

the rating category of this surgeon would change from high-rating to low-rating surgeon in the new sample. Although this assumption for positive fake reviews is quite extreme, the test may nevertheless offer a glimpse of possible existence and impact of positive fake reviews.

The estimation results are reported in Column (6) and show some interesting information. First, and most importantly, after excluding those five-star reviews, the coefficient of *high-rating* in the linear probability model is -0.09 and remains significant. The magnitude of this coefficient becomes significantly larger than that estimated by our main sample. Second, we find that high-rating surgeons performed better than no-rating surgeons in this test. Hence, it seems possible that positive fake reviews might have added noise to the high ratings, suggesting the co-existence of information and misinformation in online physician reviews.

5.4 Unobserved Surgeon Heterogeneity

According to the data at RateMDs, cardiac surgeons each received an average of 1.7 ratings from 2004 to 2013. This indicates that the rating status does not change much over time for most surgeons. Hence, including surgeon fixed effects is problematic in our setting because of the lack of variation in ratings. However, this raises the concern of unobserved surgeon heterogeneity that affects both online surgeon ratings and patient outcomes.

The random effect (RE) model is an alternative approach of controlling for unobserved heterogeneity when this heterogeneity is constant over time. Its underlying assumption is that the unobserved individual specific effects are uncorrelated with the independent variables. Hence, it is important for us to validate this assumption in our setting. The independent variables in Equation (1) include patient characteristics and surgeon characteristics (e.g., rating status, surgeon personal characteristics and their affiliated hospital characteristics). Below we discuss if the surgeon specific effects are uncorrelated with these independent variables, especially the rating variables of our primary interests. First, in our main specification, we have assumed that ED patients are randomly matched with surgeons, which rules out patient selection bias and suggests that unobserved surgeon specific effects should be uncorrelated with patient characteristics on the right hand of Equation (1). Second, some unobserved characteristics of the hospital where a surgeon conducts a surgery might affect a patient's health outcome and correlate with surgeon characteristics. Concerns resulting from this type of omitted variables could be addressed by including hospital fixed effects. Third, an RE model is appropriate for data of many products that are "drawn from a large population" of products in different categories (Greene, 1990, p.485). In our setting, the surgeons that we studied are the entire Florida cardiac surgeon population. We notice that many structural models using random effects in the industrial organization or marketing literature are justified for the same reason (e.g., Rysman, 2004, Mitra and Golder 2006, Fan, 2013). Therefore, we believe that an RE model is worthwhile trying and reporting for statistical efficiency.

Columns (7) through (10) of Table 4 report the results of the RE model. Column (7) includes the surgeon random effects in addition to patient characteristics, surgeon characteristics, and hospital characteristics. Column (8) replaces hospital characteristics with hospital fixed effects in the RE model. Column (9) and (10) are similar to Column (7) and (8) but with a logit specification. The results are qualitatively similar as those without controlling random effects, possibly because we have included many control variables in our baseline model and the ED setting is assumed to be quasi random.

In summary, we find that surgeons with high ratings are associated with low in-hospital mortality rates compared to those with low ratings based on the assumption of no selections in the ED cases. Moreover, we surprisingly find that surgeons without ratings perform no worse than those surgeons with high ratings. The results are quantitatively and qualitatively similar across various alternative specifications.

6. The Two-Stage Approach

Most of patients do not receive CABG immediately after arriving at a hospital ED because their life is all about time, but preparation for a CABG in an operating room takes at least one hour.[35] Instead, physicians usually perform aspirational thrombectomy (sucking clot out from the vessel) to quickly establish reperfusion (flow) for these ED patients first and recommend to perform CABG later in more stable conditions if possible. In the short window between boarding in hospitals and undergoing CABG, hospital administrators assign a surgeon to take care of a given patient. Naturally, one may be concerned if hospital administrators or surgeons have incentives to select patients of certain severity levels in a certain way so as to improve their performance, measured as mortality rate, shown on the hospital or physician report cards. [36] For example, in a self-managed system, ED physicians may use their private information to assign patients (Chan, 2016). To alleviate the concerns of non-random matching based on risk factors unobservable to researchers, we propose and estimate a two-stage model as a robustness check in this section.

6.1 Patient–Surgeon Matching Model

The major concern in the surgeon performance equation (1) is the systematic correlation between online ratings and indicators of patient health and surgery intensity. In other words, the error term ε may be correlated with the surgeon rating categories R due to possibly non-random matchings between surgeons and patients. To account for this, we adapt the framework of Mroz (1999), which has been used in many settings, including car inspection stations (Hubbard 1998), daycare centers (Blau and Hagy 1998) and schools (Cameron and Taber 2004).

[35] http://patient.info/doctor/acute-myocardial-infarction-management

[36] See http://www.medicare.gov/hospitalcompare/search.html.

We model a provider's expected latent utility from assigning a surgeon in rating category *r* to treat patient i as

$$U_{ir} = \alpha_{1r} \cdot V_i + \alpha_{2r} \cdot E_i + \delta_{ir}$$

$$R_{ij} = \begin{cases} S \text{ (high ratings)} & \text{if } \mathrm{U}_{iS} = \max\{U_{iS}, U_{iM,} U_{iN}\} \\ M \text{ (low ratings)} & \text{if } \mathrm{U}_{iM} = \max\{U_{iS}, U_{iM,} U_{iN}\} \\ N \text{ (no ratings)} & \textit{Otherwise} \end{cases} \tag{2}$$

where V_i refers to a patient's observable health characteristics that affect both the matching and surgical outcomes, including age, race, risk factors, as well as a constant term. E_i refers to a set of exogenous variables, such as insurance types and median income at the zip code level, that may affect the assignment between patients and surgeons but which are not directly correlated to the surgical outcomes. For simplicity of notation, we denote the two sets of patient characteristics as $Z_i = [V, E]$. Finally, the error term δ_{ir} includes additional patient risk factor that are observed by medical providers but unobservable to researchers. The identification strategy of this two-stage model follows Hubbard (1998).

6.2 Error Structures

The error term in the patient–surgeon matching model has the same unobserved component as the error term in the surgeon performance model. This allows us to incorporate the unobserved patient characteristics into Equation (1) and estimate the parameters explaining the non-random matching between surgeons and patients. Based on the framework suggested by Mroz (1999), we impose the following error structures for Equations (1) and (2), respectively:

$$\varepsilon_i = \rho \cdot \eta_i + \varsigma_i \tag{3}$$

$$\delta_{ir} = b_r \cdot \eta_i + \xi_{ir} \tag{4}$$

where η_i captures the same unobservable patient health characteristics as in the performance model and the matching model and affects both the patient's health hazard and the patient–surgeon matching outcome. For simplicity, we use *i* as the subscript, since the unit of observation in our sample is the patient. The term ς_i in Equation (3) is a random shock that affects a patient's survival but is independent of patient–surgeon matching. We assume ς_i follows a standard logit distribution. The performance model (1) and the matching model (2), linked by the error structure, can be jointly estimated via full maximum likelihood estimation (FMLE). We describe the identification of parameters and the estimation method for this two-stage model in the Online Appendix.

6.3 Are Low-rating Surgeons Associated with High Mortality Rates?

Table 5 reports the estimation results of this two-stage model. Column (1) reports the estimation results of the surgeon performance equation (i.e., Equation (1)) and Columns (2) and (3) report the estimation results of the patient-surgeon matching equation (i.e., Equation (2)). The coefficient of *high-rating* surgeons is negative and significant (p<0.01). A counterfactual analysis suggests that, after the selection bias is corrected for, the survival rate for a representative patient treated by a *high-rating* surgeon increases by 27.2 percentage point more than by a *low-rating* surgeon.

The coefficient of *no-rating* surgeons is also negative and significant, suggesting that all else being equal, patients treated by *no-rating* surgeons have better health outcomes than those treated by surgeons with low ratings. In the patient–surgeon matching model, the parameter for unobserved patient health characteristics η is positively associated with mortality rates, suggesting the possible existence of unobserved risk factors. The coefficients are positive and significant for both b_N and b_S, together suggesting that there exists some degree of selection based on unobserved risk factors which reduces the chance of a high-risk patient being treated by surgeons with ratings below four stars.

In summary, we again find that patients treated by low-rating surgeons have a higher mortality rate than those treated by surgeons with high ratings or without rating. This result supports the validity hypothesis that online ratings for physicians do contain some information reflecting their medical quality. The results are robust when we change the number of points of support.[37]

7. Discussion and Conclusion

7.1 Sound of Silence

As we previously discussed, the theory of “sound of silence” (Dellarocas and Wood, 2008) might not necessarily apply in the context of online physician rating due to the lack of bidirectional feedback mechanism. However, the lack of good ratings could still signal lack of highly satisfied patients, which might be related to the medical quality of the physician. Gao et al (2011) uses the primary care physician (PCP) as the study subjects. In their paper, they report that 696 out of 1425 PCPs were rated before 2010, which account for almost 50% of the physician sample. Based on this sample, Gao et al. (2011) shows that PCPs perceived by patients as of low quality are most unlikely to receive online ratings, while good doctors are as likely to be rated online as average doctors are, an evidence supporting the “sound of silence” theory.

Different from Gao et al (2011), our study shows that surgeons with no ratings deliver better patient outcomes than those with low ratings and no worse than those with high ratings. We suspect that

[37] The results in Table 5 are estimated by using two points of support (K=2). For robustness checks, we increase the number of points of support to 4 and 6. The results are available in Table A5 of the Online Appendix.

three possible reasons may lead to the different results. First, the percentage of rated cardiac surgeons is relatively low. Different from PCPs, a specialty with the highest percentage of rated doctors (Gao et al. 2012), rating cardiac surgeons online is still in its infancy. In our sample, only 24% of the surgeons have been rated till 2013. It might be too early to apply the theory of "silence of sound" to the specialty of cardiac surgeons and conclude that more than three quarters of unrated surgeons are low quality surgeons just because they do not have a rating at this early stage.

Figure A3 provides supportive evidence for our argument. If the theory of "sound of silence" does hold in the setting of cardiac surgeons, we would expect the average ratings to go down as more unrated doctors become rated. Figure A3 shows that the average ratings of cardiac surgeons over time are quite stable instead of having a downward sloping trend. We regress the average quarterly ratings over a constant and a quarterly linear trend. The coefficient of the linear trend is -0.004 with p-value 0.583, suggesting that there is no downward sloping trend in the average ratings of cardiac surgeons from 2010 to 2013.

Second, physicians provide localized services. The frequency of physician visits may be the driving force behind the different findings. For example, patients may develop personal relationships with their local PCPs since they have to visit them from time to time and may thus hesitate to report unpleasant experiences online. In other words, the mechanism behind the "sound of silence" might be at work in the PCP setting because tension between a patient and his or her PCPs resulting from a negative rating could have the same effect as that of a retaliating rating in bidirectional platform such as eBay. On the contrary, a patient may see a particular cardiac surgeon only once or twice in his or her life for open heart surgery and is less embarrassed to complain if they were unhappy about the services provided. Therefore, the absence of ratings for cardiac surgeons may simply reflect people's lack of willingness to post online reviews rather than to signal a surgeon's poor quality.

To better understand the quality differences between cardiac surgeons with low ratings and those without ratings, in Figure A4 we construct the pseudo-ratings for those doctors without ratings, based on surgeon characteristics, patient composition, and so on. The pseudo rating idea is motivated by the propensity score matching method. The rationale is that an unrated doctor would have the same ratings as his matched doctor's (with the same characteristics) had he been rated. Based on this assumption, we use the rated doctors' information to construct the pseudo ratings for those unrated doctors.[38] Then we compare the distribution of the pseudo-ratings for those without ratings to the real ratings for low-rated surgeons. The figure suggests that surgeons without ratings would have received higher ratings than those with low ratings had they been rated. Besides, we regress the (pseudo) ratings over a constant and a rating

[38] The variables we used to construct pseudo ratings include patient volume, percentage of private-paying consumers, percentage of black patients, percentage of female patients, average patient age, average patient local income, patient education levels, physician experience, graduation schools, board information, honor and award, malpractice and crime records.

dummy indicating 1 for no ratings and 0 for low ratings. The coefficient is 1.424 (p<0.001), both positive and significant. Overall, consistent with our estimation results, the figure suggests that cardiac surgeons without ratings may have better quality than those with low ratings.

Third, although RateMDs takes actions to reduce the fraud reviews, we suspect that positive fraud reviews are less likely to be detected than negative fraud reviews. A physician can and will have enough incentives to appeal to RateMDs if he or she receives negative fraud reviews. On the contrary, positive fraud reviews are more difficult to be detected. Although we do not think many surgeons would actively engage in fraud reviews, they probably have the incentives to do so. If that is the case, then the informational value of high-ratings would be diluted and underestimated. This might explain the finding that there is no statistical difference between high-rating surgeons and surgeons without rating. Our test about positive fraud review in Session 5.3 has provided suggestive evidence for this argument.

To summarize, we believe that the theory of "sound of silence" should apply to physicians in certain specialties when a large portion of them get rated,[39] but it could be misleading to directly apply it to all specialties at such an early stage when the majority of physicians have not yet been rated.

7.2 Generalization of our findings

This study focuses on the performance of cardiac surgeons on CABG surgeries. Our study subjects have two important features which differ from other types of physician such as PCPs, OB doctors and psychiatrists. First, treatment outcomes are relatively observable to patients and their family members. In our setting, patients or their family members can directly observe how well the surgeries were performed, which helps them evaluate physicians' medical quality. Although they may not have medical training, they are very likely to know the severity of the patient's disease very well, which help them provide adequate reviews about the cardiac surgeons. Second, unlike PCPs, surgeons do not frequently see their patients. The relatively weak patient-physician bond will not prevent patients and their family members from sharing their experience and expressing their opinions, whether positive or negative. Therefore, our findings may be generalized to those physician types sharing these two distinct features, such as orthopedic surgeons who do knee or hip replacements.

7.3 Conclusion

We assess whether online physician ratings are informative about physicians' less observed medical quality using CABG surgeons as the study subjects. Our main findings show that patients treated by surgeons rated four stars or higher have better odds of surviving than those treated by surgeons with lower ratings, after we control for various patient risk factors (observed and unobserved) and surgeon

[39] We conduct a falsification test to determine any gradual changes in the informational value of online physician reviews. To do so, we use reviews accumulated up to the focal year of analysis and repeat our estimation on past patient outcomes for the years up to 2013. Though not reported in a table, the results seem to suggest the transitions that online reviews have become better indicators of physician medical quality as more surgeons become rated over the years.

characteristics. Surprisingly, we find that surgeons without ratings perform better than those rated below four stars and no worse than those rated four stars or higher. These results are robust to various tests.

The current theoretical framework of WOM, especially the “sound of silence” theory, is applicable to search and experience goods in general. Nevertheless, we find that the “sound of silence” theory does not work when rating physician online is still in its infancy and prone to positive fake reviews. While the traditional “sound of silence” theory (Dellarocas and Wood 2008) predicts the “negative” sound of silence, our study suggests that this theoretical framework should incorporate the time dimension when being applied to professional service whose online platforms evolve slowly to mature stages.

Below we discuss some potential limitations and possible extensions of our analysis. First, our results are based on one type of physicians and one medical performance measure. It will be interesting and important to extend the current study to other types of physicians, such as orthopedists who perform knee or hip replacement, and alternative medical performance measures.

Second, our results can benefit from increasing number of reviews and rated physicians. For example, if the vast majority of physicians would have been rated, we might measure surgeons’ online rating statuses using more flexible transformations of the raw rating to investigate different quality implications of ratings at different segments of the scale. At the moment, the large number of physicians who do not have any ratings remain largely a black box to us. Therefore, to be conservative, our findings about quality inference of rated surgeons should not be over-generalized due to the rating data limitation. We are cautiously optimistic that online physician rating system is informative for the patients.

Finally, we acknowledge the reporting bias in online physician ratings. This, along with other pitfalls of online physician reviews, is exactly what motivated us to investigate whether one can still glean quality information from those raw ratings which are directly observed and are often used by consumers. Moreover, this issue of reporting bias clearly raises more questions about online physician reviews. What drives patients to contribute a review? What types of physicians are more likely to be rated? How do online ratings and public reporting affect the physician behavior? How long would it take for actual medical quality to be reflected in the perceptual quality measures from online ratings? These are open but very important questions that need to be answered in order to further our understanding of the behavior of patients and physicians in the social media era.

Despite these limitations, the practical implications of this article are important for patients, physicians, policy makers, and online physician websites. Our findings suggest that patients should avoid physicians with low ratings, at least for cardiac surgeons, a specialty in which treatment outcomes are relatively observable to patients (and their family members). And we believe this conclusion will continue to hold when the platforms evolve to more mature stages because the same mechanism through which quality signal is revealed should continue to work, if not more effectively, as these platforms evolve.

Unlike mandatory disclosure, online ratings reveal subjective assessments of physician quality along many dimensions,[40] which helps avoid distorting physician incentives. For example, online subjective assessments of physician quality could prevent physicians from gaming the report card system by avoiding seriously ill patients, which could be incentivized by an objective quality measure.[41] Policy makers can consider online WOM a good complement to the mandatory disclosure policy while physicians could be less concerned to treat seriously ill patients. Physicians should also learn from the online reviews rather than complaining about negative reviews (Jain 2010). They should also appeal fake negative reviews to protect themselves, which also help improve the information quality of online physician rating platforms.

Our findings are generally encouraging for online physician review platforms by suggesting that patients can treat online physician reviews as a reliable source for obtaining physician information. However, our study also suggests the importance and challenge of weeding out positive fake reviews which dilutes the informational value of high ratings and yet is much more difficult to detect than negative fake reviews. Nevertheless, in the future, with the accumulation of sufficient online physician reviews, information should be able to dominate misinformation and online physician review sites may become extremely valuable to consumers searching for physicians.

In order for this day to come and to see the full potential of this online information channel, online physician review sites and policy makers must work together to better understand the drivers and inhibitors for patients and their family to contribute online physician reviews so as to better design systems and policies to facilitate the generation of these “public goods.” Probably the most important practical implication of our work is to demonstrate that consumer ratings of physicians *can* carry information about their medical quality. Therefore, as a society, we should embrace and help grow this new information channel, rather than demolish and forsake it while it is still in its infancy.

References

Arrow, K. J. 1963. "Uncertainty and the Welfare Economics of Medical-Care." *American Economic Review*, 53(5), 941-73.

[40] For example, the disclosing authority has to select some quality dimensions to release in the report cards. This incurs teaching-to-the-test behavior, as Lu (2012, 2016) documents.

[41] Shunning patients with a poor prognosis may improve the mortality scores on the report card but may offend patients and their family members and result in bad online reviews. Therefore, cherry-picking patients may backfire and hurt a physician’s online ratings.

Berger, J. and K. L. Milkman. 2012. "What Makes Online Content Viral?" *Journal of Marketing Research*, 49(2), 192-205.

Blau, D. M. and A. P. Hagy. 1998. "The Demand for Quality in Child Care." *Journal of Political Economy*, 106(1), 104-146.

Cabral, L. and A. Hortacsu. 2010. "The Dynamics of Seller Reputation: Evidence from Ebay." *Journal of Industrial Economics*, 58(1), 54-78.

Cameron, S. V. and C. Taber. 2004. "Estimation of Educational Borrowing Constraints Using Returns to Schooling." *Journal of Political Economy*, 112(1), 132-182.

Chan D.C. 2016. "Teamwork and Moral Hazard: Evidence from the Emergency Department", *Journal of Political Economy*, 124(3), 734-770

Chevalier, J. A. and D. Mayzlin. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews." *Journal of Marketing Research*, 43(3), 345-354.

Chintagunta, P. K., S. Gopinath and S. Venkataraman. 2010. "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation across Local Markets." *Marketing Science*, 29(5), 944-957.

Clark, J.R. and Huckman, R.S. 2012. "Broadening Focus: Spillovers, Complementarities, and Specialization in the Hospital Industry." *Management Science*, 58(4), 708-722.

Dai, W., G.Z. Jin, J. Lee and M. Luca. 2013. "Optimal Aggregation of Consumer Ratings: An Application to Yelp.com," *Working Paper.*

Dellarocas, C. and C. A. Wood. 2008. "The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias." *Management Science*, 54(3), 460-76.

Dhar, V. and E. A. Chang. 2009. "Does Chatter Matter? The Impact of User-Generated Content on Music Sales." *Journal of Interactive Marketing*, 23(4), 300-307.

Dranove, D. 2008. *Code Red: An Economist Explains How to Revive the Healthcare System without Destroying It*. Princeton University Press.

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

Duan, W. J., B. Gu and A. B. Whinston. 2008. "Do Online Reviews Matter? - an Empirical Investigation of Panel Data." *Decision Support Systems*, 45(4), 1007-1016.

Ellimoottil, C., A. Hart, K. Greco, M. L. Quek and A. Farooq. 2013. "Online Reviews of 500 Urologists." *Journal of Urology*, 189(6), 2269-2273.

Emmert, M. and F. Meier. 2013. "An Analysis of Online Evaluations on a Physician Rating Website: Evidence from a German Public Reporting Instrument." *Journal of Medical Internet Research*, 15(8).

Emmert, M., F. Meier, F. Pisch and U. Sander. 2013a. "Physician Choice Making and Characteristics Associated with Using Physician-Rating Websites: Cross-Sectional Study." *Journal of Medical Internet Research*, 15(8).

Emmert, M., U. Sander and F. Pisch. 2013b. "Eight Questions About Physician-Rating Websites: A Systematic Review." *Journal of Medical Internet Research*, 15(2).

Fan, Y. 2013. "Ownership Consolidation and Product Characteristics: A Study of the US Daily Newspaper Market." *American Economic Review* 103(5): 1598–1628.

Gao, G., J. S. McCullough, R. Agarwal and A. K. Jha. 2012. "A Changing Landscape of Physician Quality Reporting: Analysis of Patients' Online Ratings of Their Physicians over a 5-Year Period." *Journal of Medical Internet Research*, 14(1).

Gao, G., B. Greenwood, J. McCullough and R. Agarwal. 2011. "A Digital Soapbox? The Information Value of Online Physician Ratings," *Working Paper.*

Godes, D. and J. C. Silva. 2012. "Sequential and Temporal Dynamics of Online Opinion." *Marketing Science*, 31(3), 448-473.

Greaves, F., U. J. Pape, H. Lee, D. M. Smith, A. Darzi, A. Majeed and C. Millett. 2012. "Patients' Ratings of Family Physician Practices on the Internet: Usage and Associations with Conventional Measures of Quality in the English National Health Service." *Journal of Medical Internet Research*, 14(5).

Greene, W. H. 1990. "Econometric Analysis". Macmillan, New York.

Hannan, E. L., Kilburn, H., Jr., Racz, M., Shields, E., and Chassin, M. R. (1994). Improving the outcomes of coronary artery bypass surgery in New York State. *Journal of American Medical Association*, 271(10), 761-766.

Hartz, A. J., Pulido, J. S., and Kuhn, E. M. (1997). "Are the best coronary artery bypass surgeons identified by physician surveys?" *American Journal of Public Health*, 87(10), 1645-1648.

Hubbard, T. N. 1998. "An Empirical Examination of Moral Hazard in the Vehicle Inspection Market." *Rand Journal of Economics*, 29(2), 406-426.

Huckman, R. S. and G. P. Pisano. 2006. "The Firm Specificity of Individual Performance: Evidence from Cardiac Surgery." *Management Science*, 52(4), 473-488.

Hurwitz J.E, Lee J.A, Lopiano K.K, Mckinley S.A, Keesling J. 2014. "A Flexible Simulation Platform to Quantify and Manage Emergency Department Crowding." *BMC Medical Informatics and Decision Making*, 14(50)

Jain, S. 2010. "Googling Ourselves - What Physicians Can Learn from Online Rating Sites." *New England Journal of Medicine*, 362(1), 6-7.

KC, D.S. and C. Terwiesch. 2011. "The Effect of Focus on Performance: Evidence from California Hospitals." *Management Science*, 57(11), 1897-1912.

Lagu, T., N. S. Hannon, M. B. Rothberg and P. K. Lindenauer. 2010. "Patients' Evaluations of Health Care Providers in the Era of Social Networking: An Analysis of Physician-Rating Websites." *Journal of General Internal Medicine*, 25(9), 942-946.

Liu, Y. 2006. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue." *Journal of Marketing*, 70(3), 74-89.

Lovett, M. J., R. Peres and R. Shachar. 2013. "On Brands and Word of Mouth." *Journal of Marketing Research*, 50(4), 427-444.

Lu, S.F. "Information Disclosure, Multitasking and Product Quality: Evidence from Nursing Homes," *Journal of Economics & Management Strategy*. 2012. 21(3), 673-705,

Lu, S.F. "The Role of Donations in Quality Disclosure: Evidence from Nonprofit Nursing Homes" *American Journal of Health Economics,* 2016, 2(4),431-462

Luca, M. 2011. "Reviews, Reputation, and Revenue: The Case of Yelp.Com", *Working Paper.*

Luca, M. and S. Vats. 2013. "Digitalizing Doctor Demand: The Impact of Online Reviews on Doctor Choice," *Working Paper.*

Luca, M., and G. Zervas. 2016. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." *Management Science*, Forthcoming

Mitra, D and P. N. Golder (2006), "How Does Objective Quality Affect Perceived Quality? Short-Term Effects, Long-Term Effects, and Asymmetries," *Marketing Science*, 25(3), 230-247.

Mroz, T. A. 1999. "Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome." *Journal of Econometrics*, 92(2), 233-274.

Mudambi, S. M. and D. Schuff. 2010. "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.Com." *MIS Quarterly*, 34(1), 185-200.

Rui, H. X., Y. Z. Liu and A. Whinston. 2013. "Whose and What Chatter Matters? The Effect of Tweets on Movie Sales." *Decision Support Systems*, 55(4), 863-870.

Rysman, M., 2004. "Competition between Networks: A Study of the Market for Yellow Pages." *Review of Economic Studies*, 71(2), 483–512.

Segal, J., M. Sacopulos, V. Sheets, I. Thurston, K. Brooks and R. Puccia. 2012. "Online Doctor Reviews: Do They Track Surgeon Volume, a Proxy for Quality of Care?" *Journal of Medical Internet Research*, 14(2).

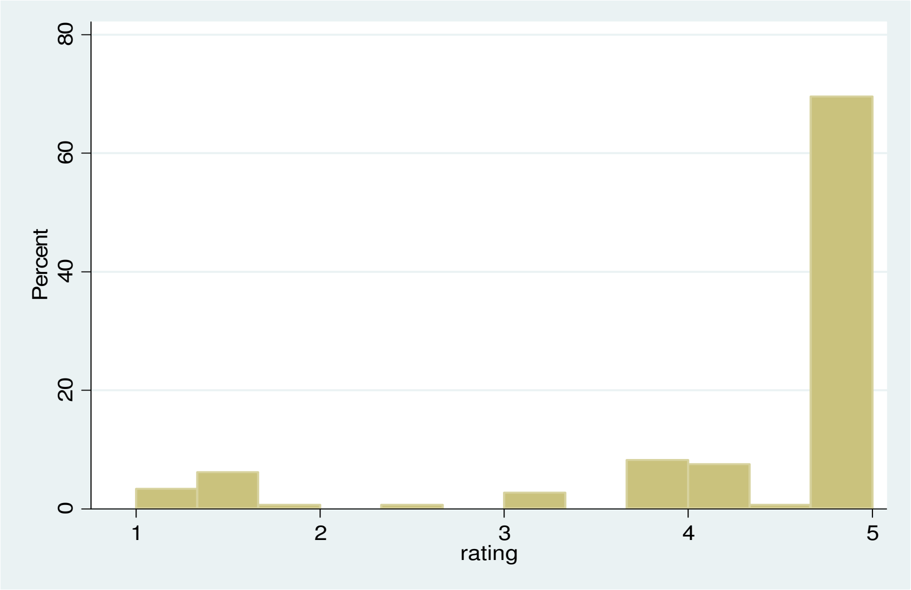
Serruys, P. W., M. C. Morice, A. P. Kappetein, A. Colombo, D. R. Holmes, M. J. Mack, E. Stahle, T. E. Feldman, M. van den Brand, E. J. Bass, et al. 2009. "Percutaneous Coronary Intervention Versus Coronary-Artery Bypass Grafting for Severe Coronary Artery Disease." *New England Journal of Medicine*, 360, 961-972.

Vermeulen, I. E. and D. Seegers. 2009. "Tried and Tested: The Impact of Online Hotel Reviews on Consumer Consideration." *Tourism Management*, 30(1), 123-127.

Yang, S., M. T. Hu, R. S. Winer, H. Assael and X. H. Chen. 2012. "An Empirical Study of Word-of-Mouth Generation and Consumption." *Marketing Science*, 31(6), 952-963.

Wooldridge, J. 2003. “Cluster-Sample Methods in Applied Econometrics”, *American Economic Review* 93(2),133-138.

Figure 1: The Distribution of Surgeon Ratings



Notes: The sample includes 624 quarterly ratings for surgeons from September 2012 to August 2013.

Table 1: Summary Statistics

This table reports summary statistics and definitions of key variables used in this study. The sample includes 77 hospitals, 200 surgeons, 624 quarterly ratings and 3,819 CABG surgeries conducted by cardiac surgeons on patients transferred from ED in Florida in 2013.

Variable	Obs	Mean	SD	Definition
Performance Measure (at the patient level)				
In-hospital Mortality (%)	3819	2.62	15.97	1 if a patient died before being discharged
Ratings (at the surgeon-quarter level)				
High-rating	624	0.18	0.39	1 if a surgeon is rated at least four stars
Low-rating	624	0.05	0.22	1 if a surgeon is rated below four stars
No-ratings	624	0.77	0.42	1 if a surgeon has no reviews
CABG volume	624	16.1	10.5	the number of CABG surgeries conducted last quarter
Surgeon Characteristics (at the surgeon level)				
Experience	200	18.0	9.9	number of years since a surgeon graduated from medical school
Attending elite schools (%)	200	0.5	7.1	1 if a surgeon attended an elite medical school
Malpractice Claim (%)	200	1.0	10.0	1 if a surgeon involved in a malpractice case
Patient Characteristics (at the patient level)				
Age	3819	65.5	11.2	age of a patient
White (%)	3819	77.1	42.0	1 if a patient is white
Black (%)	3819	9.9	29.9	1 if a patient is black
Other races (%)	3819	12.9	33.6	1 if a patient is neither white nor black
Female	3819	0.3	0.5	1 if a patient is female
Medicare	3819	0.5	0.5	1 if a patient is covered by Medicare
Medicaid	3819	0.1	0.3	1 if a patient is covered by Medicaid
Private insurance	3819	0.2	0.4	1 if a patient is covered by private insurance plans
Other insurance types	3819	0.2	0.4	1 if the patient is not covered by Medicare, Medicaid and private
Log of income	3819	10.6	1.3	median household income at the zip code level
Travel time	3819	20.2	16.9	minutes used from a patient's home to a given hospital
CABG	3819	0.02	0.13	1 if patient has history of CABG or valve surgery
Heart	3819	0.26	0.44	1 if patient has heart failure
Vascular	3819	0.12	0.32	1 if patient has history of peripheral vascular disease
Kidney	3819	0.17	0.38	1 if patient has chronic kidney disease
AMI	3819	0.57	0.50	1 if patient has AMI initial episode of care
Obesity	3819	0.04	0.19	1 if patient has morbid obesity
Nutrition	3819	0.03	0.17	1 if patient has malnutrition
Valve	3819	0.06	0.23	1 if the procedure is CABG with Valve
Hvd	3819	0.02	0.15	1 if a patient has heart valve disease
Liver	3819	0.02	0.15	1 if a patient has liver disease
type1	3819	0.16	0.37	1 if coronary bypass of one coronary artery
type2	3819	0.33	0.47	1 if coronary bypass of two coronary arteries
type3	3819	0.28	0.45	1 if coronary bypass of three coronary arteries
type4	3819	0.12	0.32	1 if coronary bypass of four or more coronary arteries
type5	3819	0.11	0.32	1 if single internal mammary-coronary artery bypass
Hospital Characteristics (at the hospital level)				
Hospital Rankings	77	14.79	1.56	measured by hospital mortality rate (in percentage) due to heart attack
CICU availability	77	0.7	0.4	1 if a cardiac intensive care unit is available in a hospital
Beds	77	490	389	the number of beds in a hospital

Table 2: General Information about Ratings and Surgeons across Categories

Rating information and surgeon characteristics based on CABG surgeries performed on ED patients in 2013

Panel A: Rating Information (unit of observation: surgeon-quarter)

	Number of Observations	Percentage (%)	Total Number of Surgeries	Average Number of ED Surgeries	Average Number of Surgeries (ED and non-ED)
High-rating Surgeons	113	18%	707	6.3	17.3
Low-rating Surgeons	32	5%	204	6.4	16.4
Surgeons without Ratings	479	77%	2,908	6.1	15.7
Total	624	100%	3,819	6.1	16.1

Panel B: Surgeon Characteristics (unit of observation is surgeon and rating is based on the last quarter of 2013)

	Number of Observations	Mortality rate (%)	Average Number of Years since graduation	Number of Surgeons with Malpractice Claim	Number of Surgeons Attending elite schools
High-rating Surgeons	33	2.9	19.6	1	0
Low-rating Surgeons	11	6.3	19.2	0	0
Surgeons without Ratings	156	2.3	17.5	1	1
Total	200	2.6	18	2	1

Table 3: A Diagnostic Test of Selection Based on Patient Preoperative Risk

	Predicted Preoperative Mortality Rates			
	Risk Factors Only		Risk Factors +Socioeconomics	
VARIABLES	(1)	(2)	(3)	(4)
Ratings (continuous)	0.002		0.0003	
	(0.001)		(0.001)	
High-rating		0.004		0.004
		(0.003)		(0.003)
No-rating		0.003		0.002
		(0.002)		(0.002)
Hospital Dummy	Y	Y	Y	Y
Prob > F (High-rating - No-rating)		0.316		0.264
Observations	911	3,819	911	3,819
R-squared	0.087	0.057	0.077	0.055

Standard errors clustered by surgeon.*** p<0.01, ** p<0.05, * p<0.1

*** p<0.01, ** p<0.05, * p<0.1

Table 4: Are High-rating Surgeons Associated with Lower Mortality Rates?

Dependent Variable: Mortality	Main				Robustness		Random Effects			
Mortality	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
High-rating	-0.034*	-0.035*	-0.881**	-0.900**	-0.042**	-0.058**	-0.035***	-0.047***	-0.900**	-1.322**
	(0.018)	(0.018)	(0.419)	(0.409)	(0.019)	(0.018)	(0.013)	(0.018)	(0.400)	(0.603)
No-rating	-0.035**	-0.035**	-0.932**	-0.967***	-0.041**	-0.029***	-0.035***	-0.042**	-0.967***	-1.161**
	(0.017)	(0.017)	(0.365)	(0.353)	(0.018)	(0.015)	(0.012)	(0.016)	(0.345)	(0.554)
Patient characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Hospital characteristics	Y	Y	Y	Y	Y	Y	Y	N	Y	N
Surgeon characteristics	N	Y	N	Y	Y	Y	Y	Y	Y	Y
Quarter dummies	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
F test: High Rating - No Rating	0.00	0.00	0.05	0.07	-0.01	-0.03***	0.00	-0.01	0.07	-0.16
F test: Prob>chi 2	0.96	0.94	0.86	0.82	0.91	<0.01	0.94	0.55	0.81	0.65
Observations	3819	3819	3734	3674	3270	3154	3819	3819	3819	3819

Standard errors are clustered by surgeon for columns (1) through (6). *** p<0.01, ** p<0.05, * p<0.1

Notes: (1) We do not cluster the standard errors for the random effect model since the cluster is at the single level (Wooldridge, 2003).
(2) Column (8) and (10) replace the hospital characteristics with hospital fixed effects.
(4) Patient characteristics includes age, race, gender, CABG, travel time, comorbidities and so on. Hospital characteristics includes hospital ranking, beds and CICU availability. Surgeon characteristics includes experience, education and malpractice information.

Table 5: Robustness Checks using the Two-Stage Model

	Performance	Matching	
VARIABLES	Mortality rates	No-rating	High-rating
	(1)	(2)	(3)
High-rating	-3.206***		
	(0.993)		
No-rating	-2.835***		
	(1.023)		
η	7.294***		
	(1.330)		
b_N		0.107***	
		(0.040)	
b_S			0.166**
			(0.072)

Standard errors are clustered by surgeon. *** p<0.01, ** p<0.05, * p<0.1. Control variables include patient characteristics, surgeon characteristics and relevant quadratic terms, hospital characteristics, and quarter dummies.

Online Appendix for “Can We Trust Online Physician Ratings? Evidence from Cardiac Surgeons in Florida”

Appendix A: Maximum Likelihood Estimation of the Two-Stage Model

Assume that the probability of observing death the in hospital after a CABG surgery is

$$Pr(Y_i = 1|X_i, \eta_i) = Pr(-\varsigma_i \leq X_i'\beta + \eta_i) = \frac{e^{X_i'\beta+\eta_i}}{1+e^{X_i'\beta+\eta_i}},$$

where the factor loading ρ is normalized to 1 for identification purpose in the surgeon performance model.

Assuming the error term ς that affects a patient’s health hazard, conditional on X, R, and η, is independent of the error term ξ that affects the patient–surgeon matching outcome, conditional on Z and η, we have

$$\Pr(R = r, Y = y|X, Z, \eta) = \Pr(Y = y|X, R = r, \eta) \cdot \Pr(R = r|Z, \eta),$$

which implies the following log-likelihood function for estimation:

$$\log L = \sum_{i=1}^{I} \sum_{r\in \mathrm{A}} \sum_{y\in\{0,1\}} \mathbf{1}_{R_i=r, Y_i=y} \log\Pr(r, y|X, Z, \eta)$$

where I is the number of patients, A is the set of surgeon categories, and Y is the surgical outcome. The index function equals one when R = r and Y = y.

The random shock ξ_i is assumed to follow the standard Type-1 extreme value distribution. Standard argument then leads to the following conditional probability that a patient will be matched to a surgeon of category r:

$$\Pr(R_i = r|Z_i, \eta_i) = \Pr(U_{ir} > U_{il}, \forall l \neq r)$$

$$= \Pr\left(\xi_{il} - \xi_{\mathrm{ir}} < Z_i'(\alpha_r - \alpha_l) + \eta_i(b_r - b_l), \forall l \neq r\right)$$

$$= \frac{e^{Z_i'\alpha_r + b_r\eta_i}}{\sum_{r\in A} e^{Z_i'\alpha_r + b_r\eta_i}}. \qquad (8)$$

We use the rating category M, surgeons with ratings below four stars, as the benchmark and assume $\alpha_M = b_M = 0$ so that the model is identified.

The parameters of this model include those that determine the distribution of the discrete factor η as well as those associated with the right-hand-side variables in each of the two stages. Following the literature, we assume that the unobserved factor η follows a discrete distribution with K points of support and p_k, $k = 1, \cdots, K$ are the corresponding probabilities. To obtain the unconditional probability, we integrate over the discrete distribution of η:

$$\Pr(R = r, Y = y|X, Z) = \sum_{k=1}^{K} p_k \cdot \Pr(Y|X, \eta_k) \cdot \Pr(R = r|Z, \eta_k) \ .$$

(9)

The log-likelihood function, by substituting (7) and (8) into (9), can be written as

$$\log L = \sum_{i=1}^{I} \sum_{r \in \mathrm{A}} \mathbf{1}_{R_i=r} \{ y_i \log(\sum_{k=1}^{K} p_k (\frac{e^{X_i'\beta+\eta_k}}{1+e^{X_i'\beta+\eta_k}} \cdot \frac{e^{Z_i'\alpha_r+b_r\eta_k}}{1+\sum_{r\in A_0} e^{Z_i'\alpha_r+b_r\eta_k}})) \\ +(1-y_i)\log(\sum_{k=1}^{K} p_k (\frac{1}{1+e^{X_i'\beta+\eta_k}} \cdot \frac{e^{Z_i'\alpha_r+b_r\eta_k}}{1+\sum_{r\in A_0} e^{Z_i'\alpha_r+b_r\eta_k}}))\} \ .$$

(10)

We use the full information maximum likelihood method (FMLE) to estimate the parameters of this full model $\{\beta, (\alpha_r, b_r)_{r\in A_0}, (p_k, \eta_k)_{k=1}^{K}\}$. Of primary interest, are the coefficients of rating categories R, i.e. β_1 in Equation (1), which indicate whether online physician reviews reflect surgeons' medical skills or not.

We normalize $E(\eta)=0$ because one cannot separately identify the location of the discrete distribution and the constant term in the surgeon performance equation. To reduce the computational burden, we assume a symmetric distribution with $K = 2$ and $\eta = \eta_1 = -\eta_2$ to better identify parameters of interest [1]. Hence, the log-likelihood function is as follows:

$$\mathrm{Log}L = \sum_{i=1}^{I} \sum_{r \in \mathrm{A}} \mathbf{1}_{R_i=r} \{ y_i \log(p (\frac{e^{X_i'\beta+\eta}}{1+e^{X_i'\beta+\eta}} \cdot \frac{e^{Z_i'\alpha_r+b_r\eta}}{1+\sum_{r\in A_0} e^{Z_i'\alpha_r+b_r\eta}}) + (1-p)(\frac{e^{X_i'\beta-\eta}}{1+e^{X_i'\beta-\eta}} \\ \cdot \frac{e^{Z_i'\alpha_r-b_r\eta}}{1+\sum_{r\in A_0} e^{Z_i'\alpha_r-b_r\eta}})) + (1-y_i)\log(p (\frac{1}{1+e^{X_i'\beta+\eta}} \cdot \frac{e^{Z_i'\alpha_r+b_r\eta}}{1+\sum_{r\in A_0} e^{Z_i'\alpha_r+b_r\eta}}) + (1 \\ - p)(\frac{1}{1+e^{X_i'\beta-\eta}} \cdot \frac{e^{Z_i'\alpha_r-b_r\eta}}{1+\sum_{r\in A_0} e^{Z_i'\alpha_r-b_r\eta}}))\}$$

The model parameters $\theta = (\beta, (\alpha_r, b_r)_{r\in \mathrm{A}_0}, \eta, p)$ can be identified after we treat the category M as the base outcome and normalize the corresponding loading factor $b_M = 0$. We report the results using and $p_1=p_2=0.5$. The point estimates and standard errors on the parameters of interest do not change when this restriction is relaxed, but near-multicollinearity may make the standard errors on the constants and the distributional parameters very large.

[1] The underlying assumption is that the qualitative results do not change when increasing K, which has been verified by Mroz (1999) by simulation and is assumed in Hubbard (1998).

Given the complexity of the model, we conduct a simulation by generating a dataset and estimate it again to see if the model can recover the simulation parameters. We use the mean and standard deviation of the variables in our sample to generate a pseudo dataset. The simulation result suggests that our estimation procedure can recover the coefficients with reasonable accuracy.

One might be concerned that the death of a patient is a rare event, even for CABG procedures. According to King and Zeng (2001), the estimation of a choice model for rare event can be biased. However, when the sample size is large (over a few thousand), the issue of rare event is less concerned. To alleviate this concern, we aggregate data from 2012 to 2013 and estimate our model. Our results remain robust. [2]

[2] Both the simulation results and the large sample results are available upon request. Simulation results for similar models can also be found in Mroz (1999).

Appendix B: Additional Figures and Tables

Figure A1: Accumulative Number of Reviews and Physicians Reviewed in Florida on RateMDs.com

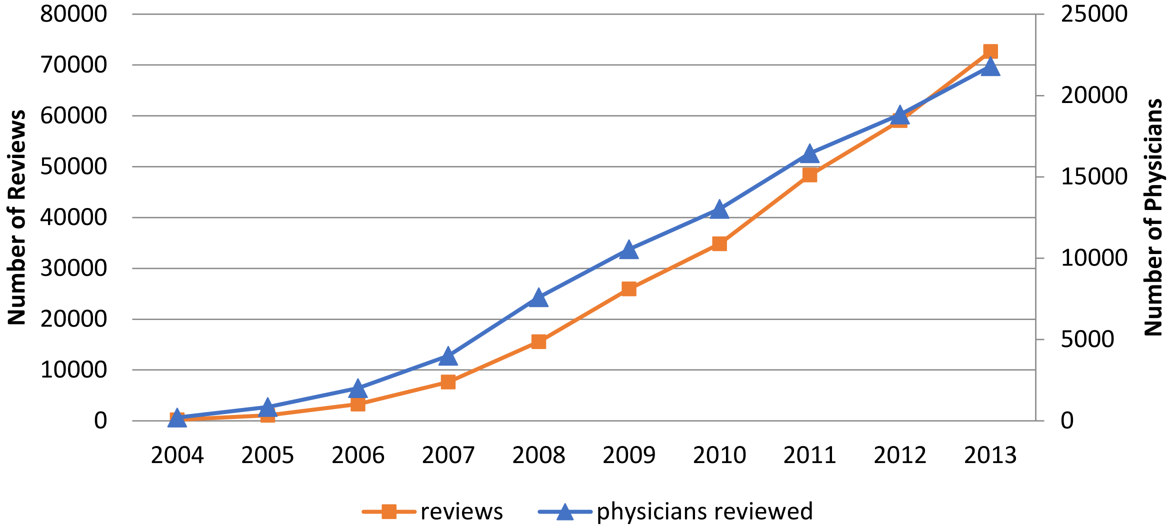


Figure A2: Word cloud from the text of patient reviews (words stemmed)

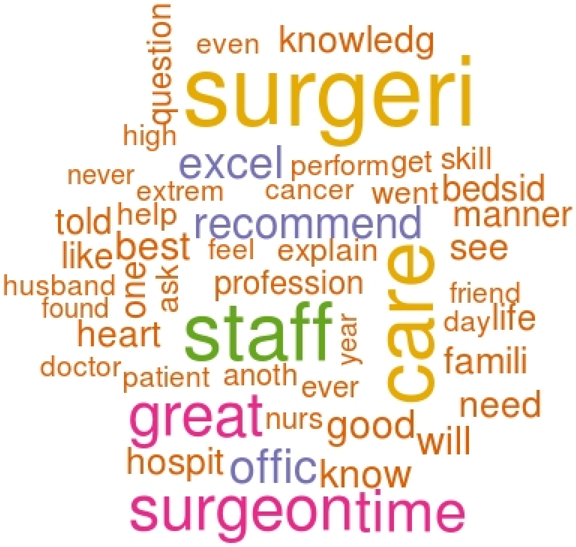


Figure A3: Average Ratings of Cardiac Surgeons

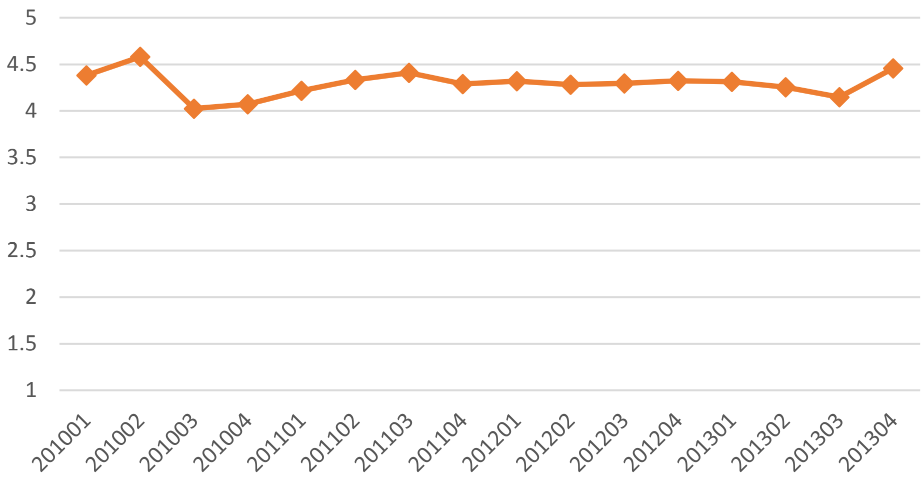


Figure A4: Quality Differences between Surgeons with No Ratings and Low Ratings

The figure compares quality differences between cardiac surgeons with no ratings and low ratings. The blue line represents for surgeons with low ratings and the red line for surgeons without ratings in 2013. We first imputed the pseudo ratings for those doctors without ratings based on surgeon characteristics, patient composition and so on. The underlying assumption is that an unrated surgeon would get the same rating as the other rated surgeons who share similar surgeon characteristics and patient composition. Then we compare the distribution of pseudo ratings for these without ratings to the real ratings for low-rating surgeons. The figure suggests that surgeons without ratings would have higher ratings than those with low ratings had they been rated. Overall, Figure 2 suggests that surgeons with no ratings have higher quality than those with low ratings.

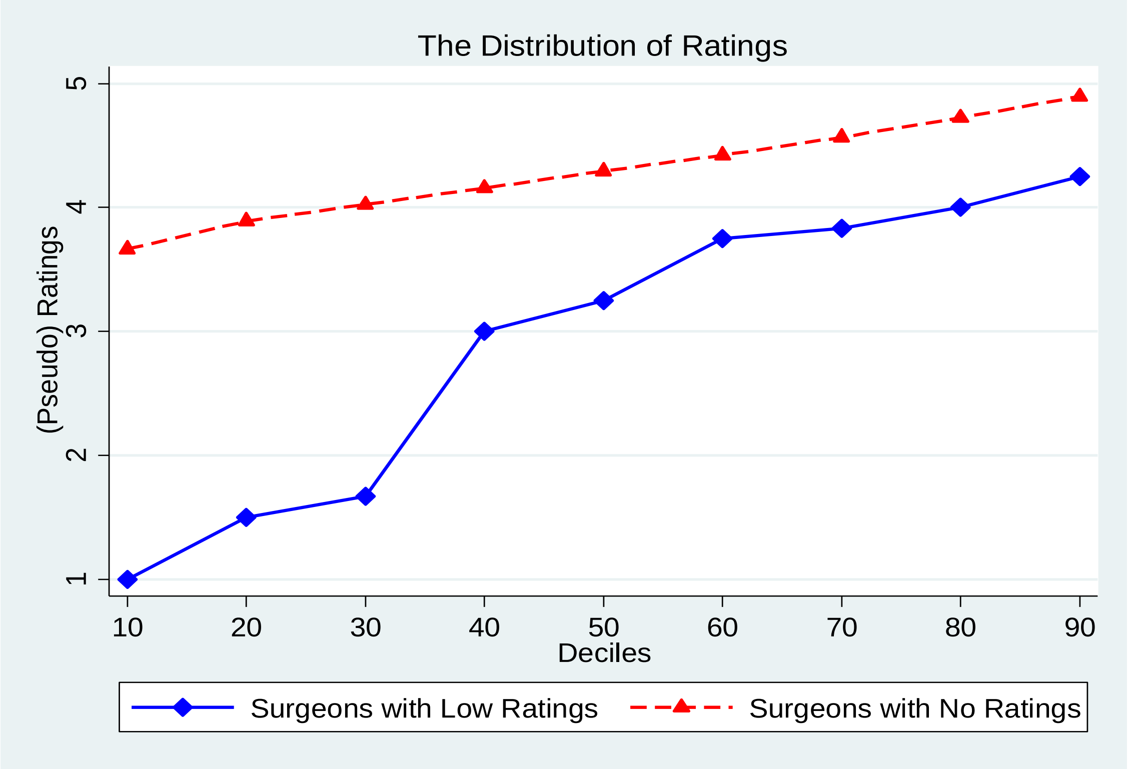


Table A1: Examples of the Text of Patient Reviews

Review ID	Review Text
1	My mom, at 83, had quadruple cardiac bypass surgery performed by Dr. Accola. I knew his reputation as a surgeon because he did a quintuple cardiac bypass on my father-in-law last year. His bedside manner was the best and my mom went into surgery very relaxed and confident. All went great and she's back to most of her activities after only 3 1/2 weeks.
2	Terrific Doctor. My wife had 2x bi pass surgery. He followed up after operation and made sure hosp. steff took care. The BEST.
3	Performed heart surgery that was a complete failure. Had to wait 6 months for another surgeon to redo the operation.
4	Put back together a heart that was badly broken after others said it couldn't be done and the nurses have gone out of their way to help me with at home needs after sugery
5	He's a great physician..He performed a thoracotomy precidure on me.I am very pleased Even though I was anxious to go gome home,he convinced me that it was best that I stay so I wouldn't have any problems..July 7,2011 was the date of my surgery,and since I feel great. He is very stern but caring.I wouls recommend him in a heart beat.
6	very courteous and knowledgable. Great bedside manners. saved my life while in the hospital

Table A2: Estimation of a Spline Model

	Spline Model	
VARIABLES	LPM	Logit
Segment 1: [1, 4)	-0.357***	-0.962***
	(0.121)	(0.369)
Segment 2: [4, 5]	0.018	0.822
	(0.015)	(0.648)
Patient Characteristics	Y	Y
Surgeon Characteristics	Y	Y
Hospital Characteristics	Y	Y
Quarter Dummies	Y	Y
Observations	911	911

Standard errors are clustered by surgeon

*** p<0.01, ** p<0.05, * p<0.1

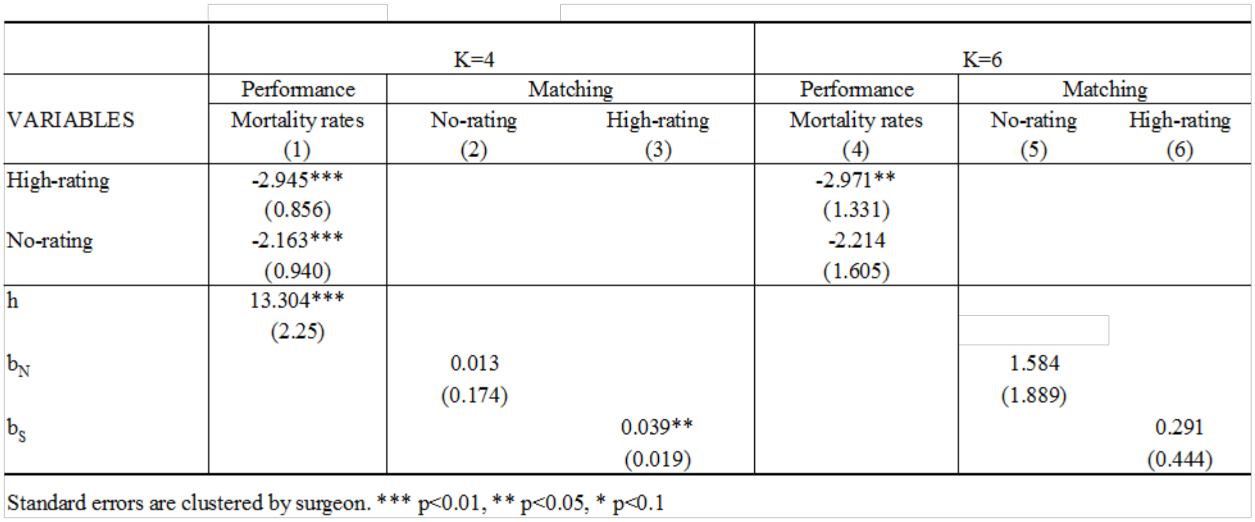
Table A3: Correlation Matrix of Key Independent Variables

	High-rating	No-rating	Experience	Malpractice	Age	Female	CABG	Heart	Vascular	Kidney	AMII	Obesity	Nutrition	Valve	HVD	Liver	Type1	Type2	Type3	Type4	Type5	Hospital ranking	CICU	Beds	Travel time
High-rating	1																								
No-rating	-0.85	1																							
Experience	0.1	-0.08	1																						
Malpractice	0.04	-0.02	-0.01	1																					
Age	0.01	0	0.06	-0.03	1																				
Female	0.01	0	-0.01	0.02	0.08	1																			
CABG	0.02	-0.02	0.01	0	0.04	-0.02	1																		
Heart	0.02	-0.02	0.05	-0.02	0.12	0.06	0.02	1																	
Vascular	0.03	-0.03	0.02	0.01	0.12	0.02	0.03	0.08	1																
Kidney	0.05	-0.04	0.04	-0.02	0.18	-0.01	0.02	0.21	0.1	1															
AMII	0.01	-0.01	-0.02	0.01	0.04	-0.02	0	0.1	-0.02	0.04	1														
Obesity	0	0.01	-0.01	-0.01	-0.05	0.06	0.03	0.07	0	0.04	0.01	1													
Nutrition	0.05	-0.04	-0.01	0	0.05	0.03	-0.01	0.09	0.03	0.05	0.03	0.01	1												
Valve	0.02	-0.06	0	0.01	0.12	0.01	0.05	0.18	0.02	0.06	0.01	0.01	0.03	1											
HVD	0.01	-0.04	0.03	-0.01	0.17	0.04	0.05	0.22	0.09	0.1	0.03	0	0.03	0.48	1										
Liver	0.05	-0.03	-0.01	-0.01	-0.04	-0.03	-0.02	0.02	0	-0.01	0	0.01	0.02	-0.01	0.02	1									
Type1	-0.03	0.02	-0.02	0.03	-0.01	0.06	0.03	0	-0.03	-0.02	-0.05	0.02	0.01	0.06	0.02	0.01	1								
Type2	-0.03	0.02	-0.05	-0.02	0.05	0.02	0	0.01	0.03	0.01	0.03	-0.01	-0.03	0.02	0.03	0.01	-0.3	1							
Type3	0.01	-0.01	0.01	-0.05	-0.01	-0.02	-0.02	0	-0.01	0.02	0.04	-0.01	0	-0.04	-0.02	-0.03	-0.27	-0.43	1						
Type4	0.09	-0.07	0.05	-0.05	-0.02	-0.06	0	0	-0.02	0	0.01	0.03	0.02	-0.02	-0.01	0	-0.16	-0.25	-0.22	1					
Type5	-0.02	0.03	0.02	0.11	-0.02	0	-0.01	-0.01	0.03	-0.01	-0.04	-0.01	0.01	-0.03	-0.02	0.02	-0.15	-0.25	-0.22	-0.13	1				
Hospital ranking	-0.12	0.11	-0.05	-0.07	-0.01	0.03	-0.04	-0.02	-0.01	-0.02	-0.01	0.01	-0.04	0.01	-0.03	-0.01	0.01	0.02	0.04	-0.04	-0.05	1			
CICU	0.05	-0.02	-0.03	-0.13	-0.02	0	-0.01	0	0.01	0.01	-0.04	0.01	0.04	-0.03	-0.03	-0.01	0	0.01	0.02	0.03	-0.06	-0.02	1		
Beds	0.22	-0.2	0.09	0.16	-0.08	0.01	0.01	0.03	0.04	0.01	-0.05	0.02	-0.02	0.01	0.02	-0.01	0	0	-0.02	-0.01	0.02	-0.12	0.24	1	
Travel time	-0.01	0.03	-0.05	0	-0.04	0	0.01	-0.03	-0.01	-0.03	-0.05	-0.01	-0.02	-0.02	-0.02	-0.01	-0.01	0.02	-0.02	-0.03	0.04	0.01	0.04	0.07	1

Table A4: Pairwise t tests of Patient Characteristics across Surgeon Categories

	low-rating - high-rating		low-rating - no-rating		no-rating - high-rating	
	difference	p-value	difference	p-value	difference	p-value
Age	-0.62	0.47	-0.42	0.59	-0.20	0.66
White	-0.05	0.21	-0.12	<0.01	0.07	<0.01
Black	0.03	0.32	0.04	0.08	-0.02	0.21
Female	-0.04	0.23	-0.04	0.28	-0.01	0.69
Medicare	-0.01	0.90	-0.03	0.50	0.02	0.35
Medicaid	-0.01	0.54	0.01	0.61	-0.03	0.05
Private insurance	0.04	0.23	0.04	0.20	0.00	0.98
Log of income	-0.04	0.73	-0.01	0.96	-0.04	0.51
CABG	0.00	0.70	0.00	0.70	-0.01	0.19
Heart	-0.02	0.65	0.01	0.71	-0.03	0.14
Vascular	-0.03	0.32	0.00	0.92	-0.03	0.05
Kidney	-0.06	0.06	-0.01	0.66	-0.04	0.01
AMI	-0.02	0.67	-0.01	0.87	-0.01	0.60
Obesity	-0.01	0.29	-0.01	0.26	0.00	0.93
Nutrition	-0.03	0.07	0.00	0.84	-0.02	0.01
Valve	0.07	<0.01	0.09	<0.01	-0.02	0.13
HVD	0.07	0.02	0.09	<0.01	-0.02	0.23
Liver	-0.03	<0.01	-0.02	0.01	-0.02	0.03
type1	0.04	0.23	0.01	0.75	0.03	0.06
type2	0.04	0.24	0.01	0.72	0.03	0.10
type3	0.02	0.63	0.03	0.45	-0.01	0.69
type4	-0.10	<0.01	-0.02	0.35	-0.08	<0.01
type5	-0.01	0.80	-0.03	0.25	0.02	0.14

Table A5: Estimation of the Two-Stage Model with Larger Value of K



	K=4			K=6		
	Performance	Matching		Performance	Matching	
VARIABLES	Mortality rates	No-rating	High-rating	Mortality rates	No-rating	High-rating
	(1)	(2)	(3)	(4)	(5)	(6)
High-rating	-2.945***			-2.971**		
	(0.856)			(1.331)		
No-rating	-2.163***			-2.214		
	(0.940)			(1.605)		
h	13.304***					
	(2.25)					
b_N		0.013			1.584	
		(0.174)			(1.889)	
b_S			0.039**			0.291
			(0.019)			(0.444)

Standard errors are clustered by surgeon. *** p<0.01, ** p<0.05, * p<0.1

Appendix C: A Mathematical Model for Hypothesis A

We assume a patient's initial health condition, denoted by h, is either good (h_1) with probability λ or poor (h_0) with probability $1-\lambda$. Here $h_1 > h_0$. After receiving treatment from the health care provider, the patient's health condition, denoted by y, is either y_1 or y_0 where $y_1 > y_0$. The realization of y depends both on the initial health condition h and the skill of the care provider, s, which takes two values, s_1 or s_0. Define the distribution of y conditional on h and s as the following:

$$\mathbf{P}(y = y_1 | h = h_1, s = s_1) = P_{11},\ \mathbf{P}(y = y_1 | h = h_0, s = s_1) = P_{01},$$

$$\mathbf{P}(y = y_1 | h = h_1, s = s_0) = P_{10},\ \mathbf{P}(y = y_1 | h = h_0, s = s_0) = P_{00}.$$

Assumption 1 (Monotonicity in Skill) *Better medical skill always leads to higher probability of better outcome.*

Without loss of generality, we assume

$$P_{11} > P_{10} \text{ and } P_{01} > P_{00}$$

and interpret s_1 as high type (i.e., better skill) and s_0 as low type.

The patient and his/her family members have imperfect information both about the patient's initial health condition and the patient's post-treatment health condition. We model this fact by assuming that the patient and his/her family members observe two signals

$$\hat{h} = h + e_1,\ \text{and } \hat{y} = y + e_2$$

where e_1 and e_2 are two random variables with mean zero. For simplicity, we assume e_1 and e_2 are independent of h and y.

Upon observing $\hat{y}$ after receiving the medical service, the patient or his/her family members will give the medical service provider a high rating R_1 if their observed health improvement, $\hat{y} - \hat{h}$, is above a threshold, $\bar{r} + e_3$ where e_3 is a zero-mean random variable capturing the patient-specific disturbance to the threshold $\bar{r}$. On the other hand, the patient or his/her family members will give the medical service provider a low rating R_{-1} if the health improvement is below a threshold, $\underline{r} + e_4$ where e_4 is a zero-mean random variable capturing the patient-specific disturbance to the threshold $\underline{r}$. Underlying this construction is the following assumption.

Assumption 2 (Monotonicity in Rating)

- *For any patient, if he/she will give the high rating R_1 to a provider when his/her improvement in health condition is x_1, then he/she will also give the high rating if his/her improvement in health condition is $x > x_1$.*
- *For any patient, if he/she will give the low rating R_{-1} to a provider when his/her improvement in health condition is x_0, then he/she will also give the low rating if his/her improvement in health condition is $x < x_0$.*

Therefore, the medical service provider will receive a high rating if $\hat{y} - \hat{h} \geq \bar{r} + e_3$, or equivalently $e_2 - e_1 - e_3 \geq \bar{r} + h - y$, and will receive a low rating if $\hat{y} - \hat{h} \leq \underline{r} + e_4$, or equivalently $e_2 - e_1 - e_4 \leq \underline{r} + h - y$. Define $\bar{e} \equiv e_2 - e_1 - e_3$ and denote its cumulative distribution function by F which is assumed to be continuous and include the interval $[\bar{r} + h_0 - y_1, \bar{r} + h_1 - y_0]$ in its support. Similarly, define $\underline{e} \equiv e_2 - e_1 - e_4$ and denote its cumulative distribution function by G which is assumed to be continuous and include the interval $[\underline{r} + h_0 - y_1, \underline{r} + h_1 - y_0]$ in its support. Hence, the medical service provider will receive a high rating with probability $1 - F(\bar{r} + h - y)$ and receive a low rating with probability $G(\underline{r} + h - y)$.

Proposition 1 (Informativeness of Online Rating) *The conditional probability that a provider is a high type is higher if the rating is higher, i.e.,*

$$\mathbf{P}(s = s_1 | R = R_1) > \mathbf{P}(s = s_1 | R = R_{-1}).$$

Proof. We will first show that

$$\mathbf{P}(R = R_1 | s = s_1) > \mathbf{P}(R = R_1 | s = s_0) \text{ and } \mathbf{P}(R = R_{-1} | s = s_1) < \mathbf{P}(R = R_{-1} | s = s_0),$$

and then we will use the Bayes' rule to show the conclusion. We organize the proof into three steps accordingly.

First, we show $\mathbf{P}(R = R_1 | s = s_1) > \mathbf{P}(R = R_1 | s = s_0)$.

Note that

$$\mathbf{P}(R = R_1 | h = h_1, s = s_1) = 1 - [P_{11} F(\bar{r} + h_1 - y_1) + (1 - P_{11}) F(\bar{r} + h_1 - y_0)]$$

$$\mathbf{P}(R = R_1 | h = h_0, s = s_1) = 1 - [P_{01} F(\bar{r} + h_0 - y_1) + (1 - P_{01}) F(\bar{r} + h_0 - y_0)],$$

hence,

$$\begin{aligned}\mathbf{P}(R = R_1 | s = s_1) = 1 &- \lambda \left(P_{11} F(\bar{r} + h_1 - y_1) + (1 - P_{11}) F(\bar{r} + h_1 - y_0)\right) \\ &-(1 - \lambda) \left(P_{01} F(\bar{r} + h_0 - y_1) + (1 - P_{01}) F(\bar{r} + h_0 - y_0)\right).\end{aligned}$$

Similarly, we have

$$\mathbf{P}(R = R_1 | h = h_1, s = s_0) = 1 - [P_{10} F(\bar{r} + h_1 - y_1) + (1 - P_{10}) F(\bar{r} + h_1 - y_0)]$$

$$\mathbf{P}(R = R_1 | h = h_0, s = s_0) = 1 - [P_{00} F(\bar{r} + h_0 - y_1) + (1 - P_{00}) F(\bar{r} + h_0 - y_0)]$$

hence,

$$\begin{aligned}\mathbf{P}(R = R_1 | s = s_0) = 1 &- \lambda \left(P_{10} F(\bar{r} + h_1 - y_1) + (1 - P_{10}) F(\bar{r} + h_1 - y_0)\right) \\ &-(1 - \lambda) \left(P_{00} F(\bar{r} + h_0 - y_1) + (1 - P_{00}) F(\bar{r} + h_0 - y_0)\right).\end{aligned}$$

So $\mathbf{P}(R = R_1 | s = s_1) > \mathbf{P}(R = R_1 | s = s_0)$ if and only if

$$\lambda (P_{11} - P_{10}) \left(F(\bar{r} + h_1 - y_0) - F(\bar{r} + h_1 - y_1)\right) + (1 - \lambda)(P_{01} - P_{00}) \left(F(\bar{r} + h_0 - y_0) - F(\bar{r} + h_0 - y_1)\right) > 0,$$

which is true because $P_{11} > P_{10}$, $P_{01} > P_{00}$, and $y_1 > y_0$.

Second, we show $\mathbf{P}(R = R_{-1} | s = s_1) < \mathbf{P}(R = R_{-1} | s = s_0)$.

Note that

$$\mathbf{P}(R = R_{-1} | h = h_1, s = s_1) = P_{11} G(\underline{r} + h_1 - y_1) + (1 - P_{11}) G(\underline{r} + h_1 - y_0)$$

$$\mathbf{P}(R = R_{-1} | h = h_0, s = s_1) = P_{01} G(\underline{r} + h_0 - y_1) + (1 - P_{01}) G(\underline{r} + h_0 - y_0),$$

hence,

$$\begin{aligned}\mathbf{P}(R = R_{-1} | s = s_1) = &\lambda \left(P_{11} G(\underline{r} + h_1 - y_1) + (1 - P_{11}) G(\underline{r} + h_1 - y_0)\right) \\ &+(1 - \lambda) \left(P_{01} G(\underline{r} + h_0 - y_1) + (1 - P_{01}) G(\underline{r} + h_0 - y_0)\right).\end{aligned}$$

Similarly, we have

$$\mathbf{P}(R = R_{-1} | h = h_1, s = s_0) = P_{10} G(\underline{r} + h_1 - y_1) + (1 - P_{10}) G(\underline{r} + h_1 - y_0)$$

$$\mathbf{P}(R = R_{-1} | h = h_0, s = s_0) = P_{00} G(\underline{r} + h_0 - y_1) + (1 - P_{00}) G(\underline{r} + h_0 - y_0),$$

hence,

$$\begin{aligned}\mathbf{P}(R = R_{-1} | s = s_0) = &\lambda \left(P_{10} G(\underline{r} + h_1 - y_1) + (1 - P_{10}) G(\underline{r} + h_1 - y_0)\right) \\ &+(1 - \lambda) \left(P_{00} G(\underline{r} + h_0 - y_1) + (1 - P_{00}) G(\underline{r} + h_0 - y_0)\right).\end{aligned}$$

So $\mathbf{P}(R = R_{-1} | s = s_1) < \mathbf{P}(R = R_{-1} | s = s_0)$ if and only if

$$\lambda (P_{11} - P_{10}) \left(G(\underline{r} + h_1 - y_0) - G(\underline{r} + h_1 - y_1)\right) + (1 - \lambda)(P_{01} - P_{00}) \left(G(\underline{r} + h_0 - y_0) - G(\underline{r} + h_0 - y_1)\right) > 0,$$

which is true because $P_{11} > P_{10}$, $P_{01} > P_{00}$, and $y_1 > y_0$.

Finally, by Bayes' rule, $\mathbf{P}(s = s_1|R = R_1) > \mathbf{P}(s = s_1|R = R_{-1})$ is equivalent to

$$\begin{aligned} & \frac{\mathbf{P}(s = s_1)\mathbf{P}(R = R_1|s = s_1)}{\mathbf{P}(s = s_1)\mathbf{P}(R = R_1|s = s_1) + \mathbf{P}(s = s_0)\mathbf{P}(R = R_1|s = s_0)} \\ > \quad & \frac{\mathbf{P}(s = s_1)\mathbf{P}(R = R_{-1}|s = s_1)}{\mathbf{P}(s = s_1)\mathbf{P}(R = R_{-1}|s = s_1) + \mathbf{P}(s = s_0)\mathbf{P}(R = R_{-1}|s = s_0)} \\ \Leftrightarrow \quad & \frac{\mathbf{P}(s = s_0)\mathbf{P}(R = R_1|s = s_0)}{\mathbf{P}(s = s_1)\mathbf{P}(R = R_1|s = s_1)} < \frac{\mathbf{P}(s = s_0)\mathbf{P}(R = R_{-1}|s = s_0)}{\mathbf{P}(s = s_1)\mathbf{P}(R = R_{-1}|s = s_1)}. \end{aligned}$$

The last inequality is true because

$$\frac{\mathbf{P}(R = R_1|s = s_0)}{\mathbf{P}(R = R_1|s = s_1)} < 1 < \frac{\mathbf{P}(R = R_{-1}|s = s_0)}{\mathbf{P}(R = R_{-1}|s = s_1)}$$

from the first two steps. ■

The above proposition gives a formal justification of the informational value of online physician ratings. However, for some type of medical services patients may have little information about their health condition before or after receiving service. We formally examine this by analyzing the extreme case when the signal-to-noise ratio observed by patients approaches zero. To do so, we parameterize F and G using normal distribution:

$$F(x) \sim N(0, \sigma_F^2), \; G(x) \sim N(0, \sigma_G^2).$$

Corollary 1 (Non-informativeness of Online Rating) $\mathbf{P}(s = s_1|R = R_1) = \mathbf{P}(s = s_1|R = R_{-1})$ *if and only if* $\sigma_F = \sigma_G = \infty$.

Proof. The conclusion follows directly from the following facts:

$$\lim_{\substack{\sigma_F \to \infty \\ \sigma_G \to \infty}} \mathbf{P}(R = R_1|s = s_0) = \mathbf{P}(R = R_1|s = s_1), \;\; \lim_{\substack{\sigma_F \to \infty \\ \sigma_G \to \infty}} \mathbf{P}(R = R_{-1}|s = s_0) = \mathbf{P}(R = R_{-1}|s = s_1) = 1,$$

which can be easily shown from the proof of Proposition 1. ■

Therefore, from a theoretical point of view, online rating is completely non-informative only when patients have absolutely no information at all about their health condition before or after receiving medical service.