



Information Systems Research

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Are Online Reviews of Physicians Reliable Indicators of Clinical Outcomes? A Focus on Chronic Disease Management

Danish H. Saiffee, Zhiqiang (Eric) Zheng, Indranil R. Bardhan, Atanu Lahiri

To cite this article:

Danish H. Saiffee, Zhiqiang (Eric) Zheng, Indranil R. Bardhan, Atanu Lahiri (2020) Are Online Reviews of Physicians Reliable Indicators of Clinical Outcomes? A Focus on Chronic Disease Management. Information Systems Research 31(4):1282-1300. <https://doi.org/10.1287/isre.2020.0945>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2020, INFORMS

Please scroll down for article—it is on subsequent pages







With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes. For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Are Online Reviews of Physicians Reliable Indicators of Clinical Outcomes? A Focus on Chronic Disease Management

Danish H. Saifee,^a Zhiqiang (Eric) Zheng,^b Indranil R. Bardhan,^c Atanu Lahiri^b

^a Culverhouse College of Business, University of Alabama, Tuscaloosa, Alabama 35487; ^b Naveen Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080; ^c McCombs School of Business, University of Texas at Austin, Austin, Texas 78705

Contact: dhsaifee@cba.ua.edu,  <https://orcid.org/0000-0002-2058-4117> (DHS); ericz@utdallas.edu,  <https://orcid.org/0000-0001-8483-8713> (Z(E)Z); indranil.bardhan@mcombs.utexas.edu,  <https://orcid.org/0000-0002-4265-5780> (IRB); atanu.lahiri@utdallas.edu,  <https://orcid.org/0000-0003-3174-8014> (AL)

Received: July 20, 2017

Revised: October 29, 2018; August 21, 2019; March 19, 2020

Accepted: May 16, 2020

Published Online in Articles in Advance: September 23, 2020

<https://doi.org/10.1287/isre.2020.0945>

Copyright: © 2020 INFORMS

Abstract. Current trends on patient empowerment indicate that patients who play an active role in managing their health also seek and use information obtained from online reviews of physicians. However, it is far from certain whether patient-generated online reviews accurately reflect the quality of care provided by physicians, especially in the context of chronic disease care. Because chronic diseases require continuous care, monitoring, and multiple treatments over extended time periods, it can be quite hard for patients to assess the effectiveness of a particular physician accurately. Given this credence nature of chronic disease care, the research question is the following: what is the information value associated with online reviews of physicians who treat chronic disease patients? We address this issue by examining the link between online reviews of physicians and their patients' actual clinical outcomes based on a granular admission–discharge data set. Contrary to popular belief, our study finds that there is no clear relationship between online reviews of physicians and their patients' clinical outcomes, such as readmission risk or emergency room visits. Our findings have two major implications: (a) online reviews may not be helpful in the context of healthcare services with credence aspects; (b) because treatments of chronic diseases have more credence good characteristics when compared with surgeries or other acute care services, one should not extrapolate research on surgeries and acute care services to chronic disease care. Rather, one should acquire a better understanding of the information conveyed in online reviews regarding a physician's ability to deliver certain clinical outcomes before drawing inferences. Our findings have important ramifications for all stakeholders including hospitals, physicians, patients, payers, and policymakers.

History: Anindya Ghose, Senior Editor; Ashish Agarwal, Associate Editor.

Funding: Support from the National Natural Science Foundation of China [Grant 91646206] funded Z. Zheng's travel.

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/isre.2020.0945>.

Keywords: online reviews and ratings • healthcare • sentiment analysis • topic modeling • clinical outcomes • credence goods • chronic disease care • chronic obstructive pulmonary disease • social media

1. Introduction

Online consumer reviews play an important role in almost every market today, and the healthcare industry is no exception. Online reviews of physicians, in particular, have the potential to reduce information asymmetry between healthcare providers and patients, empowering patients to make better decisions. A simple but fundamental question then is whether and to what extent patients can rely on online reviews of physicians. We examine this question in the context of chronic disease care. It is of particular interest because recent estimates by the Centers for Disease Control indicate that 90% of the \$3.5 trillion that the United States spends on healthcare are on people with chronic diseases and mental health conditions.¹ These staggering costs have already led to

calls for additional research in this domain (Bardhan et al. 2020).

1.1. Theoretical Background

Researchers have long recognized the trichotomy of search, experience, and credence qualities of products and services (Nelson 1970, Darby and Karni 1973). Search goods are those that exhibit search qualities that consumers can gauge ahead of their purchases. In contrast, experience goods involve mostly experiential characteristics that are revealed to consumers only after purchase and consumption. At the end of this spectrum are credence goods, whose quality cannot be ascertained even after the purchase and direct personal experience with the product (Darby and Karni 1973). Broadly, credence goods are “goods which we don't know

whether we need, and even once we've consumed them, still don't know if they were good value. In economic terms, these goods suffer from the worse of possible information failures, particularly with respect to the asymmetry of information between the seller (in this case the doctor) and the consumer" (Smith 2014).

This trichotomy is not only significant from a conceptual standpoint; it also has material implications for the efficacy of online reviews. For example, in their study of Amazon's online reviews, Mudambi and Schuff (2010) demonstrate that the perceived usefulness of online consumer reviews of experience goods is not the same as that of their counterparts for search goods, echoing the idea that firsthand information gathering—as opposed to vicarious learning from reviews—likely plays a more significant role in the case of experience goods. Concerns about the usefulness of reviews only grow when we consider credence goods. First, prior research on search and experience goods generally finds online reviews to be effective (Chevalier and Mayzlin 2006, Chen and Xie 2008, Chintagunta et al. 2010, Zhu and Zhang 2010), but does this insight also extend to credence goods? Second, if consumers find it hard to assess such goods even after consumption, is it not natural to surmise that consumer feedback would be relatively less useful in their case? A major focus of this research is to empirically examine these questions in a rigorous manner to better understand the role of online reviews in the relatively underexplored context of credence goods with chronic disease care serving as a context-specific example.

The defining feature of a credence good is that it is very difficult to accurately assess its quality even after consuming it. Sloan (2001) makes forceful arguments as to why healthcare consumption is largely of credence nature and why perhaps so for chronic disease patients: (1) Patients are not well informed about their own health and healthcare, at least much less informed than are physicians, or in other words, they do not have "as good or nearly as good an understanding of the utility of the product as the producer" (Arrow 1963, p. 952). (2) The demand for care is probabilistic depending on the person's health state, which itself is also probabilistic. (3) Patients often need to trust the experience of healthcare professionals to make healthcare decisions rather than rely on their own knowledge. Through direct personal experience of seeing patients, physicians learn what does or does not work and under what conditions a particular treatment is needed, but consumers of healthcare do not acquire such experience through exposure to repeated trials (of treating hundreds of patients under different conditions). Therefore, although physicians also face uncertainty, they have an institutional information advantage, "not only in terms of having had professional education in

healthcare but also because of extensive learning-by-doing" (Sloan 2001, p. 900). This information asymmetry gives rise to the credence nature of healthcare services. Finally, many health service markets are far from being perfectly competitive, often because of government regulations. This lack of competition further leads to nontransparency of care quality as providers are not obliged to differentiate themselves from one another by fully disclosing their actual care quality.

Although all healthcare services have some credence elements in general, certain types of treatments and services exhibit greater credence characteristics compared with others. For example, consider the context of surgeries discussed in the recent literature (Segal et al. 2012, Lu and Rui 2018). At a minimum, the outcome of a surgery is often easy to understand and assess. As noted by Lu and Rui (2018, p. 2560), "Family members can observe at least whether their beloved survived after the surgery in the hospital or not, and can therefore infer how well the surgery was performed and how good the surgeon is, although their inferences may not be completely accurate." In short, care quality is observable after the fact—at least partially if not fully—which simply means that surgeries do have a significant experience component.

Now, contrast this with the case of a chronic condition, such as chronic obstructive pulmonary disease (COPD), the context of our study. First, unlike ailments that require surgeries, chronic conditions can be slow to develop and may go undiagnosed for substantial periods of time. In other words, the consumer faces greater uncertainties about the need to seek care—if someone breaks a leg, there is no uncertainty as to whether to consult a surgeon (given the state of the person's health), but the need to see a physician is not always obvious in the case of chronic conditions. Second, chronic diseases are typically treatable but not curable per se and require continuous care management throughout a patient's life (Clarke et al. 2017). This also means that evaluating the postconsumption care quality is not as easy from a patient's perspective as evaluating a medical procedure with visible recovery cues. The task of attributing success or failure of a physician is further compounded by the fact that chronic disease patients often visit multiple hospitals and providers over the course of their treatment. This particular aspect also presents a challenge to researchers as it now becomes necessary to track a patient's journey across multiple hospitals and physicians to properly evaluate their quality of care, a challenge we duly undertake in this study.

Finally, a number of social, behavioral, psychological, and economic determinants play significant roles in the long-term management of chronic diseases, making it nearly impossible for a patient to isolate and decipher a physician's performance from

these confounding factors. Friis et al. (2016, p. 1) aptly recognize this patient–provider information gap and observe that “compared to the general population, people with long-term conditions report more difficulties in understanding health information and engaging healthcare providers.” Overall, it should be apparent that chronic disease care contains far more credence aspects than other healthcare services and, for that matter, other search or experience goods studied in the prior literature. This is precisely why it is important to examine the efficacy of online reviews in the setting of chronic disease care without any presumption about the applicability of prior research to this context.

1.2. Practical Relevance

Ascertaining the usefulness of online physician reviews is important because of their increasing popularity among consumers (Ellimoottil et al. 2013, Emmert et al. 2013). According to Hanauer et al. (2014), approximately 60% of patients now consider online reviews to be an important factor in their selection of physicians. Similarly, a recent survey of healthcare consumers finds that more than 80% utilized online review websites, often viewing or posting comments about their interactions with clinical staff (Hedges 2019). In another survey, 28% strongly agreed that a positive online review of a physician would cause them to seek care, and another 27% indicated that a negative review would lead them to avoid that physician altogether (Burkle and Keegan 2015).

Not surprisingly, physicians have also begun monitoring online reviews and ratings closely while looking for ways to boost their ratings on review sites such as Yelp, Vitals, and RateMDs (Ornstein 2016). Some providers have even gone a step further to the extent of using online feedback to better understand patients’ concerns (Jain 2010, Emmert et al. 2016). At the same time, however, there is increasing wariness among healthcare providers about the veracity of online reviews; there have even been instances in which physicians have filed defamation lawsuits over negative patient reviews (Goldman 2015, O’Donnell and Alltucker 2018).

Despite the growing relevance of online physician reviews and assertions from the review websites, it is not clear to us if such reviews are actually reliable.² Specifically, it is hard to draw inferences about the generalizability of prior studies to chronic diseases. This is precisely where we contribute. We are among the first to track patients’ clinical journeys across multiple providers as necessary to comprehensively assess the quality of chronic disease care.

1.3. Research Overview

We synthesized two unique data sets, one capturing online reviews (textual comments and ratings) of physicians and another containing actual clinical outcomes

experienced by their patients. To elaborate, we first obtained a longitudinal, multihospital, admission–discharge data set of patients suffering from COPD. This data set spans a 10-year period and includes patient admissions across 80 hospitals in North Texas. It enables us to objectively measure physician performance using clinical outcomes, such as their patients’ future readmissions and emergency room (ER) visit rates.³ Rolling up these measures to the physician level allows us to measure the overall quality of care provided by each physician. In this context, we need to reemphasize that COPD is not curable but treatable. The condition can be monitored in a number of ways, including X-rays, CT scans, and blood tests. Treatments involve medications, such as bronchodilators, oral and inhalable steroids, and antibiotics, and lung therapies, such as oxygen therapy, among others. If the severity of a patient’s condition is not assessed regularly and appropriate treatments are not provided in time, the patient can relapse and be hospitalized. In other words, by offering preventive care, physicians can help patients avoid unnecessary ER visits and readmissions. Therefore, our clinical outcomes capture a physician’s ability to prevent costly hospital admissions and ER visits.

Second, we extracted online reviews from a popular website and employed sentiment-mining tools to score the overall sentiments expressed in the text of each review. We then combined these two data sets to examine whether sentiments expressed in review texts and their accompanying star ratings are reliable indicators of clinical outcomes delivered by a physician. In other words, if a physician receives relatively positive (negative) online reviews, does that necessarily mean that patients can expect relatively good (bad) clinical outcomes?

Our results suggest that COPD patients under the care of physicians with more positive online reviews may not necessarily experience better clinical outcomes compared with patients who receive care from physicians with worse review scores. The implications are clear. Despite their popularity, online physician reviews may not be uniformly informative to prospective patients. Evidently, chronic disease treatments replete with credence characteristics are different from surgeries or acute care treatments studied in prior research, and insights from prior studies do not extend to chronic disease management. From a policy perspective, our results imply that patients with chronic diseases should be provided with alternative sources of information that accurately reflect care quality so as to help them make informed physician choices.

2. Literature Review

Our study is related to several key streams of literature. The first stream examines off-line survey data

collected by hospitals and clinics for gauging patient experience. The second stream primarily deals with user-generated online content by healthcare consumers. The third stream is the broader literature on online ratings and reviews, which spans contexts other than healthcare. The fourth and final stream of interest is the economics literature on credence goods.

2.1. Off-Line Patient Perception

Several studies have examined the relationship between patient experience and clinical outcomes, such as mortality rate, 30-day readmission rate, clinical safety, and effectiveness of procedures. For example, Glickman et al. (2010) investigate whether patient satisfaction is associated with adherence to practice guidelines and clinical outcomes for acute myocardial infarctions. Boulding et al. (2011) find that higher overall patient satisfaction is associated with lower 30-day, risk-standardized readmission rates. A literature survey by Doyle et al. (2013) indicates that patient perception, in general, is positively associated with patient safety, clinical effectiveness, and adherence to clinical guidelines as well as objectively measured clinical outcomes, such as the repeat-visit rate or reduction of symptoms.

We differ from this stream in a number of ways. These studies typically rely on off-line surveys to solicit patient opinion, whereas we use patients' online reviews of physicians. The use of online reviews enables us to study free-form textual feedback through sentiment-mining and topic-modeling techniques. Further, many of these studies use hospital-level data to construct their clinical outcome variables, whereas we draw upon a patient-admission level data set that tracks inpatient admissions and discharges across multiple hospitals. This allows us to study the relationship between patient feedback and care quality at a much more granular level.

2.2. Online Content Generated by Healthcare Consumers

Our research framework is grounded in the burgeoning literature on health IT, particularly the branch of user-generated content in healthcare. Agarwal et al. (2010) provide a comprehensive review of the early health IT literature, noting the richness of research opportunities at the intersection of the internet and healthcare while highlighting the need for a closer examination of the quality of online health information. Recent research attempts to address these gaps. For example, Gao et al. (2012) examine online ratings of physicians and report a positive, albeit small, association between online ratings and physician characteristics, such as experience, board certification, and education as well as malpractice claims. In contrast, our focus is on the association between online reviews

and clinical outcomes. We control for physician characteristics as we intend to study whether online reviews can provide informative signals about future clinical outcomes beyond those already provided by such characteristics.

Gao et al. (2015) examine how perceived physician quality affects the likelihood of a physician being rated online and, conditional on getting rated, how perceived quality is reflected in online ratings. They observe that physicians who are rated as lower quality in off-line surveys are less likely to be rated online, suggesting that online ratings are not unduly negative. A key distinction between our paper and theirs is our focus on the ability of reviews to inform prospective patients about clinical outcomes, such as the readmission risk or ER visit rate.

Among closely related works, Gray et al. (2015) find no evidence of a relationship between online physician ratings and clinical measures, such as blood pressure or low-density lipoprotein. In contrast, Lu and Rui (2018), who examine the validity of physician ratings in the context of coronary artery bypass graft surgeries, observe that patients of cardiac surgeons with low ratings show a higher likelihood of mortality compared with those with high or no ratings. Likewise, Bardach et al. (2012) find a positive association between ratings of physicians on Yelp and clinical outcomes. In short, the evidence remains mixed and inconclusive.

Given such conflicting evidence and the distinctive nature of chronic disease care, it is imperative that we thoroughly reevaluate whether online reviews indeed signal a physician's clinical performance. To this end, we focus on individual physicians as opposed to hospitals and limit our analysis to only patients suffering from COPD, to eliminate any confounding factors that may vary from one condition to another. More importantly, recognizing that chronic disease management is long term in nature, we do not limit ourselves to a single hospital. Instead, we track each patient's journey across multiple hospitals over time to accurately estimate the readmission rate and other outcomes. We also go beyond star ratings and consider sentiments and topics expressed in the rich textual reviews. Finally, we perform a number of robustness checks to rule out alternative explanations, such as review bias, patient self-selection, and review manipulation. Interestingly, we do not find physician reviews to be reliable indicators of future clinical outcomes across different analyses. Our empirical findings indicate that the efficacy of online reviews is far from certain in the context of healthcare—potentially much less so in the context of chronic disease care with a credence component—and, thus, enrich prior literature on the usefulness of online physician reviews.

2.3. Online Reviews

Over the past two decades, the economics of consumer reviews has been extensively studied from various angles. One stream examines the supply side and addresses a wide variety of questions related to the production of reviews, such as word-of-mouth generation and social influence (Dellarocas and Narayan 2006, Moe and Trusov 2011, Berger and Iyengar 2012, Muchnik et al. 2013), evolution of ratings and number of reviews (Godes and Silva 2012), potential self-selection and other biases Gao et al. 2015, Chen et al. 2016), and fake reviews and intentional manipulation (Mayzlin et al. 2014). The demand side, that is, how such reviews are consumed in practice, also receives considerable attention. The general consensus is that potential consumers read and rely on such reviews, and product ratings/reviews do correlate with actual sales (Chevalier and Mayzlin 2006, Clemons et al. 2006, Chen and Xie 2008, Forman et al. 2008, Chintagunta et al. 2010, Sun 2012). More recently, researchers have focused on the issue of perceived and actual usefulness of online reviews, addressing questions such as what makes them appealing to potential consumers (Mudambi and Schuff 2010) or whether they are indeed useful in predicting quality (Lu and Rui 2018). Our study fits within this stream as it focuses on the ability of online reviews to capture care quality with a unique focus on chronic disease care.

2.4. Credence Goods

The term “credence” was originally proposed by Darby and Karni (1973) to characterize goods and services whose quality information is never completely revealed to consumers. This has several implications. Specifically, fraud and malpractice become likely; hiding information on quality may allow a seller to sell to consumers who may not have purchased if they had full information. For example, a healthcare provider may promote the most profitable treatment(s) to patients even when cheaper alternatives are available (Dulleck and Kerschbamer 2006).

Realizing that such inefficiencies are likely to exist in the case of credence goods, researchers also investigate economic remedies, such as liability, verification, reputation, and competition (Dulleck et al. 2011). More recently, it has been suggested that consumer reviews and word-of-mouth can play significant roles in the context of professional services that possess many credence elements (Gao et al. 2015). We augment this line of research by examining the case of COPD, a context that is significantly more credence than other nonchronic diseases.

3. Research Framework

Online reviews of physicians contain more information than just numeric (star) ratings. For example,

they often help prospective patients glean information about experiences of past patients regarding various aspects of physician care, including but not limited to the time spent with a patient, bedside manner, and the doctor’s ability to explain diagnoses and procedures involved. Some aspects of textual reviews, such as detailed accounts of procedures, may also provide useful cues about the clinical aspects of care. It is, therefore, not surprising that online reviews often influence patient choice, and prospective patients expect physicians with largely positive reviews to deliver better clinical outcomes (Hanauer et al. 2014, Burkle and Keegan 2015). However, there is only limited data-driven evidence that such expectations hold in the case of chronic disease care or, for that matter, any service with a significant credence component.

There is indeed some legitimacy to concerns about the reliability of online reviews of physicians. This is because a typical patient, who lacks comprehensive medical training, may not be well equipped to ascertain the clinical proficiency of a physician (Castaneda 2018). Although, in some cases, such as surgeries, a patient can ascertain the success or failure of the procedures performed, it is unlikely that chronic disease patients are able to assess the quality of care provided by their physicians just as easily. The prolonged nature of treatments and numerous interactions involved can make it extremely difficult for patients to accurately track and evaluate changes to their medical condition. In other words, the credence nature of chronic disease care could very well obfuscate patients’ judgments about the quality of care delivered by their physician(s). This leads to our primary research question (RQ):

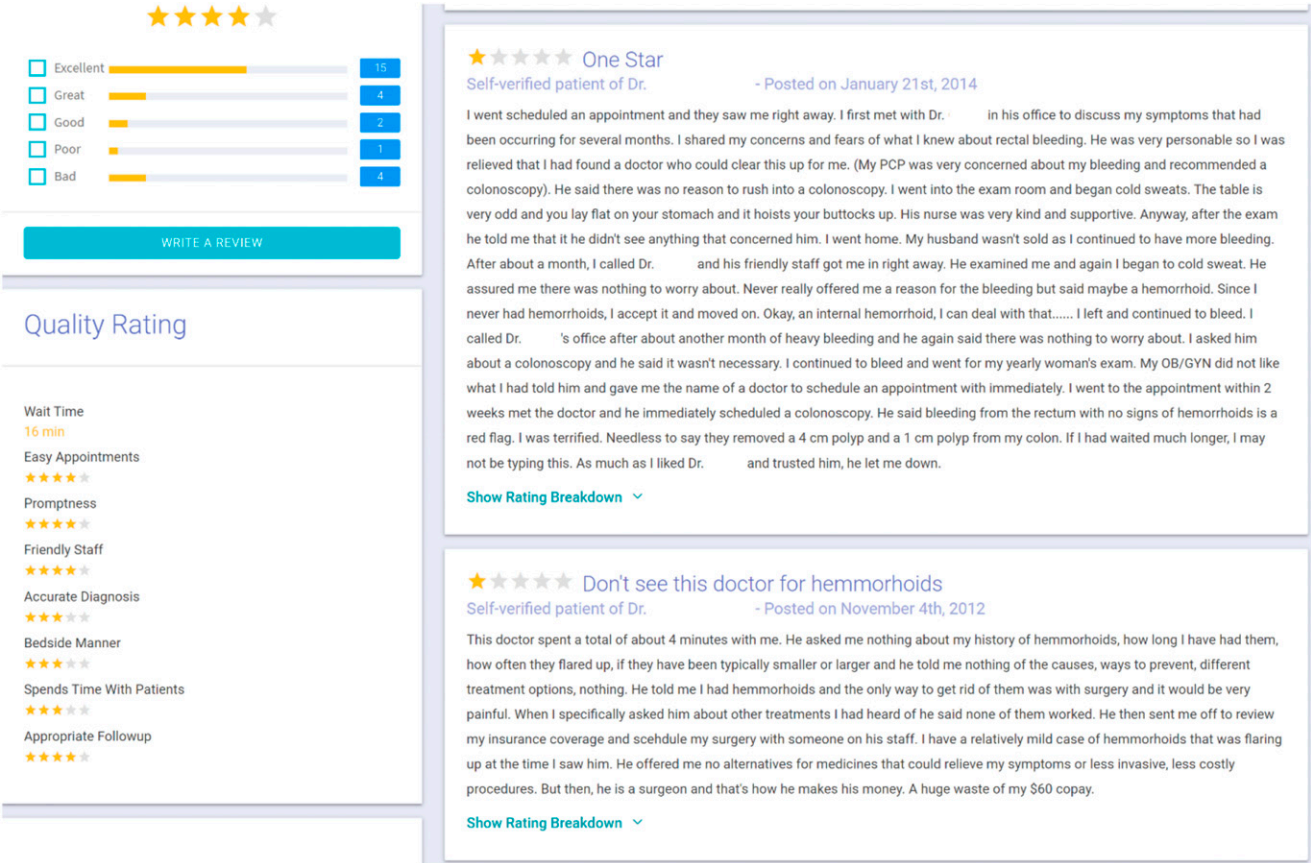
RQ1. Are online reviews reliable indicators in the case of chronic disease care?

a. Are physicians who receive better online reviews, in terms of sentiments expressed in textual reviews, more likely to deliver better clinical outcomes for their patients?

b. Are physicians who receive more positive online reviews, measured in terms of higher overall star ratings, more likely to deliver better clinical outcomes?

With respect to star ratings, patients rate their experience across multiple dimensions, some related to the physician and others related to their interactions with staff as shown in Figure 1. The bottom left corner of Figure 1 shows the seven dimensions tracked by Vitals.com. Dimensions such as accurate diagnosis, time spent with the patient, bedside manner, and appropriate follow-up are directly related to clinical aspects of care delivery. At the same time, however, patients can also rate their experience with respect to the ease of appointment scheduling and promptness and friendliness of the office staff. It is only natural to

Figure 1. (Color online) Screenshot of Online Reviews and Ratings from Vitals.com



surmise that these staff and nonclinical ratings may not contain the same signals on care quality as physician-related ones and, therefore, should be analyzed separately. Viewed another way, the aggregate ratings may not be reliable indicators of clinical outcomes even when physician-related dimensions are actually so. If this is indeed the case, online reviews could be useful even when the aggregate star rating turns out to be not. Hence, we ask the following:

RQ2. Can physician-related ratings serve as reliable indicators of their patients' clinical outcomes? That is, are physician-specific ratings more informative of clinical outcomes when compared with staff-specific ones?

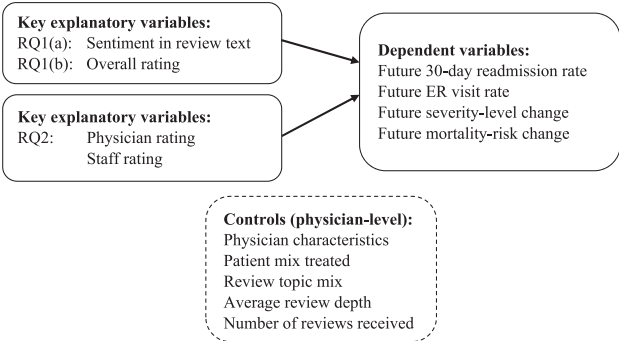
These research questions are depicted in the research framework shown in Figure 2. Our key explanatory variables include sentiments expressed in textual reviews and numeric star ratings, whereas dependent variables include actual clinical outcomes. We consider a total of four outcome variables: we use future 30-day readmission risk and future ER visit rate in our main analyses and use two additional ones, future severity-level and future mortality-risk change, in robustness analyses. Our primary aim is to study the association between these dependent and explanatory variables without any precept about the nature of

the relationship, that is, whether it is positive, negative, or insignificant (RQ1). In addition, we seek to understand whether star ratings on certain dimensions are informative even if the overall rating is not (RQ2).

3.1. Research Data

We obtained research data from two primary sources: (1) the Dallas Fort Worth Hospital Council (DFWHC) Research Foundation database and (2) Vitals.com, a publicly accessible health infomediary. The first data set is the source of our dependent or clinical outcome

Figure 2. Research Framework



variables, and the second one provides us with key explanatory variables.

The DFWHC data set spans a 10-year period from 2006 to 2015, from which we extracted approximately 630,000 inpatient admission–discharge records of COPD patients. Each inpatient record contains information on patient health conditions, demographics, diagnoses and procedures, and the identity of the physician. We tracked patient visits across multiple hospitals by matching the regional master patient index (REMPI), a unique identification number assigned to each patient. The REMPI enabled us to obtain patients' entire (inpatient) admission–discharge histories and accurately compute clinical outcomes based on their readmissions and ER visits across multiple hospitals in the North Texas region (Bao et al. 2020).

The second data set, collected from Vitals.com (during December 2015–January 2016), provides data on physician characteristics and online reviews, including textual reviews and star ratings. This data set spans nine years of online physician reviews, from 2007 to 2015. We integrated this data set with the DFWHC data by first rolling up each data set to the physician-quarter level (discussed in the following section) and then matching physician attributes across the two data sets. The attributes used for matching are the first name, middle initial (if available), last name, and location of the physician. The resulting unbalanced panel data set contains approximately 2,000 physicians⁴ and stretches over 35 time periods (quarters).

3.2. Dependent Variables

All clinical outcome variables are constructed from the DFWHC data. We first describe the ones used in the main analysis. At any given point in time and a focal physician, *Future30DayReadm* measures the proportion of future admissions (handled by the physician) that are followed by a readmission within 30 days of discharge. For a given physician and time period (each panel record), we calculated *Future30DayReadm* as follows. We first constructed a binary variable that equals one for a patient-admission record only if the patient's next admission (readmission) date is within 30 days of discharge, irrespective of the attending physician handling the readmission. Next, for each attending physician and quarter combination, we calculated the forward-looking average of this binary variable across future admissions (admissions in the same quarter or later). Intuitively, this average represents the *fraction* of all future admissions handled by the physician that eventually resulted in a 30-day readmission. The use of 30-day readmissions to measure quality of clinical care is consistent with its definition by the Centers for Medicare and Medicaid Services (CMS) (Bardhan et al. 2015). We constructed *FutureERVisit* in a similar manner—an admission was

assigned a value of one only if the patient visited an ER during the next admission (readmission), irrespective of the attending physician handling the readmission. We then calculated the forward-looking average of this binary variable across future admissions handled by the physician to obtain the proportion of future admissions that were succeeded by an admission involving an ER visit.

A few points are in order here. First, we use *Future30DayReadm* and *FutureERVisit* as our primary outcome variables because they represent key metrics monitored by the CMS under the scope of the Hospital Readmission Reduction Program for chronic disease management (Bardach et al. 2012, Agarwal et al. 2016).⁵ Second, these variables are fractions that range from zero to one with higher values representing worse clinical outcomes. Third, they are constructed in a forward-looking manner because our objective is to study whether reviews available at a given time are good indicators of future outcomes. Finally, it is possible to construct these variables differently and consider other outcomes as we show later in the robustness analyses.

3.3. Explanatory Variables

Our key explanatory variables are constructed based on data collected from Vitals.com. On this website, a patient can rate a physician on a five-point scale along seven dimensions in addition to writing a textual review; see Figure 1. The *OverallRating* of a physician at a given time is then calculated as the average of all seven dimensions across all reviews available until that point in time. Recall that, of the seven dimensions, three are related to services provided by the physician's staff, and the other four are mostly related to clinical care provided by the physician. Accordingly, we construct *StaffRating* by averaging the three dimensions related to the staff and *PhysicianRating* from the other four dimensions.

To measure patients' sentiments expressed in textual reviews, we constructed *SentimentScore* using sentiment analysis. Such analysis is often used to extract user opinions, sentiments, evaluations, and emotions from written language (Liu 2012). We used the classic bag-of-words sentiment analysis approach that employs a dictionary based on Nielsen (2011) and classifies words and phrases into four categories: very positive, positive, negative, and very negative. Using this dictionary, we first scored sentiment words in a review and then added those scores to obtain the overall sentiment score for the review.⁶ To compute the *SentimentScore* of a physician at any given time (quarter), we averaged sentiment scores across all reviews till that quarter. In the robustness section, we consider an alternative sentiment analysis tool that incorporates recent advancements in sentiment mining.

3.4. Control Variables

3.4.1. Review Topics. We control for the mix of topics underlying textual reviews received by each physician. To identify the latent topics, as described in OnlineAppendix B, we used the latent Dirichlet allocation (LDA) method (Blei et al. 2003, Tirunillai and Tellis 2014). The smallest number of topics that are reasonable in our case is four as shown in Table 1.

Next, we classified our reviews based on the underlying topics. Because a review may discuss multiple topics, we simply focused on the most prominent topic discussed in each review. Accordingly, for a given physician and quarter, *TopicSurgery* is the proportion of all textual reviews (written on or before that quarter) for which the prominent theme is the surgical competence of the physician's service. *TopicPhysician* is the proportion of reviews in which the prominent theme is interaction with the physician (time spent, treatments offered, and listening skills of the physician). *TopicPromptness* and *TopicOverallCare* represent the proportions for which the dominant theme is staff promptness and the overall care provided by the physician, respectively. These topic variables are needed as controls because reviews on a particular topic can be more negative than on the others. Typically, reviews concerning promptness are more negative vis-à-vis the rest; see Figure B1 in Online Appendix B for an in-depth discussion.

3.4.2. Review Depth and Volume. It is important to account for the depth or comprehensiveness of reviews (Mudambi and Schuff 2010). Hence, we constructed two variables, namely *ReviewWordsNum* and *SentimentVariance*. For each physician, *ReviewWordsNum*

represents the average length (in words) of the reviews up to the time of interest. Likewise, *SentimentVariance* is the average sentiment variance within a review⁷ over all reviews up to that time and accounts for the possibility that reviews with low levels of positive and negative word-of-mouth may have the same sentiment score as others with high levels of each (East et al. 2007). To control for review volume, we used the number of textual reviews received, *ReviewsNum*.

3.4.3. Patient Mix. To account for differences in patient mix across physicians, we employed several control variables that capture possible heterogeneity among physicians with respect to their patient populations (Lu and Rui 2018). For a physician, *LOS* is defined as the average length of hospital stay (discharge date minus the admission date) of patients. *MortMajExt* is defined as the proportion of admissions for which the patient has a major or extreme mortality risk, and *SevMajExt* is the proportion for which the severity level is major or extreme. *Expired* represents the proportion of patient admissions having a discharge status containing the term "expired." We also control for the proportion of admissions for which the patient was previously admitted at a different hospital or hospital system because patients' switching decisions may be related to changes in their health condition(s). *SwitchHosp* and *SwitchHospSys*, respectively, represent the proportions of such records for a physician. As Table 2 suggests, on average, these proportions are 36.7% and 28.9%, respectively, indicating that a significant number of patients in our sample switched hospitals and hospital systems.

Finally, we control for the ethnicity, gender, and age of patients under the care of a physician because they may influence the quality of physician-patient relationships (Ferguson and Candib 2002, Lu and Rui 2018). *EthnHisp* represents the proportion of admissions for which the patient is Hispanic and *RaceWhite* the proportion for which the patient is of Caucasian origin. *PtAge* records the average patient age. *Female* measures the proportion of patients that are female.

3.4.4. Physician Attributes. It is important to control for additional physician characteristics, such as experience (Gao et al. 2012). We use *VisitsNum*, which represents the number of patient admissions for a given physician up to the focal quarter as a proxy for physician experience in terms of the number of cases handled. Note that there is no need to separately control for the physician's experience in years as it is already subsumed by two-way fixed effects.

To control for prior clinical outcomes, we constructed *30DayReadm*, the average 30-day readmission rate for admissions handled by the physician in the past; we used

Table 1. Top Latent Topics Based on 20 Closest Word Stems

Topic	Surgery	Physician	Promptness	Overall care
1	surgeri	doctor	offic	staff
2	pain	patient	call	care
3	life	time	wait	recommend
4	back	year	appoint	great
5	year	good	time	feel
6	procedur	care	nurs	alway
7	sever	medic	back	question
8	work	problem	anoth	friend
9	result	ever	hour	high
10	right	know	rude	concern
11	surgeon	treat	day	profession
12	hospit	help	tri	answer
13	day	treatment	test	explain
14	husband	listen	minut	best
15	month	issu	room	experi
16	first	person	new	manner
17	thank	mani	insur	anyon
18	follow	physician	schedul	excel
19	abl	son	seem	famili
20	two	well	staff	love

Table 2. Descriptive Statistics

Type	Variable	Mean	Median	Standard deviation	Minimum	Maximum
Clinical outcome (dependent variable)	<i>Future30DayReadm</i>	0.056	0.000	0.161	0	1
	<i>FutureERVisit</i>	0.637	0.833	0.413	0	1
Ratings/sentiment (key explanatory variable)	<i>SentimentScore</i>	2.527	2.500	2.862	−23	23
	<i>OverallRating</i>	4.066	4.429	1.033	1	5
	<i>PhysicianRating</i>	4.063	4.550	1.133	1	5
	<i>StaffRating</i>	4.071	4.333	0.995	1	5
Review topics (control)	<i>TopicSurgery</i>	0.253	0.000	0.343	0	1
	<i>TopicPhysician</i>	0.269	0.101	0.341	0	1
	<i>TopicPromptness</i>	0.193	0.000	0.297	0	1
	<i>TopicOverallCare</i>	0.285	0.143	0.346	0	1
Review depth and volume (control)	<i>ReviewWordsNum</i>	63.753	55.000	43.525	4	626
	<i>SentimentVariance</i>	1.498	0.800	2.282	0	46
	<i>ReviewsNum</i>	4.396	2.500	9.380	1	507
Patient mix (control)	<i>LOS</i>	4.090	3.600	2.659	0	90
	<i>SevMajExt</i>	0.237	0.176	0.230	0	1
	<i>MortMajExt</i>	0.123	0.048	0.170	0	1
	<i>Expired</i>	0.012	0.000	0.042	0	1
	<i>SwitchHosp</i>	0.367	0.333	0.313	0	1
	<i>SwitchHospSys</i>	0.289	0.227	0.290	0	1
	<i>EthnHisp</i>	0.069	0.027	0.117	0	1
	<i>RaceWhite</i>	0.782	0.835	0.213	0	1
	<i>PtAge</i>	51.231	59.000	20.531	1	96
	<i>Female</i>	0.610	0.598	0.258	0	1
	<i>VisitsNum</i>	64.814	24.000	123.398	1	1,680
	<i>30DayReadm</i>	0.080	0.053	0.111	0	1
Physician attributes (control)	<i>ERVisit</i>	0.604	0.625	0.272	0	1

a similar procedure to construct *ERVisit*. These variables as well as other explanatory variables are backward-looking as opposed to forward-looking outcomes, such as *Future30DayReadm* or *FutureERVisit*.

4. Estimation Models and Results

In this section, we first address our research questions, RQ1 and RQ2. We then report results from additional analyses necessary to rule out alternative explanations and endogeneity concerns.

4.1. Reliability of Sentiment Score and Overall Rating

We use the following two-way fixed-effects panel model to empirically test RQ1:

$$\text{ClinicalOutcome}_{i,t \rightarrow T} = \text{Review}_{i,0 \rightarrow t} \beta + \text{Controls}_{i,0 \rightarrow t} \delta + \mu_i + \nu_t + \varepsilon_{it}.$$

$\text{ClinicalOutcome}_{i,t \rightarrow T}$ denotes the forward-looking outcome variable, which is operationalized by either *Future30DayReadm* or *FutureERVisit*. It represents the average performance of physician *i* from time (quarter) *t* onward. As far as explanatory variables are concerned, $\text{Review}_{i,0 \rightarrow t}$ represents physician *i*'s online stock of reputation at time *t* and is calculated based on all reviews written up until quarter *t*. More specifically, to answer RQ1(a), we consider *SentimentScore* as a proxy for online reputation, and *OverallRating* is

used for RQ1(b). $\text{Controls}_{i,0 \rightarrow t}$ is the vector of all control variables, constructed in a similar manner based on reviews and admissions data for physician *i* up until *t*.

The use of a forward-looking outcome variable, along with backward-looking explanatory variables, is necessary to examine whether a physician's cumulative review score at a given time can inform prospective patients about care quality to be expected in the future. Doing so also mitigates possible biases that may arise from simultaneity between our explanatory and outcome variables. In addition, μ_i , the time-invariant fixed effect for physician *i*, takes into account unobserved time-invariant heterogeneity across physicians. Likewise, ν_t , the time fixed effect for quarter *t*, accounts for unobserved time shocks. Finally, ε_{it} is the idiosyncratic error term.

Table 3 shows the coefficient estimates. Throughout the paper, we report results based on heteroscedasticity-consistent estimates of standard errors, clustered by physicians.⁸ The first two columns of Table 3 show the regression estimates of *Future30DayReadm* on *SentimentScore* and *OverallRating*, respectively, whereas the last two columns report the same for *FutureERVisit*.

Because higher values of the outcome variables indicate worse clinical outcomes, the coefficients of the main explanatory variables—*SentimentScore* or *OverallRating*—need to be negative and statistically significant for online reviews to be useful to prospective patients. However, as we can observe, online reviews

Table 3. Two-Way Fixed Effect Estimation for RQ1

Variables	Future30DayReadm		FutureERVisit	
	RQ1a	RQ1b	RQ1a	RQ1b
<i>SentimentScore</i>	0.002 ⁺ (0.001)		0.002 (0.002)	
<i>OverallRating</i>		0.001 (0.002)		0.004 (0.005)
<i>TopicSurgery</i>	−0.003 (0.008)	−0.005 (0.008)	−0.037 ⁺ (0.021)	−0.043 [*] (0.021)
<i>TopicPromptness</i>	−0.002 (0.008)	−0.002 (0.008)	−0.010 (0.020)	−0.011 (0.021)
<i>TopicOverallCare</i>	0.001 (0.007)	0.002 (0.007)	−0.048 [*] (0.020)	−0.050 [*] (0.020)
<i>ReviewWordsNum</i>	0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>SentimentVariance</i>	−0.002 [*] (0.001)	−0.001 (0.001)	0.003 (0.002)	0.005 [*] (0.002)
<i>ReviewsNum</i>	−0.000 ⁺ (0.000)	−0.000 ⁺ (0.000)	0.000 (0.000)	0.000 (0.000)
<i>LOS</i>	−0.000 (0.001)	−0.000 (0.001)	−0.002 (0.007)	−0.001 (0.008)
<i>SevMajExt</i>	0.030 (0.048)	0.048 (0.046)	−0.044 (0.108)	−0.045 (0.108)
<i>MortMajExt</i>	−0.122 [*] (0.055)	−0.140 [*] (0.054)	−0.115 (0.136)	−0.118 (0.137)
<i>Expired</i>	−0.206 (0.296)	−0.211 (0.298)	0.093 (0.363)	0.105 (0.367)
<i>SwitchHosp</i>	0.024 (0.018)	0.021 (0.018)	−0.022 (0.051)	−0.023 (0.052)
<i>SwitchHospSys</i>	−0.009 (0.020)	−0.007 (0.02)	0.024 (0.055)	0.026 (0.055)
<i>EthnHisp</i>	0.002 (0.035)	0.005 (0.035)	−0.103 (0.124)	−0.092 (0.124)
<i>RaceWhite</i>	0.012 (0.022)	0.014 (0.023)	0.056 (0.069)	0.056 (0.069)
<i>PtAge</i>	0.000 (0.001)	0.001 (0.001)	0.001 (0.002)	0.001 (0.002)
<i>Female</i>	0.009 (0.032)	0.009 (0.032)	−0.074 (0.074)	−0.079 (0.075)
<i>VisitsNum</i>	−0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>30DayReadm</i>			−0.013 (0.126)	−0.012 (0.127)
<i>ERVisit</i>	0.002 (0.027)	−0.001 (0.027)		
<i>R²</i>	0.8639	0.8645	0.9047	0.9048
Root mean square error	0.0603	0.0598	0.1302	0.1301
Number of physicians	2,028	1,998	1,975	1,946

Note. Standard errors within parentheses.

⁺*p* < 0.1, ^{*}*p* < 0.05, ^{**}*p* < 0.01, ^{***}*p* < 0.001.

of physicians are not reliable indicators of future clinical outcomes. These results point to the fact that chronic disease care is likely different from search and experience goods discussed in the prior literature and that it is presumptuous to expect online reviews to be equally useful in a context replete with credence characteristics.

4.2. Components of Star Rating

To answer RQ2, we use the same regression model with a minor change: $Review_{i,0 \rightarrow t}$ is now replaced with either *PhysicianRating* or *StaffRating*. The results, as shown in Table 4, indicate that the coefficients of *PhysicianRating* and *StaffRating* are insignificant for both clinical outcomes. Hence, physician and staff ratings are not reliable indicators either as far as their ability to signal care quality is concerned.

To examine the relationship between ratings and clinical outcomes at a more granular level, we ran the same model separately for each rating component. Recall that these components are the star ratings on the following dimensions: diagnosis, follow-up, time spent with patients, bedside manner of the physician, ease of appointment scheduling, promptness of the staff, and courteousness of the staff. We omit the detailed results for the sake of brevity. Overall, for both outcome variables, the coefficients of individual

ratings are insignificant. This further corroborates that neither online ratings of physicians and their staff nor sentiments expressed in textual reviews are useful indicators of clinical outcomes experienced by their patients.

4.3. Endogeneity

Next, we address possible endogeneity concerns to ensure that our results are not biased.

4.3.1. Self-Selection by Patients. First, we consider the possibility that patients who are sicker to begin with may deliberately seek physicians who are likely to deliver better clinical outcomes (Dranove et al. 2003). Although we partially address this issue by controlling for the patient mix treated by each physician, further analysis is warranted to ensure that such self-selection is not biasing our results. To that end, we here perform a subpanel analysis focusing only on nonelective admissions, that is, admissions for which patients have no control over their choice of physicians. This idea of focusing on patients who do not have a chance to select their physicians is based on similar logic by Lu and Rui (2018). To elaborate, almost all admission records in our data set are categorized as either emergency, urgent, or elective visits. The first two categories are considered nonelective,

Table 4. Two-Way Fixed Effects Estimation with Physician and Staff Ratings

Variables	<i>Future30DayReadm</i>		<i>FutureERVisit</i>	
	RQ2	RQ2	RQ2	RQ2
<i>PhysicianRating</i>	−0.000 (0.002)		0.003 (0.005)	
<i>StaffRating</i>		0.000 (0.002)		0.003 (0.006)
<i>TopicSurgery</i>	−0.006 (0.008)	−0.004 (0.008)	−0.044* (0.021)	−0.038 ⁺ (0.022)
<i>TopicPromptness</i>	−0.004 (0.008)	−0.002 (0.008)	−0.014 (0.021)	−0.015 (0.021)
<i>TopicOverallCare</i>	0.002 (0.007)	0.003 (0.008)	−0.052** (0.020)	−0.047* (0.020)
<i>ReviewWordsNum</i>	−0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>SentimentVariance</i>	−0.001 (0.001)	−0.001 (0.001)	0.005* (0.002)	0.005* (0.002)
<i>ReviewsNum</i>	−0.000 ⁺ (0.000)	−0.000 ⁺ (0.000)	0.000 (0.000)	0.000 (0.000)
<i>LOS</i>	−0.000 (0.001)	−0.000 (0.001)	−0.000 (0.008)	−0.001 (0.008)
<i>SevMajExt</i>	0.049 (0.046)	0.047 (0.046)	−0.035 (0.108)	−0.047 (0.111)
<i>MortMajExt</i>	−0.140* (0.055)	−0.143** (0.055)	−0.128 (0.137)	−0.118 (0.139)
<i>Expired</i>	−0.207 (0.298)	−0.211 (0.300)	0.109 (0.366)	0.112 (0.368)
<i>SwitchHosp</i>	0.020 (0.018)	0.020 (0.018)	−0.025 (0.052)	−0.022 (0.051)
<i>SwitchHospSys</i>	−0.007 (0.02)	−0.007 (0.020)	0.029 (0.055)	0.026 (0.055)
<i>EthnHisp</i>	0.004 (0.035)	0.004 (0.035)	−0.095 (0.124)	−0.099 (0.124)
<i>RaceWhite</i>	0.014 (0.023)	0.011 (0.023)	0.055 (0.069)	0.053 (0.071)
<i>PtAge</i>	0.001 (0.001)	0.001 (0.001)	0.001 (0.002)	0.000 (0.002)
<i>Female</i>	0.009 (0.032)	0.009 (0.033)	−0.084 (0.074)	−0.090 (0.077)
<i>VisitsNum</i>	−0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>30DayReadm</i>			−0.014 (0.127)	−0.010 (0.127)
<i>ERVisit</i>	−0.002 (0.027)	−0.000 (0.028)		
<i>R</i> ²	0.8649	0.8614	0.9048	0.9045
Root mean square error	0.0597	0.0600	0.1302	0.1306
Number of physicians	1,989	1,980	1,937	1,929

Note. Standard errors within parentheses.

⁺*p* < 0.1, **p* < 0.05, ***p* < 0.01, ****p* < 0.001.

and they comprise about 80% of all admissions, and elective admissions comprise 20%. We mitigate possible bias resulting from patient self-selection by excluding the elective admissions from our final panel construction. As shown in Table 5, our results remain qualitatively similar to our main results in Table 3 even after elective admissions are excluded from our analysis.

4.3.2. Omitted Variables. To address endogeneity concerns that might stem from omission of variables correlated with both the explanatory and dependent variables, we need to construct two instrument variables (IV). To this end, we utilized a binary variable *Awards* that equals one for physicians who have received one or more awards in the following categories: patient satisfaction, patients' choice, compassion, patient relation, or punctuality. Approximately 60% of our physicians won at least one such award. Note that we explicitly exclude the category of awards that may be directly related to care quality, such as the Outstanding Physician Award by the Quality Management Committee. To construct the two IVs necessary for our panel analysis, we used the interaction of *Awards* × *SentimentScoreLag* as the IV for *SentimentScore* and *Awards* × *OverallRatingLag* for *OverallRating*, where *SentimentScoreLag* and *OverallRatingLag* are lagged

values (lagged by four quarters) of *SentimentScore* and *OverallRating*, respectively.

The rationale for using these instruments is that physician awards are conspicuously displayed, and award-winning physicians are likely to have better online reviews compared with nonwinners. However, it is very unlikely that these awards would directly drive actual clinical outcomes that a physician delivers to patients. These IVs are, in fact, quite strong; the results from the first stage indicate that the instruments are strongly associated with the review variables. Further, to test the validity of the IVs, we conducted exogeneity and exclusion tests using the Sargan–Hansen test for over-identification. The results of the tests indicate the IVs are indeed valid; refer to Online Appendix C for further details. The second-stage results are shown in Table 6. There is still no evidence that physicians who receive more favorable textual comments or higher star ratings compared with their peers deliver better clinical outcomes to their patients.

4.3.3. Review Manipulation. Physicians who do not receive good patient ratings might have incentives to engage in review manipulation (Lagu et al. 2010), which can confound our analysis. To investigate potential review manipulation, we obtained additional data on online reviews from Yelp. Yelp flags potentially

Table 5. Two-Way Fixed Effects Estimation to Address Patient Self-Selection

Variables	<i>Future30DayReadm</i>		<i>FutureERVisit</i>	
	RQ1a	RQ1b	RQ1a	RQ1b
<i>SentimentScore</i>	0.002* (0.001)		0.000 (0.002)	
<i>OverallRating</i>		0.002 (0.002)		0.004 (0.005)
<i>TopicSurgery</i>	0.003 (0.008)	0.001 (0.008)	−0.024 (0.021)	−0.030 (0.020)
<i>TopicPromptness</i>	−0.000 (0.008)	0.002 (0.008)	−0.023 (0.019)	−0.018 (0.019)
<i>TopicOverallCare</i>	−0.000 (0.008)	−0.000 (0.008)	−0.040* (0.018)	−0.043* (0.018)
<i>ReviewWordsNum</i>	−0.000 (0.000)	−0.000+ (0.000)	−0.000 (0.000)	−0.000 (0.000)
<i>SentimentVariance</i>	−0.001 (0.001)	0.000 (0.001)	0.003 (0.002)	0.003 (0.002)
<i>ReviewsNum</i>	−0.000 (0.000)	−0.000 (0.000)	−0.000 (0.000)	−0.000 (0.000)
<i>LOS</i>	0.001 (0.002)	0.001 (0.002)	−0.022* (0.011)	−0.021+ (0.011)
<i>SevMajExt</i>	0.027 (0.054)	0.051 (0.051)	0.056 (0.111)	0.061 (0.112)
<i>MortMajExt</i>	−0.125+ (0.065)	−0.149* (0.064)	−0.111 (0.133)	−0.116 (0.134)
<i>Expired</i>	−0.098 (0.291)	−0.111 (0.296)	−0.006 (0.399)	0.103 (0.394)
<i>SwitchHosp</i>	0.021 (0.020)	0.016 (0.019)	0.030 (0.069)	0.026 (0.069)
<i>SwitchHospSys</i>	0.008 (0.026)	0.011 (0.026)	−0.046 (0.070)	−0.036 (0.070)
<i>EthnHisp</i>	0.007 (0.034)	0.008 (0.034)	−0.064 (0.144)	−0.032 (0.142)
<i>RaceWhite</i>	0.021 (0.024)	0.025 (0.024)	0.018 (0.074)	0.015 (0.074)
<i>PtAge</i>	−0.001 (0.001)	−0.001 (0.001)	0.003 (0.002)	0.003 (0.002)
<i>Female</i>	0.031 (0.035)	0.032 (0.035)	−0.047 (0.091)	−0.050 (0.092)
<i>VisitsNum</i>	−0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>30DayReadm</i>			−0.032 (0.129)	−0.005 (0.128)
<i>ERVisit</i>	0.020 (0.030)	0.017 (0.030)		
<i>R²</i>	0.8792	0.8808	0.9175	0.9180
Root mean square error	0.0534	0.0525	0.1095	0.1092
Number of physicians	1,766	1,737	1,701	1,673

Note. Standard errors within parentheses.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

manipulated reviews of physicians and labels them as “not recommended” (Luca and Zervas 2016).⁹ Reviews that Yelp finds trustworthy are labeled “recommended.” Although Yelp’s algorithm is not a perfect classifier, it provides us with a better understanding of reviews that are impacted by manipulation and those that are not—fewer “not recommended” reviews for a physician implies a lower likelihood that manipulation has an impact on the physician’s reputation. To rule out manipulation as a probable cause of bias, we deployed a subsample analysis based solely on physicians for whom there are no “not recommended” Yelp reviews. The results from this analysis are shown in Table 7.

We observe from Table 7 that coefficients of the two key explanatory variables—*SentimentScore* and *OverallRating*—are qualitatively similar to their counterparts in Table 3. Only the coefficient of *SentimentScore* in the first column is positive and significant at $p < 0.05$ (indicating worse outcomes), and all others are insignificant. Hence, even after accounting for review manipulation, our results suggest that online reviews of physicians are not reliable indicators of their patients’ clinical outcomes.

4.3.4. Review Bias. We now examine the possibility of a review bias, the possibility that dissatisfied patients

are more likely to write online reviews to vent their frustration (Chen et al. 2016). If such behavior is systematic, it might unduly affect the valence of online comments and ratings, eventually biasing the coefficient estimates of our key explanatory variables. We approach the issue as follows. If patients indeed behave this way, we can expect systematic differences between the frequencies with which physicians delivering different levels of clinical outcomes actually receive reviews. Specifically, physicians who delivered worse outcomes would receive reviews more frequently. We investigated whether this is the case by examining the association between past clinical outcomes delivered by a physician and the future intensity of reviews. In other words, is there evidence to suggest that physicians with higher *30DayReadm* and *ERVisit* measures at a certain times were reviewed more frequently thereon?

Our results are as shown in Table 8. The insignificant coefficients in the first two rows indicate a lack of association between future review frequency (*FutureReviewNumIntensity*) and past clinical outcomes (*30DayReadm* and *ERVisit*). Hence, there is no evidence to suggest that there is review bias in our context, which is consistent with prior literature (Gao et al. 2015).

Table 6. Two-Way Fixed Effects Estimation with Instrumental Variables

Variables	<i>Future30DayReadm</i>		<i>FutureERVisit</i>	
	RQ1a	RQ1b	RQ1a	RQ1b
<i>SentimentScore</i>	0.003 (0.004)		0.020 ⁺ (0.012)	
<i>OverallRating</i>		−0.022 (0.014)		0.040 (0.031)
<i>TopicSurgery</i>	−0.001 (0.010)	−0.008 (0.009)	−0.016 (0.030)	−0.035 (0.028)
<i>TopicPromptness</i>	−0.005 (0.014)	−0.044 ⁺ (0.023)	0.032 (0.038)	0.038 (0.051)
<i>TopicOverallCare</i>	0.004 (0.008)	0.010 (0.008)	−0.070 ^{**} (0.026)	−0.075 ^{**} (0.026)
<i>ReviewWordsNum</i>	0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>SentimentVariance</i>	−0.002 (0.002)	0.000 (0.001)	−0.008 (0.007)	0.004 (0.003)
<i>ReviewsNum</i>	−0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>LOS</i>	0.002 (0.003)	0.003 (0.004)	−0.004 (0.009)	−0.007 (0.012)
<i>SevMajExt</i>	0.034 (0.064)	0.036 (0.065)	−0.072 (0.154)	−0.089 (0.153)
<i>MortMajExt</i>	−0.111 (0.070)	−0.135 ⁺ (0.070)	−0.174 (0.195)	−0.165 (0.196)
<i>Expired</i>	−0.476 (0.488)	−0.389 (0.499)	−0.566 (0.771)	−0.707 (0.796)
<i>SwitchHosp</i>	0.034 (0.030)	0.032 (0.030)	0.063 (0.068)	0.069 (0.068)
<i>SwitchHospSys</i>	−0.012 (0.032)	−0.018 (0.031)	−0.062 (0.076)	−0.061 (0.076)
<i>EthnHisp</i>	0.008 (0.055)	0.007 (0.056)	−0.050 (0.184)	−0.027 (0.182)
<i>RaceWhite</i>	0.003 (0.028)	−0.002 (0.028)	0.087 (0.098)	0.088 (0.098)
<i>PtAge</i>	0.000 (0.001)	0.001 (0.001)	−0.002 (0.003)	−0.002 (0.003)
<i>Female</i>	−0.002 (0.049)	−0.005 (0.049)	−0.067 (0.113)	−0.068 (0.114)
<i>VisitsNum</i>	−0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>30DayReadm</i>			0.207 (0.227)	0.168 (0.230)
<i>ERVisit</i>	0.003 (0.040)	−0.006 (0.040)		
<i>R²</i>	0.8880	0.8883	0.9200	0.9203
Root mean square error	0.0558	0.0552	0.1227	0.1225
Number of physicians	1,888	1,855	1,846	1,814

Note. Standard errors within parentheses.

⁺*p* < 0.1, ^{*}*p* < 0.05, ^{**}*p* < 0.01, ^{***}*p* < 0.001.

5. Robustness Checks

We now discuss robustness of our results with respect to the choice and construction of clinical outcomes, choice of estimation model, and finally text-mining methodology.

5.1. Alternative Specification of ER Visit Rate

Recall that we defined the ER visit rate as the proportion of admissions for which the following admission (i.e., readmission) involved a visit to an emergency room. This definition does not place any restriction on the date of the immediate readmission and counts it as an ER visit even if it occurred long after the original discharge. However, some may argue that a readmission occurring long after the discharge does not accurately reflect the quality of care provided during the index admission. Hence, we now employ an alternative specification that only counts the proportion of admissions followed by an ER visit within 90 days of the discharge and, accordingly, introduce a new outcome variable *Future90DayERVisit* (as well as its backward-looking explanatory counterpart *90DayERVisit*). As evident from Table A1 in Online Appendix A, replacing *FutureERVisit* and *ERVisit* with these new variables does not change our results qualitatively; the key findings from Table A1 are similar to those in Table 3.

5.2. Alternative Panel Construction

In our main analysis, for period *t*, all explanatory variables are constructed based on data up until *t*, and the dependent variables are constructed based on data from time *t* onward. A concern with this approach could be that reviews written long back may not be informative of future clinical outcomes. Likewise, outcomes delivered far out in the future may not have any connection with reviews written today. To address the concern, we here construct the dependent and independent variables based on one-year time windows. In other words, all variables are now based on their average values for a one-year period, and the explanatory variables are lagged by one year. The results, as reported in Table A2 (Online Appendix A), are similar to those reported in Table 3. Hence, our earlier insights are robust to this alternative panel construction.

5.3. Other Clinical Outcomes

Earlier, we considered two clinical outcome variables, *Future30DayReadm* and *FutureERVisit*. To ensure that our main insights are not restricted to this choice, we here consider two additional outcomes, *Future30DaySeverityTransition* and *Future30DayMortalityTransition*. The intuition behind these outcomes is as follows. In addition to studying 30-day readmissions, we can also measure any changes in patients' severity

Table 7. Two-Way Fixed Effects Estimation to Address Review Manipulation

Variables	<i>Future30DayReadm</i>		<i>FutureERVisit</i>	
	RQ1a	RQ1b	RQ1a	RQ1b
<i>SentimentScore</i>	0.003* (0.001)		0.004 (0.004)	
<i>OverallRating</i>		0.004 (0.003)		0.012 (0.010)
<i>TopicSurgery</i>	0.005 (0.013)	−0.001 (0.013)	0.019 (0.041)	0.007 (0.04)
<i>TopicPromptness</i>	−0.014 (0.015)	−0.010 (0.014)	−0.055 (0.039)	−0.053 (0.041)
<i>TopicOverallCare</i>	0.003 (0.016)	0.001 (0.016)	−0.031 (0.036)	−0.034 (0.036)
<i>ReviewWordsNum</i>	0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>SentimentVariance</i>	−0.003* (0.001)	−0.001 (0.001)	−0.002 (0.004)	0.000 (0.004)
<i>ReviewsNum</i>	−0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>LOS</i>	0.001 (0.003)	0.001 (0.006)	−0.004 (0.009)	−0.011 (0.014)
<i>SevMajExt</i>	0.142 (0.128)	0.152 (0.130)	0.191 (0.251)	0.203 (0.253)
<i>MortMajExt</i>	−0.180 (0.136)	−0.201 (0.137)	−0.545* (0.260)	−0.552* (0.265)
<i>Expired</i>	−0.614 (0.688)	−0.676 (0.699)	−0.921 (1.133)	−0.814 (1.102)
<i>SwitchHosp</i>	0.031 (0.039)	0.027 (0.040)	−0.008 (0.097)	−0.005 (0.097)
<i>SwitchHospSys</i>	−0.024 (0.038)	−0.023 (0.038)	−0.045 (0.101)	−0.044 (0.102)
<i>EthnHisp</i>	−0.078 (0.069)	−0.082 (0.069)	0.280 (0.239)	0.3000 (0.236)
<i>RaceWhite</i>	−0.059 (0.045)	−0.058 (0.044)	0.126 (0.179)	0.137 (0.180)
<i>PtAge</i>	−0.000 (0.001)	−0.000 (0.001)	0.003 (0.004)	0.003 (0.004)
<i>Female</i>	−0.054 (0.067)	−0.049 (0.069)	−0.347+ (0.182)	−0.343+ (0.180)
<i>VisitsNum</i>	−0.000 (0.000)	−0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
<i>30DayReadm</i>			0.564* (0.270)	0.556* (0.274)
<i>ERVisit</i>	0.018 (0.067)	0.012 (0.068)		
<i>R²</i>	0.8906	0.8934	0.9074	0.9083
Root mean square error	0.0564	0.0554	0.1302	0.1297
Number of physicians	573	568	557	552

Note. Standard errors within parentheses.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

and mortality risk from the time of the previous admission. Such measures can reflect the quality of care received during the previous admission as they represent disease progress between the times of admission and readmission. Accordingly, for each physician and time period, we construct *Future30DaySeverityTransition* and *Future30DayMortalityTransition* by computing the forward-looking averages of changes in severity and mortality risk across all future 30-day readmissions attributed to the physician's care. These variables are also continuous, and higher values indicate worse outcomes. The results shown in Table A3 indicate that our original results about the lack of efficacy of online reviews extend to the new outcome variables. Specifically, the coefficients of *SentimentScore* and *OverallRating* are statistically insignificant for both new outcomes.

5.4. Digital Divide

Many COPD patients are senior citizens, and they could be averse to participating in online forums. The same could be true for very young patients as well. If this is indeed the case, online reviews may be less informative of clinical outcomes of patients in very senior or very young age groups. To account for this possibility, we now employ a subpanel analysis based on patient age. Because the mean age of our patients

(*PtAge*) is 51 years, we restrict the analysis to only those physicians whose average patient's age is at most 51 years but no less than 18 years ($18 \leq PtAge \leq 51$). In fact, a substantial number of physicians, approximately 22%, fall into this category. The regression results based on this subsample are shown in Table A4. There is still no indication that reviews

Table 8. Two-Way Fixed Effects Estimation for Review Bias

Variables	<i>FutureReviewNumIntensity</i>
<i>30DayReadm</i>	0.207 (0.155)
<i>ERVisit</i>	−0.070 (0.092)
<i>LOS</i>	−0.003 (0.003)
<i>SevMajExt</i>	0.232* (0.110)
<i>MortMajExt</i>	−0.130 (0.137)
<i>Expired</i>	0.045 (0.492)
<i>SwitchHosp</i>	0.090 (0.067)
<i>SwitchHospSys</i>	−0.077 (0.074)
<i>EthnHisp</i>	0.142 (0.163)
<i>RaceWhite</i>	0.052 (0.100)
<i>PtAge</i>	0.002 (0.003)
<i>Female</i>	−0.114 (0.128)
<i>VisitsNum</i>	0.000 (0.000)
<i>R²</i>	0.8407
Root mean square error	0.2370
Number of physicians	1,694

Note. Standard errors within parentheses.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

are useful for physicians who primarily treat certain age groups.

5.5. Estimation without Physician Fixed Effects

One concern about our estimation model could be that review scores and clinical outcomes for a given physician may not change significantly over time. In such a scenario, physician fixed effects may subsume a portion of the association, if any, between the review scores and clinical outcomes. To address this issue, we now run our earlier regressions without physician fixed effects. The coefficients in the first two rows of Table A5 are statistically insignificant, implying that our insight about a lack of usefulness of online reviews is robust to this approach.

5.6. Nonlinear Relationships

To account for possible nonlinear effects, we simply rerun our original model with additional quadratic regression terms, the squared values of *SentimentScore* and *OverallRating*. The findings with respect to *SentimentScore* and *OverallRating*, as shown in Table A6, are similar to those in Table 3. Also, the coefficients of *SentimentScoreSquared* and *OverallRatingSquared* are both statistically insignificant, suggesting that nonlinear effects are not a concern in our context.

5.7. Alternative Sentiment-Scoring Algorithm

Often, to correctly identify the sentiments expressed in a word, we need to consider other words in the vicinity of that sentiment word using text-mining techniques with *n*-grams (Liu 2012). For example, the word “good” is a positive-sentiment word, but when used in “not good,” it becomes a negative-sentiment word. To account for such effects, we here use the *Sentiment* function in the sentiment-mining package in R called *SentimentR*, which includes various types of valence shifters, including negators, amplifiers, and deamplifiers.¹⁰ By default, it treats five words before and two words after each polarized word as possible valence shifters. We construct an alternative sentiment score variable, *SentimentScoreAlt*, using this tool. The results based on this new sentiment metric are shown in Table A7 in Online Appendix A. It is evident that the results from our earlier analysis are robust to this alternative sentiment-mining technique.

5.8. Reliability and Review Topics

Online reviews may not necessarily emphasize the procedural side of care delivery. For example, patients’ perceptions, as reflected in their textual comments, may overemphasize nonclinical factors, such as flexibility in scheduling, promptness and courteousness of staff, responsiveness and professionalism of the medical team, etc. These factors are not necessarily

indicative of the quality of clinical care provided and may not be accurate indicators of future clinical outcomes. We now investigate whether topics discussed in text reviews moderate—either suppress or amplify—their reliability, if any. Specifically, are text reviews any more reliable for physicians whose comments have a greater emphasis on topics such as physician and surgery?

To answer this question, we need to conduct topic modeling at a level that is more granular than that of a review. The reason is that, within a textual review, there can be multiple sentences expressing different underlying themes. In order to account for these latent themes or topics at the sentence level, we classify the sentences in review texts into the topics discussed earlier, namely physician, surgery, promptness, and overall care. For each physician, *SentPhysician* denotes the proportion of sentences that discuss the topic of the physician; this proportion is computed based on reviews until time *t*. *SentSurgery*, *SentPromptness*, and *SentOverallCare* are then determined similarly.

Because we need to examine the moderating effects of topics on sentiments at the sentence level, we also conduct sentiment scoring at the sentence level. To this end, we construct a new variable, *SentenceScore*, using sentiment mining at the sentence level and then averaging over review sentences up until time *t*. For sentiment mining, we have used the tool discussed in Section 5.7. To investigate possible moderation effects, we consider interaction terms between *SentenceScore* and the topic variables from earlier. Because the sum of *SentPhysician*, *SentSurgery*, *SentPromptness*, and *SentOverallCare* is exactly one, we need only three of these topic variables in our final analysis. Table A8 shows the results of our two-way fixed effects estimation.

In Table A8, the coefficient of *SentenceScore* is either insignificant or significant and positive for both *Future30DayReadm* and *FutureERVisit*. Hence, even if *SentPhysician* = 1 and *SentSurgery* = *SentPromptness* = *SentOverallCare* = 0 were to hold for a given physician, the sentiment score would not be informative of the physician’s future clinical outcomes. Further, the sum of coefficients of *SentenceScore* and *SentenceScore* × *SentSurgery* is statistically insignificant for both *Future30DayReadm* and *FutureERVisit*, implying that reviews would not be reliable even if they were entirely about the topic of surgery. Overall, there is no evidence that reviews emphasizing certain topics are useful indicators.

5.9. Physicians with No Ratings

In this section, we present another robustness check with a panel that includes physicians and time periods for which no reviews are available. This approach is similar to that of Lu and Rui (2018) and uses categorical rating variables. We employ one-year time

windows for this analysis. Specifically, for each record in the panel, we construct three binary variables to indicate whether the physician is highly rated (*HighRatingBinary* = 1), lowly rated (*LowRatingBinary* = 1), or not rated (*NoRatingBinary* = 1) based on the average rating received by the physician in that year. We treat records with median (approximately 4.6 for this panel with one-year windows) or higher ratings as highly rated, those with lower ratings as lowly rated, and the rest as not rated. Because we now have records with no ratings, it is no longer possible to use many of the review controls used earlier. The results from this analysis are shown in Table A9. Evidently, compared with patients of lowly rated physicians, patients of highly rated or not rated physicians do not experience better clinical outcomes. In sum, this analysis corroborates our earlier findings.

5.10. Dynamic Panel Estimation

As a final robustness check, we utilize dynamic panel estimation to account for the possibility that past clinical outcomes may also be driving future clinical outcomes. To carry out this analysis, we first include, as an additional explanatory variable, the lagged value of the dependent variable in a physician fixed-effects model. We then first-difference all variables (initially constructed based on one-year windows) to remove physician fixed effects. Finally, for the panel with *future readmission rate* as the dependent variable, we use the first and second lags of the first-differenced readmission rate (the new explanatory variable) as well as the third lag of the readmission rate as our IVs. We use similar IVs for the panel with *future ER visit rate* as the dependent variable. This analysis again yields results that are qualitatively similar, as shown in Table A10. Specifically, neither the sentiment score nor the overall rating is significantly associated with the clinical outcome variables.

6. Conclusions

The question concerning the usefulness of online physician reviews is of obvious importance, especially in the context of chronic diseases. Chronic diseases may not have time boundaries or clear recovery cues and, hence, are fundamentally different from surgeries and treatments for curable acute conditions. Further, a number of social, behavioral, psychological, and economic determinants play significant roles in the long-term management of chronic diseases, making it harder for patients to isolate and decipher a physician's performance from confounding factors. In sum, chronic disease care is inherently more credence than episodic and time-bound treatments for curable acute conditions studied previously.

Another factor that sets chronic conditions apart is that they often require multiple visits to multiple

providers. Because it is hard to track patient visits across different providers, a limitation of prior research has been a lack of accurate measurement of readmissions. In contrast, the granularity of our data allows tracking of patient readmissions across hospitals, and our results better reflect the realities of chronic disease care and the context of credence goods in general.

Our results suggest that one needs to have a better grasp of the information conveyed by online reviews when extrapolating prior research to a credence context in healthcare. Contrary to some earlier findings, we do not observe any evidence of a clear relationship between online reviews of physicians and their patients' clinical outcomes. In other words, although online reviews and ratings may have the ability to tell a story with respect to certain aspects of care (such as efficiency, courteousness, or promptness), they are simply not reliable indicators of clinical performance of physicians, at least not for chronic diseases such as COPD. Not only are the overall review rating and sentiment score uninformative, but all clinical ratings—including those focusing on physicians—are equally lacking in signals of future clinical outcomes. Further, there is no evidence to suggest that textual reviews that emphasize certain topics, such as physicians or surgeries, are any more useful. Overall, the lack of usefulness of all forms of patient feedback is conspicuous. This is consistent with the economic theory of credence goods.

To improve the rigor of our empirical investigation, we examine a number of alternative explanations for our results. Specifically, we rule out review manipulation, patient self-selection, omitted variable bias, age-based digital divide, and review bias as probable causes; throughout these additional analyses, the lack of efficacy of online reviews was uniformly conspicuous. The broader takeaway is apparent. Although researchers have found online reviews to be quite useful for hospitality services (such as hotels and restaurants) and information goods (such as books and movies), our research indicates that they may not be useful in evaluating physician care quality for credence healthcare services, such as chronic disease care and management.

Our findings are relevant to settings that share similar features: (a) treatments with credence characteristics for which quality of care is not easily discernible; (b) diseases for which deterioration in patient health is a reliable metric of clinical outcomes; and (c) conditions that require admissions to multiple hospitals, making it harder to attribute success or failure to individual physicians.

6.1. Practical Implications

Our results have important policy implications. First and foremost, healthcare consumers should not simply

rely on online physician reviews and ratings to form opinions about the quality of care provided. Likewise, hospitals and clinics should not evaluate physician performance solely based on these reviews. Rather, one should be mindful that, in the context of chronic diseases, physicians who receive more positive online reviews cannot be assumed to provide better care quality as measured by their patients' clinical outcomes.

In a similar vein, the main lesson for policymakers is that, in certain healthcare contexts, they cannot and should not rely on public online review websites to close information gaps that still persist between patients and providers. Because reducing these information asymmetries is important to achieving an efficient marketplace and establishing a patient-centric approach to healthcare, policymakers should make additional information available to healthcare consumers. Not coincidentally, in recent years, the CMS has sought greater transparency on provider quality by initiating the Physician Quality Reporting System—a program designed to incentivize physicians to report their adherence to evidence-based clinical guidelines—and by making the reported data publicly available. Based on data submitted by physicians, the CMS assigns them an adherence score, a metric that can be used to objectively assess care quality. However, many patients are not aware of this information or how best to use it. This may explain why patients remain overwhelmingly reliant on public data sources. Whether intended or not, our findings suggest that this overreliance is clearly not without risks. To effect meaningful changes and improve public access to reliable data, governmental agencies need to create greater awareness among consumers. For example, relevant statistics on health outcomes, such as readmission rates and mortality risk, could be made widely available along with appropriate visualization tools.

6.2. Limitations and Future Research

We consistently find that online reviews are not as reliable for chronic disease care as they are in other contexts. Although this is in line with what we know about credence goods and the difficulties that consumers face in assessing them, what makes online reviews unreliable remains difficult to answer. In particular, a lack of data on how patients read and write reviews and make decisions on choosing physicians makes it difficult to pinpoint the exact mechanisms. Although we have ruled out many alternative explanations for our results and found the main insight to be quite robust, additional research is necessary to unearth the exact drivers.

Another concern is that our analysis is based on online reviews from Vitals, and it is not immediately

known whether the results derived would be applicable to other data sources. Fortunately, there is evidence that physician ratings provided by other review platforms are positively and significantly correlated with data from Vitals (Daskivich et al. 2018). This suggests that that our data source is not atypical, and our results are reasonably reliable. Nevertheless, additional validation studies could be helpful.

Finally, our clinical data set focuses on physicians who treat patients suffering from a specific chronic condition. Even though focusing on COPD allows us to rule out idiosyncratic effects associated with different health conditions, doing so also limits the generalizability of our findings to other conditions. Additional validation studies are, therefore, needed to ascertain the generalizability of our findings to a broader context.

Acknowledgments

This research was completed based on data obtained partially under institutional review board approval MR15-395 at the University of Texas at Dallas. The authors gratefully acknowledge the guidance received from the senior editor, associate editor, and the reviewer team. They are also grateful to the participants of the 2016 INFORMS Conference on Information Systems and Technology, Nashville, TN; the 2017 Conference on Health IT and Analytics, Washington, DC; the 2017 Workshop on Information Systems and Economics, Seoul; and the 2017 International Conference on Smart Health (ICSH), Hong Kong. An early version of the paper won the best paper award at ICSH 2017. Finally, they thank the seminar participants at the University of Texas at Austin, University of Iowa, University of Washington at Tacoma, Wichita State University, University of Nottingham at Ningbo, and Shanghai Jiaotong University.

Endnotes

¹Information on costs can be found at <https://www.cdc.gov/chronicdisease/about/costs/index.htm>, last accessed on 7/2/2019.

²Online review websites claim to provide quality information through their platforms. For example, the home page of Vitals.com displayed the following (on September 30, 2018): “Our vision is to create a competitive health care marketplace for consumers. We’re building better tools and offering consumers cost and quality information that provides value throughout the patient journey and connects them to the right provider.” —Heyward Donigan, President and CEO.

³Using clinical outcomes to measure the quality of care is indeed important in healthcare. Angst et al. (2011) and Kc and Terwiesch (2009) consider patient length of stay in hospitals as a proxy for quality, and Bardhan et al. (2015), Kc and Terwiesch (2009), and Senot et al. (2016) use readmission rates.

⁴The exact number changes slightly from one regression to another, depending on the model and variables considered. For example, in Table 3, it ranges from 1,946 to 2,028 because some reviews are incomplete (missing some ratings), and about 5% of the physicians in our sample contain missing data. Statistically, there is no difference between physicians who have missing data and those with complete data.

⁵ This information can be found at <https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html>.

⁶ *Sentiment score of a review* $= 2 \times \text{number of very positive words} + 1 \times \text{number of positive words} - 1 \times \text{number of negative words} - 2 \times \text{number of very negative words}$.

⁷ *Sentiment variance within a review* $= \frac{V}{\text{The length of the review}}$,

where $V = \text{number of very positive words} \times (2 - \text{sentiment score of the review})^2$

$+ \text{number of positive words} \times (1 - \text{sentiment score of the review})^2$

$+ \text{number of negative words} \times (-1 - \text{sentiment score of the review})^2$

$+ \text{number of very negative word} \times (-2 - \text{sentiment score of the review})^2$.

⁸ In addition to physician-clustered errors, we tested a model with first-differenced variables to ensure that serial correlation is not an issue in our context. The key findings obtained from this analysis are similar to those in Table 3.

⁹ There is another category of red-flagged reviews called “removed.” We treat such reviews also as “not recommended” for the purpose of this analysis.

¹⁰ For details, please refer to <https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf> as well as <https://rdrr.io/cran/sentimentr/man/sentiment.html> (last accessed on July 10, 2019).

References

- Agarwal A, Zhang W, Kuo Y, Sharma G (2016) Process and outcome measures among COPD patients with a hospitalization cared for by an advance practice provider or primary care physician. *PLoS One* 11(2):e0148522.
- Agarwal R, Gao GG, DesRoches C, Jha AK (2010) The digital transformation of healthcare: Current status and the road ahead. *Inform. Systems Res.* 21(4):796–809.
- Angst CM, Devaraj S, Queenan CC, Greenwood B (2011) Performance effects related to the sequence of integration of healthcare technologies. *Production Oper. Management* 20(3):319–333.
- Arrow KJ (1963) Uncertainty and the welfare economics of medical care. *Amer. Econom. Rev.* 53(5):941–973.
- Bardach NS, Asteria-Penaloza R, Boscardin WJ, Dudley RA (2012) The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ Quality Safety* 22(3):194–202.
- Bao C, Bardhan IR, Singh H, Meyer BA, Kirksey K (2020) Patient-provider engagement and its impact on health outcomes: A longitudinal study of patient portal use. *Management Inform. Systems Quart.* 44(2):699–723.
- Bardhan I, Chen HC, Karahanna E (2020) The role of information systems and analytics in chronic disease management. *Management Inform. Systems Quart.* 44(1):185–200.
- Bardhan I, Oh C, Zheng Z, Kirksey K (2015) Predictive analytics for readmission of patients with congestive heart failure. *Inform. Systems Res.* 26(1):19–39.
- Berger JA, Iyengar R (2012) How interest shapes word-of-mouth over different channels. <https://ssrn.com/abstract=2013141>.
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.
- Boulding W, Glickman SW, Manary M, Schulman KA, Staelin R (2011) Relationship between patient satisfaction with inpatient care and hospital readmission within 30 Days. *Amer. J. Managed Care* 17(1):41–48.
- Burkle CM, Keegan MT (2015) Popularity of internet physician rating sites and their apparent influence on patients’ choices of physicians. *BMC Health Services Res.* 15:416.
- Castaneda R (2018) Rating doctors: What you need to know. *U.S. News Online* (February 15), <https://health.usnews.com/health-care/patient-advice/articles/rating-doctors-what-you-need-to-know>.
- Chen H, Zheng Z, Ceran Y (2016) Debiasing the reporting bias in social media analytics. *Production Oper. Management* 25(5):849–865.
- Chen Y, Xie J (2008) Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Sci.* 54(3):477–491.
- Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.
- Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.
- Clarke JL, Bourn S, Skoufalos A, Beck EH, Castillo DJ (2017) An innovative approach to healthcare delivery for patients with chronic conditions. *Population Health Management* 20(1):23–30.
- Clemons EK, Gao GG, Hitt LM (2006) When online reviews meet hyperdifferentiation: A study of the craft beer industry. *J. Management Inform. Systems* 23(2):149–171.
- Darby MR, Karni E (1973) Free competition and the optimal amount of fraud. *J. Law Econom.* 16(1):67–88.
- Daskivich TJ, Houman J, Fuller G, Black JT, Kim HL, Spiegel B (2018) Online physician ratings fail to predict actual performance measures on quality, value, and peer review. *J. Amer. Medical Informatics Assoc.* 25(4):401–407.
- Dellarocas C, Narayan R (2006) What motivates consumers to review a product online? A study of the product-specific antecedents of online movie reviews. Workshop on Information Systems and Economics, Evanston, IL.
- Doyle C, Lennox L, Bell D (2013) A systematic review of evidence on the links between patient satisfaction and clinical safety and effectiveness. *BMJ Open* 3(1):1–18.
- Dranove D, Kessler D, McClellan M, Satterthwaite M (2003) Is more information better? The effects of “report cards” on healthcare providers. *J. Political Econom.* 111(3):555–588.
- Dulleck U, Kerschbamer R (2006) On doctors, mechanics, and computer specialists: The economics of credence goods. *J. Econom. Literature* 44(1):5–42.
- Dulleck U, Kerschbamer R, Sutter M (2011) The economics of credence goods: An experiment on the role of liability, verifiability, reputation, and competition. *Amer. Econom. Rev.* 101(2):526–555.
- East R, Hammond K, Wright M (2007) The relative incidence of positive and negative word of mouth: A multi-category study. *Internat. J. Res. Marketing* 24(2):175–184.
- Ellimoottil C, Hart A, Greco K, Quek ML, Farooq A (2013) Online reviews of 500 urologists. *J. Urology* 189(6):2269–2273.
- Emmert M, Meszmer N, Sander U (2016) Do healthcare providers use online patient ratings to improve the quality of care? Results from an online-based cross-sectional study. *J. Medical Internet Res.* 18(9):e254.
- Emmert M, Sander U, Pisch F (2013) Eight questions about physician-rating websites: A systematic review. *J. Medical Internet Res.* 15(2):e24.
- Ferguson WJ, Candib LM (2002) Culture, language, and the doctor-patient relationship. *Family Medicine Community Health Publications* 34(5):353–361.
- Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3):291–313.
- Friis K, Lasgaard M, Osborne RH, Maindal HT (2016) Gaps in understanding health and engagement with healthcare providers across common long-term conditions: A population survey of health literacy in 29,473 Danish citizens. *BMJ Open* 6(1):e009627.
- Gao GG, Greenwood BN, Agarwal R, McCullough JS (2015) Vocal minority and silent majority: How do online ratings reflect

- population perceptions of quality? *Management Inform. Systems Quart.* 39(3):565–589.
- Gao GG, McCullough JS, Agarwal R, Jha A (2012) A changing landscape of physician quality reporting: Analysis of patients' online ratings of their physicians over a 5-year period. *J. Medical Internet Res.* 14(1):e38.
- Glickman SW, Boulding W, Manary M, Staelin R, Roe MT, Wolosin RJ, Ohman EM, Peterson ED, Schulman KA (2010) Patient satisfaction and its relationship with clinical quality and inpatient mortality in acute myocardial infarction. *Circulation Cardiovascular Quality Outcomes* 3(2):188–195.
- Godes D, Silva JC (2012) Sequential and temporal dynamics of online opinion. *Marketing Sci.* 31(3):448–473.
- Goldman E (2015) Another failed doctor lawsuit against a patient for online reviews—Brandner v. Molonguet. Technology & Marketing Law Blog (January 8), <http://blog.ericgoldman.org/archives/2015/01/another-failed-doctor-lawsuit-against-a-patient-for-online-reviews-brandner-v-molonguet.htm>.
- Gray B, Vandergrift JL, Gao GG, McCullough JS, Lipner RS (2015) Website ratings of physicians and their true quality of care. *JAMA Internal Medicine* 175(2):291–293.
- Hanauer DA, Zheng K, Singer DC, Gebremariam A, Davis MM (2014) Public awareness, perception, and use of online physician rating sites. *JAMA.* 311(7):734–735.
- Hedges L (2019) How patients use online reviews. Software Advice. Accessed July 10, 2019, <https://www.softwareadvice.com/resources/how-patients-use-online-reviews/>.
- Jain S (2010) Googling ourselves—What physicians can learn from online rating sites. *New England J. Medicine* 362(1):6–7.
- Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.
- Lagu T, Hannon NS, Rothberg MB, Lindenauer PK (2010) Patients' evaluations of healthcare providers in the era of social networking: An analysis of physician-rating websites. *J. General Internal Medicine* 25(9):942–946.
- Liu B (2012) *Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies* (Morgan & Claypool Publishers, San Rafael, CA).
- Lu SF, Rui H (2018) Can we trust online physician ratings? Evidence from cardiac surgeons in Florida. *Management Sci.* 64(6):2557–2573.
- Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Sci.* 62(12):3412–3427.
- Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *Amer. Econom. Rev.* 104(8):2421–2455.
- Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *J. Marketing Res.* 48(3):444–456.
- Mudambi SM, Schuff D (2010) Research note: What makes a helpful online review? A study of customer reviews on Amazon.com. *Management Inform. Systems Quart.* 34(1):185–200.
- Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: A randomized experiment. *Sci.* 341(6146):647–651.
- Nelson P (1970) Information and consumer behavior. *J. Political Econom.* 78(2):311–329.
- Nielsen FÅ (2011) A New ANEW: Evaluation of a word list for sentiment analysis in microblogs. Rowe M, Stankovic M, Dadzie AM, Hardey M, eds. *Proc. ESWC2011 Workshop Making Sense Microposts: Big Things Come Small Packages* (CEUR Workshop Proc., No. 718) (CEUR, Herakleion, Greece), 93–98.
- O'Donnell J, Alltucker K (2018) Doctors, hospitals sue patients who post negative comments, reviews on social media. *USA Today Online* (July 18), <https://www.usatoday.com/story/news/politics/2018/07/18/doctors-hospitals-sue-patients-posting-negative-online-comments/763981002/>.
- Ornstein C (2016) Doctors fire back at bad Yelp reviews—And reveal patients' information online. *Washington Post Online* (May 27), https://www.oregonlive.com/today/2015/11/doctor_sues_patient_over_negat.html.
- Segal J, Sacopoulos M, Sheets V, Thurston I, Brooks K, Puccia R (2012) Online doctor reviews: Do they track surgeon volume, a proxy for quality of care? *J. Medical Internet Res.* 14(2):e50.
- Senot C, Chandrasekaran A, Ward PT, Tucker AL, Moffatt-Bruce SD (2016) The impact of combining conformance and experiential quality on hospitals' readmissions and cost performance. *Management Sci.* 62(3):829–848.
- Sloan FA (2001) Arrow's concept of the consumer of medical care. *J. Health Politics Policy Law* 26(5):899–910.
- Smith Y (2014) Credence goods: The unique economics of healthcare. *Naked Capitalism* (May 23), <https://www.nakedcapitalism.com/2014/05/credence-goods-unique-economics-health-care.html>.
- Sun M (2012) How does the variance of product ratings matter? *Management Sci.* 58(4):696–707.
- Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Marketing Res.* 51(4):463–479.
- Zhu F, Zhang X (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Marketing* 74(2):133–148.