Do Online Reviews of Physicians Reflect Healthcare Outcomes?

Danish H. Saifee, Indranil Bardhan, and Zhiqiang (Eric) Zheng(✉)

Jindal School of Management, University of Texas at Dallas,
Richardson, TX 75080, USA
{danish.saifee, bardhan, ericz}@utdallas.edu

Abstract. Patients are increasingly using online reviews to choose physicians. However, it is not known whether online reviews accurately capture the true quality of care provided by physicians. This research addresses this issue by empirically examining the link between online reviews of a physician and the actual clinical outcomes of patients treated by the physician. Specifically, this study uses online reviews from Vitals.com, and combines that data with patient health outcomes data collected from Dallas-Fort Worth Hospital Council. Our econometric analyses show that there is no clear relationship between online reviews of physicians and their patients' health outcomes, such as readmission and ER visit rates. Our results imply that online reviews may not be as helpful in the context of healthcare as they are for other experience goods such as books, movies, or hotels. Our findings have important implications for healthcare providers, healthcare review websites, and healthcare consumers.

Keywords: Online reviews · Healthcare · Clinical outcomes · Topic modeling · Sentiment analysis · Chronic obstructive pulmonary disease (COPD)

1 Introduction

Online reviews of physicians have the potential to reduce information asymmetry between healthcare providers and patients, empowering patients to make better decisions. A pertinent question, then, is *whether, and to what extent, patients benefit from online reviews of physicians*.

Ascertaining this efficacy is important because of the greater role that online reviews play in patients' decisions about which physicians to see and which ones to avoid (Hanauer et al. 2014). In fact, many physicians monitor their reviews and ratings closely and try to boost their ratings on review sites such as Yelp, Vitals, and RateMDs.[1] There are even numerous instances in which physicians have filed defamation lawsuits over negative patient reviews.[2] Evidently, patients are increasingly using online reviews to select physicians as well as other healthcare providers,

[1] https://www.washingtonpost.com/news/to-your-health/wp/2016/05/27/docs-fire-back-at-bad-yelp-reviews-and-reveal-patients-information-online, last accessed 03/31/2017.

[2] http://www.oregonlive.com/today/index.ssf/2015/11/doctor_sues_patient_over_negat.html and http://blog.ericgoldman.org/archives/2015/01/another-failed-doctor-lawsuit-against-a-patient-for-online-reviews-brandner-v-molonguet.htm, last accessed on 03/31/2017.

H. Chen et al. (Eds.): ICSH 2017, LNCS 10347, pp. 161–168, 2017.
https://doi.org/10.1007/978-3-319-67964-8_15

prompting providers to take these reviews rather seriously. Despite the increasing importance of online reviews in healthcare, it is not at all clear that these reviews are actually leading to better patient choices. Put differently, the relationship between physician reviews and quality of physician care remains largely unexplored. A major challenge lies in the difficulty of accurately measuring the quality of care provided by physicians. Some researchers have used surveys to assess patients' perceptions of physicians to construct a proxy for physician quality (Doyle et al. 2013). However, patient perception may not be the same as reality.

To address this challenge, we obtained research data for this study from two sources. The first dataset (spanning from 2006 to 2015) was obtained from Dallas Fort Worth Hospital Council's (DFWHC) Research Foundation database on COPD patients. This dataset consists of approximately 630,000 inpatient admission-discharge records, 10,200 attending physicians, and 330,000 patients. The second dataset of about 14,500 physicians in North Texas (spanning from 2007 to 2015) was collected from Vitals.com. This dataset provides data on physician characteristics and online reviews, including textual reviews, review ratings, and years of physician practice. We integrated the data from these two data sets, using physician names, to create a unique dataset that provides patient health outcomes for physicians who are also rated and reviewed by their patients and examine whether online reviews of physicians are reliable predictors of their patients' clinical outcomes. In other words, if a physician receives very positive online reviews, does that also mean that her patients also exhibit good health outcomes?

Our results show that patients under the care of physicians with better online reviews may not necessarily experience better clinical outcomes, compared to patients receiving care from physicians with worse review ratings. Our results have broader implications for healthcare providers and consumers.

2 Literature Review

A few recent studies in the information systems area examine online ratings and reviews of care providers. For example, Bardach et al. (2013) suggest that reviewers on Yelp may possess knowledge on important aspects of care. Gao et al. (2015) find that physicians who are rated lower in quality (by the patient population) in offline surveys are less likely to be rated online and online ratings are positively correlated with patient reviews, and that online ratings tend to be exaggerated at the upper end of the quality spectrum. They construct their quality measure using patient surveys conducted by Consumers' Checkbook using the instrument and procedure designed by the U.S. Agency for Healthcare Research and Quality. Gray et al. (2015) don't find a clear evidence of association between physician website ratings and traditional quality measures such as blood pressure or low-density lipoprotein controlled. Although these papers shed much needed light on patient perception of providers, they either (1) rely on limited care quality measures such as offline patient satisfaction surveys or (2) are mostly limited to aggregated numeric ratings of physicians as a surrogate for patient perception and often do not consider the rich sentiments expressed in textual reviews.

Studies in medical journals examining the relationship between patient experience and clinical outcomes, such as the mortality rate, 30-day readmission rate, and clinical safety, are also relevant (e.g., Glickman et al. 2010; Boulding et al. 2011). A comprehensive review, conducted by Doyle et al. (2013), summarizes prior research that examined the relationship between patient experience and clinical outcomes. Majority of these studies find positive connection between patient satisfaction and clinical outcomes. Although these findings provide important insights, a bulk of the studies in this literature stream rely on offline surveys to solicit patient experience, which do not allow significant parsing of the textual content through sentiment-mining and topic-modeling techniques as can be done with online reviews. These prior studies have also often relied on cross-sectional hospital- or clinic-level data, limiting the extent to which their findings can be extrapolated to the context of patient experience at the physician level. Finally, the use of these survey findings by patients is not nearly as widespread as is that of websites containing reviews of physicians.

The stream of research on online consumer reviews has generally found that online reviews of products, such as books, and services, and hospitality, enable consumers to make more informed decisions by providing them information on other consumers' perspectives (e.g., Vermeulen and Seegers 2009; Chevalier and Mayzlin 2006). However, it is not clear whether the findings in prior research relating to the efficacy and usefulness of reviews automatically are applicable to a healthcare context. That is, the true quality of healthcare services could be significantly more difficult to assess when compared to the context of hotels, restaurants, or other similar services.

3 Research Question

Online reviews of physicians can contain rich information and often provide significantly more information than numeric (star) ratings. For example, they can help users gather information about the experience of past patients of a physician including, but not limited to, bedside manners of the physician, whether she spends sufficient time with her patients, follows up after the visit, and the thoroughness of explanations (of diagnoses and procedures) provided by her or her staffs. Some aspects of online reviews, such as detailed accounts of procedures and clinical steps performed by a physician, may even provide useful cues about the clinical aspects of care. Moreover, online reviews can influence patients' choices. Based on a survey of patients, Hanauer et al. (2014) report that 35% of the respondents selected physicians with good ratings, while 37% avoided those with bad ones. Thus, it suggests that prospective patients expect physicians who receive largely positive online reviews to deliver better clinical outcomes. However, to the best of our knowledge, *there is no data-driven evidence that this is indeed the case.*

There ought to be a concern about the reliability of online reviews of a physician in predicting the quality of service delivered by the physician because a patient, who typically lacks a comprehensive medical training, may not be well equipped to ascertain the clinical proficiency of a physician.[3] Also, an online review of a physician may not necessarily provide information on the clinical characteristics of that physician's care delivery and could easily overemphasize factors such as flexibility in scheduling appointments, promptness and courteousness of the staff, receptiveness and of the medical team, etc. These factors are not necessarily indicative of the level of clinical care provided by the physician. This leads us to our central research question:

> ***Are physicians who receive better online reviews more likely to deliver better clinical outcomes for their patients?***

4 Research Framework

4.1 Variables

The two clinical outcome measures used in our study, *Future30DayReadm* and *FutureERVisit*, are constructed from the DFWHC dataset. *Future30DayReadm* is the proportion of future patient admissions within thirty days of the previous discharge date, for a given physician at a given point in time (quarter), due to the same principal diagnosis (i.e. COPD). We construct a binary variable that equals 1 for a patient visit only if that patient's next admission date is within 30 days of his current discharge date. Then, for each attending physician, we calculate the rolling average of this dummy variable, beginning from the chronologically last (most recent) inpatient admission record to obtain *Future30DayReadm*. *FutureERVisit* is the proportion of future patient admissions involving a visit to an emergency room, with construction similar to *Future30DayReadm*.

The key explanatory variables with regard to online reviews are *OverallRating* and *SentimentScore*. *OverallRating* is the average of the overall star ratings of a physician at a given time, and *SentimentScore* is the average of the sentiment score (up to a time-point) derived from textual reviews in vitals.com. The sentiment analysis technique that we applied classifies the sentiment of each word in a review into four sentiment categories: very positive, positive, very negative, and negative (based on the vocabulary provided by Nielsen 2011). Then, aggregation across all sentiment words within a review yields an overall sentiment score, *SentScorePerReview,* for the review.[4] To control for variations in clinical outcomes arising from variations in the patient-mix handled by physicians, we create several controls. (Note that these controls as well as the key explanatory variables are backward-looking, as opposed to the forward-looking outcome variables *Future30DayReadm* and *FutureERVisit*.) We also control for sentiment variance, and *latent topics* underlying the textual content of online reviews.

We, next, conduct a fine-grained textual analyses of the online reviews by deploying latent Dirichlet allocation (LDA) (Blei et al. 2003) to derive latent topics

[3] Source: http://health.usnews.com/health-news/patient-advice/articles/2014/12/19/are-online-physician-ratings-any-good%20, last accessed on 03/31/2017.

[4] $$\begin{aligned} SentScorePerReview\; = \;& 2 \times number\,of\,very\,positive\,words\; + \;1 \times number\,of\,positive\,words \\ & -1 \times number\,ofnegative\,words - 2 \times numberof\;very\,negative\,words. \end{aligned}$$

underlying the textual content in online reviews. Figure 1 plots the distribution of the sentiment category (positive, neutral, or negative) across these four latent topics.[5] Reviews under the latent topic "Overall Care" tend to be rated more positively, as opposed to reviews for the other three latent topics, while reviews for the latent topic "Promptness" tend to be more negative, compared to the rest. This provides some insights into how the types of underlying themes might be driving sentiments expressed in online reviews.

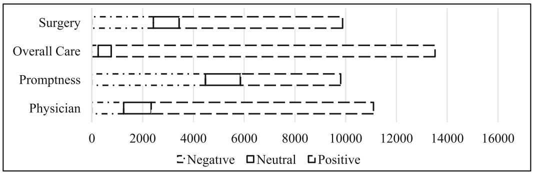

Fig. 1. Frequency of sentiments by latent topics

4.2 Estimation Model and Results

To account for potential physician-time-level fixed effects and omitted variable biases, we consider a two-stage two-way fixed-effects panel regression with instrument variables. The physician fixed effects account for time-invariant physician attributes that are not captured in our data. The use of forward-looking measures for the outcome variables helps us mitigate possible biases in coefficient estimates of our key explanatory variables, which can arise from simultaneity between these explanatory variables and clinical outcomes. We construct two instrument variables (IV), which represent the average sentiment score of online reviews and average score of online ratings received by the focal attending physician's peer physician group in the same hospital system, over the previous two and a half years (10 quarters). A physician's reviews (online perception) can be reliably predicted using the online perception of other physicians in the same hospital system, aggregated over time. But, this time-aggregated online perception of her peer group need not systematically determine clinical outcomes of her (i.e. focal physician's) patients. The first stage regression results indicate that these IVs are strong. Table 1 presents the second-stage regression estimation results.

The coefficient estimates of our key explanatory variable—*SentimentScore* and *OverallRating*—in Table 1 demonstrate that physicians who receive better online reviews or higher online star ratings, compared to their peers, do not necessarily exhibit better health outcomes as measured by the future 30-day readmission or ER visit rates of their patients. In fact, higher overall ratings are associated with a higher frequency of future ER visits, casting additional doubts on the efficacy of online reviews and ratings.

[5] If *SentScorePerReview* is greater than zero, we label the review "positive;" if it is less than zero, we label it "negative," and "neutral" if it is zero.

Table 1. Two-stage Two-way fixed effects IV estimation results (second-stage)

	Future30DayReadm		*FutureERVisit*	
	(1)	(2)	(3)	(4)
SentimentScore	−0.013 (0.037)	–	0.129^{+} (0.067)	–
OverallRating		0.151^{+} (0.081)		**0.515** (0.188)**
SentimentVariance	0.134 (0.257)	−0.111 (0.086)	−0.857 (0.471)	−0.478* (0.200)
ReviewWordsNum	0.000 (0.000)	0.000 (0.000)	−0.001 (0.000)	0.001* (0.000)
TopicSurgery	−0.049 (0.078)	0.036 (0.032)	0.284* (0.135)	0.221** (0.074)
TopicPhysician	−0.035 (0.037)	0.026 (0.027)	0.181** (0.068)	0.224** (0.064)
TopicPromptness	−0.069 (0.137)	0.234 (0.136)	0.536* (0.254)	0.929** (0.322)
Experience	−0.002 (0.004)	0.003 (0.005)	0.026* (0.011)	0.043** (0.014)
ERVisit	−0.011 (0.024)	−0.016 (0.021)	–	–
30DayReadm	–	–	0.659*** (0.171)	0.375** (0.143)
VisitsNum	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000* (0.000)
LOS	0.011** (0.004)	0.010** (0.004)	−0.021* (0.009)	−0.017 (0.009)
Expired	−0.518* (0.262)	−0.738* (0.316)	−1.895** (0.643)	−2.609*** (0.706)
PtAge	0.002 (0.002)	−0.001 (0.001)	−0.012*** (0.003)	−0.015*** (0.003)
Female	0.030 (0.035)	0.009 (0.028)	−0.32** (0.113)	−0.203** (0.075)
SevMajExt	0.030 (0.052)	0.069 (0.046)	−0.238* (0.101)	−0.200 (0.112)
MortMajExt	−0.260*** (0.058)	−0.205*** (0.051)	0.340** (0.115)	0.405** (0.128)
SwitchHospSys	−0.030 (0.034)	−0.044 (0.028)	−0.129 (0.075)	−0.308*** (0.067)
SwitchHosp	0.064* (0.031)	0.082** (0.028)	0.118 (0.072)	0.298*** (0.066)
EthnHisp	0.123* (0.050)	0.182*** (0.051)	0.124 (0.148)	0.095 (0.128)
RaceWhite	−0.047 (0.027)	−0.012 (0.028)	0.098 (0.075)	0.114 (0.073)

p < 0.10^{+}, p < 0.05*, p < 0.01**, p < 0.001***, standard error in parenthesis

Hence, our results suggest that *neither sentiments expressed in reviews nor numeric ratings are accurate predictors of actual clinical outcomes*.

5 Robustness Checks

An endogeneity concern could arise from potential self-selection by patients, i.e. patients with poor health may choose to go to physicians perceived to be of high quality. When that happens, physicians who deliver better clinical outcomes could end up receiving relatively poor reviews. To deal with possible self-selection, we apply the two-stage Heckman selection method. The results from the Heckman method do not lend any evidence to the possibility that patient self-selection is indeed driving our main finding that reviews and ratings are not as useful in predicting clinical outcomes, as commonly believed. These results are omitted due to space constraints.

Next, we consider the possibility that physicians whose patients experience poor clinical outcomes (high readmission or ER visit rates) may be involved in of review manipulation. To examine this, we divided our physicians into two groups: those whose patients have experienced below-average readmission rates (*AvgFut30DayReadm = 0*), and those whose patients have experienced above-average readmission rates (*AvgFut30DayReadm = 1*). We repeat this for ER visit rates and again create two groups for *AvgFutERVisit* = 0 and *AvgFutERVisit* = 1, respectively. We, next scrape the numbers of "recommended" and "not-recommended" reviews for physicians from Yelp. Reviews not recommended are potentially suspicious due to potential for manipulation. Thus, if we find that physicians whose patients have experienced relatively poorer clinical outcomes have a disproportionately larger number (or fraction) of such reviews, we can suspect some manipulation on Yelp, and perhaps other web sites as well. None of the t-tests' results in Table 2 suggest that physicians who deliver above-average readmission or ER visit rates receive a higher number (or fraction) of "not-recommended" reviews, compared to physicians who deliver below-average readmission or ER visit rates, not providing any evidence that physicians are engaging in active manipulation of online reviews.

Table 2. Comparison of number and percent of not-recommended yelp reviews

	Number of Not-Reco reviews				Percent of Not-Reco reviews			
	AvgFut30DayReadm		*AvgFutERVisit*		*AvgFut30DayReadm*		*AvgFutERVisit*	
	0	1	0	1	0	1	0	1
Mean (Std Dev)	0.653 (4.394)	0.268 (0.917)	0.762 (5.250)	0.310 (1.550)	37.382 (39.244)	37.734 (39.659)	37.785 (39.672)	37.557 (39.266)
No. of obs.	623	370	399	490	198	93	132	122
t	2.12		1.66		−0.07		0.05	
p	0.035		0.097		0.944		0.963	

6 Contributions and Implications

In summary, our paper contributes to and builds on prior research in the following four ways: (1) it attempts to study the relationship between online reviews of a physician and actual clinical outcomes of the physician's patients, (2) it measures clinical outcomes *objectively* based on the readmission rate and ER visit rate at the patient-admission level, (3) it analyzes the fine-grained textual content of reviews, rather than relying only on aggregated numeric ratings, in examining patients' opinions, and (4) it applies text mining techniques as well as econometric methods, including a series of robustness checks, to investigate whether the textual content in reviews of physicians is indeed a reliable predictor of clinical outcomes. To the best of our knowledge, there is no prior research that has addressed all of the above dimensions in a unified framework, as we have proposed in this paper.

Our study has several managerial and healthcare policy implications. First, healthcare consumers need to be cautious, when using online reviews and ratings to form opinions about physician quality. Physicians who receive better online reviews, may not necessarily exhibit better quality as measured by their patients' health outcomes. Second, our results suggest that online reviews require further scrutiny than what is currently done to decipher physician quality. Our study lends support to the concerns raised in the popular press about over-reliance on online reviews of physicians to assess actual physician quality particularly in the context of chronic conditions.

Third, hospitals and clinics should be careful about relying on online reviews of physicians for evaluating physician performance, since they do not serve as accurate predictors of future patient health outcomes.

References

Bardach, N.S., Asteria-Penaloza, R., Boscardin, W.J., Dudley, R.A.: The relationship between commercial website ratings and traditional hospital performance measures in the USA. BMJ Qual. Saf. **22**(3), 194–202 (2013)

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Boulding, W., Glickman, S.W., Manary, M., Schulman, K.A., Staelin, R.: Relationship between patient satisfaction with inpatient care and hospital readmission within 30 days. Am. J. Managed Care **17**(1), 41–48 (2011)

Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: online book reviews. J. Mark. Res. **43**(3), 345–354 (2006)

Doyle, C., Lennox, L., Bell, D.: A systematic review of evidence on the links between patient satisfaction and clinical safety and effectiveness. BMJ Open **3**(1), 1–18 (2013)

Gao, G.G., Greenwood, B.N., Agarwal, R., McCullough, J.S.: Vocal minority and silent majority: how do online ratings reflect population perceptions of quality? MIS Q. **39**(3), 565–589 (2015)

Glickman, S.W., Boulding, W., Manary, M., Staelin, R., Roe, M.T., Wolosin, R.J., Ohman, E. M., Peterson, E.D., Schulman, K.A.: Patient satisfaction and its relationship with clinical quality and inpatient mortality in acute myocardial infarction. Circ. Cardiovasc. Qual. Outcomes **3**(2), 188–195 (2010)

Gray, B., Vandergrift, J.L., Gao, G.G., McCullough, J.S., Lipner, R.S.: Website ratings of physicians and their true quality of care. JAMA Internal Med. **175**(2), 291–293 (2015)

Hanauer, D.A., Zheng, K., Singer, D.C., Gebremariam, A., Davis, M.M.: Public awareness, perception, and use of online physician rating sites. J. Am. Med. Assoc. **311**(7), 734–735 (2014)

Nielsen, F.Å.: A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 Workshop on ‘Making Sense of Microposts’: Big Things Come in Small Package, CEUR Workshop Proceedings, no. 718, pp. 93–98 (2011)

Vermeulen, I.E., Seegers, D.: Tried and tested: the impact of online hotel reviews on online hotel reviews on consumer consideration. Tour. Manag. **30**(1), 123–127 (2009)