Management Science

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

The Interplay Between Online Reviews and Physician Demand: An Empirical Investigation

Yuqian Xu, Mor Armony, Anindya Ghose

To cite this article:
Yuqian Xu, Mor Armony, Anindya Ghose (2021) The Interplay Between Online Reviews and Physician Demand: An Empirical Investigation. Management Science 67(12):7344-7361. https://doi.org/10.1287/mnsc.2020.3879

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 67, No. 12, December 2021, pp. 7344–7361
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Interplay Between Online Reviews and Physician Demand: An Empirical Investigation

Yuqian Xu,[a] **Mor Armony,**[b] **Anindya Ghose**[b]

[a] Department of Technology, Operations, and Statistics, Department of Business Administration, Gies College of Business, University of Illinois at Urbana-Champaign, Illinois 61820; [b] Stern Business School, New York University, New York, New York 10012

Contact: yuqian@illinois.edu, https://orcid.org/0000-0001-5733-0412 (YX); marmony@stern.nyu.edu, https://orcid.org/0000-0001-5970-3753 (MA); aghose@stern.nyu.edu, https://orcid.org/0000-0002-6499-8944 (AG)

Received: May 4, 2016
Revised: August 3, 2017; November 8, 2018; March 23, 2020; August 14, 2020
Accepted: October 3, 2020
Published Online in Articles in Advance: March 2, 2021

https://doi.org/10.1287/mnsc.2020.3879

Abstract. Social media platforms for healthcare services are changing how patients choose physicians. The digitization of healthcare reviews has been providing additional information to patients when choosing their physicians. On the other hand, the growing online information introduces more uncertainty among providers regarding the expected future demand and how different service features can affect patient decisions. In this paper, we derive various service-quality proxies from online reviews and show that leveraging textual information can derive useful operational measures to better understand patient choices. To do so, we study a unique data set from one of the leading appointment-booking websites in the United States. We derive from the text reviews the seven most frequently mentioned topics among patients, namely, bedside manner, diagnosis accuracy, waiting time, service time, insurance process, physician knowledge, and office environment, and then incorporate these service features into a random-coefficient choice model to quantify the economic values of these service-quality proxies. By introducing quality proxies from text reviews, we find the predictive power of patient choice increases significantly, for example, a 6%–12% improvement measured by mean squared error for both in-sample and out-of-sample tests. In addition, our estimation results indicate that contextual description may better characterize users' perceived quality than numerical ratings on the same service feature. Broadly speaking, this paper shows how to incorporate textual information into an econometric model to understand patient choice in healthcare delivery. Our interdisciplinary approach provides a framework that combines machine learning and structural modeling techniques to advance the literature in empirical operations management, information systems, and marketing.

History: Accepted by David Simchi-Levi, operations management.
Supplemental Material: The online supplement is available at https://doi.org/10.1287/mnsc.2020.3879.

Keywords: physician • patient choice • quality • social media • text mining • sentiment • rating • review • service quality • healthcare

1. Introduction

Patients in the healthcare market typically face uncertainty regarding a physician's quality and may have limited information on service-quality measures, such as accuracy of diagnosis and bedside manner, and operational factors, such as service time and waiting time. Researchers in the past years have dedicated much effort to improve healthcare quality transparency in hospitals (Harris and Buntin 2008), and more recent work has studied quality transparency of individual physicians (Wang et al. 2019). Nowadays, online physician reviews are gaining an increasing level of attention among patients to supplement their limited information. With the onset of this social media revolution, patients can now find physicians more easily, read reviews from other patients, see open appointment times, and book instantly. To gain a sense of the rapidly growing healthcare online market, we take a look at some statistics. First, the demand from the patients' side for this type of online information is growing: a recent survey found almost three quarters (72%) of patients use online reviews as their first step to find a new doctor.[1] Moreover, according to a recent report by Grand View Research Inc., the global digital health market size is expected to reach $509.2 billion by 2025.[2]

With the prevalence of online reviews, physicians have to pay closer attention to operational factors, because these factors are likely to become common knowledge. Moreover, the rapid growth of online information can also introduce more uncertainty to healthcare providers regarding their future demand and patient needs. Given the fact that patients are playing an increasingly active role in their own healthcare delivery process, the study of a patient choice model with online information is important to help better match healthcare supply with patient demand and support healthcare management. On a broader note, predictive modeling of patient choice leveraging online information brings value throughout the entire healthcare ecosystem, for example, by forestalling appointment no-shows, predicting patient utilization patterns, managing the health supply chain, bolstering patient engagement and satisfaction, and so on. This paper focuses on predictive modeling of patient choice with service-quality proxies derived from online reviews. In particular, we study a unique data set with physician reviews and daily booked appointments and propose a Berry–Levinsohn–Pakes (BLP)-type random-coefficient choice model (Berry et al. 1995) together with service-quality proxies derived from online text reviews to characterize patient choices.

1.1. Our Work

We integrate and extend the advances in the online marketing and healthcare operations literature, taking a step toward a more detailed understanding of patients' choice behavior in the healthcare online markets. To the best of our knowledge, this paper is the first to derive health-service proxies from online healthcare reviews and then incorporate them in predictive modeling of patient choice. Our work should be useful for healthcare providers and payers when predicting demand for providers' services. Based on our estimation results, we show that integrating quality proxies from text reviews can significantly improve the predictive power of patient choice, for example, a 6%–12% improvement measured by mean squared error for both in-sample and out-of-sample tests. In addition, we show that operational features, such as waiting time and service time, are important elements in predicting patient choices.

The key challenges of this paper are to integrate the various sources of social media (both textual and numeric information) to derive service-quality proxies and then to leverage the aggregate-level demand data into a patient choice model. Our work benefits from the rapid development of research on social media information. An increasing number of papers in information systems (IS) and marketing have incorporated text information into consumer choice models; however, limited research has looked at mining healthcare service reviews with operational factors and choice models in mind. On the other hand, empirical operations management (OM) literature has started to look into the value of social media information on operations (Cui et al. 2018), but limited research has utilized the information from the healthcare online market. One potential difficulty in mining physician service reviews is deriving the operations-based service features, such as waiting time and service time. Compared with product reviews, the features for physician service reviews have a more diverse format because of service intangibility. Our goal is to fill this gap by proposing a patient choice model that allows for physicians to be differentiated along multiple dimensions, contains richer distributions of choice parameters, and incorporates operational factors derived from online user-generated information.

Our data set consists of 872 physicians in the United States from one of the leading online appointment-booking platforms. We have 145 days of physician appointment data from November 27, 2014 to April 12, 2015 that contain the number of appointment slots offered over a 30-day window and the total number of appointments booked per day. In addition to the appointment data, our work combines three other types of data: (i) user-generated content (UGC) information, (ii) city-level health status data from the Kaiser Family Foundation, and (iii) patient income data from the U.S. census. We further derive the UGC data from text mining of the detailed reviews.

In the first step of our analysis, we merge various sources of physician-, city-, and patient-level characteristics that could potentially influence patient choices. We also use topic extraction and sentiment analysis to extract the text information. As a second step, we apply a BLP-type random-coefficient choice model (Berry et al. 1995, Nevo 2001, Berry and Pakes 2007, Song 2011) that can capture patient heterogeneity to estimate demand. As a final step, we compare our proposed model with a basic model (without text information) to illustrate the improvement in predictive power. In addition, we conduct subset analyses to show heterogeneous effects of service features derived from text-based reviews on rating categories and review quantity.

Our findings provide important information regarding how patients who have access to various sources of information choose their physicians. Our main findings could be summarized as follows. First, we identify the seven most frequently mentioned service features among patients through text mining, among which bedside manner, accuracy of diagnosis, waiting time, and service time have a statistically significant relationship with patient choices. Second, we show the significance of introducing text reviews in predicting doctors' short-term future demand. In particular, we compare our proposed model with a basic model (without text information) and find a 6%–12% increase in predictive power as measured by mean squared error, a 3%–6% increase as measured by root-mean-square error, and a 2%–5% increase as measured by mean absolute deviation, in all the in-sample and out-of-sample tests. Moreover, our estimation results indicate that contextual description may better characterize users' perceived quality than numerical ratings on the same service feature. Finally, our subset analyses show the predictive power of text review increases when the overall rating goes up and that quality features derived from contextual reviews are more important for popular doctors.

In summary, our paper makes two key contributions. First, to our knowledge, this paper is the first to derive healthcare service proxies from unstructured text reviews on multiple dimensions. Second, to our knowledge, our work is the first to leverage these service-quality proxies from online review data to predict physician demand in healthcare services. Moreover, our text-mining and choice-model results together show the importance of operational features, such as waiting time and service time, on healthcare service delivery. On a broader note, our interdisciplinary approach provides a framework that combines machine learning and structural modeling with empirical OM. This framework can help generate potential future empirical OM research with many different online platforms.

1.2. Background on the Online Platform

The platform we use in our paper is an online physician appointment-booking and review website launched in 2007. We show a sample of the platform setting in Appendix A of our online supplement. With this website, patients can search for physicians by specifying insurance, specialty, location, and so on. Based on the search criteria, the platform returns a list of physicians, and then patients can get access to physicians' appointment book and easily make an appointment online. Note that one important reason we choose this platform is that it has the competitive advantage of syncing the physician's entire appointment calendar. Therefore, what is shown online is the same as the physician's appointment book in the clinic; regardless of whether an appointment is booked through this platform under study or any other online or offline channel, we can see the appointment being booked as its availability is removed from this platform.

This platform does not charge the patients. By contrast, physicians who choose to participate pay $250 monthly. Each physician using this platform has a profile page with individual background information, along with patient reviews, ratings, and appointment availability. On the patient side, after each appointment, the platform sends a thank-you email to the patients and encourages them to review the physician. Verified patients can then rate the physicians they have visited. Once a review is submitted, anyone can access the platform and read the review for free. Thus, another important reason we choose this platform is that only verified patients can post a review.

To our knowledge, Luca and Vats (2014) is the only paper that studies the impact of online healthcare reviews on physician demand. A comparison of our approach and results with theirs is thus of interest. In general, their paper established a causal impact of overall ratings on physician demand. They used the review and appointment-booking data for primary care physicians in Manhattan from the same platform during February–May 2013 and ran a regression discontinuity (RD) analysis to study the causal relationship between overall ratings and physician demand. We also show the causal impact of the overall rating with a similar RD analysis (see Section 6.4 for details); however, the main focus of our paper is to derive service-quality proxies from unstructured text reviews and leverage these features to predict doctor demand. In addition, our data set contains physicians with 18 specialties across 31 U.S. cities and, hence, is more representative of the U.S. outpatient physician and patient population. Finally, we use the detailed text-review information in addition to ratings and extract important service features that have a significant correlation with patient choice. By incorporating text-mining techniques with econometric models in demand estimation, we further improve predictive power.

2. Related Literature

Two streams of literature are closely related to our paper: (i) patient choice models and (ii) the impact of online reviews on sales.

First, existing research has looked into patient choice with regard to hospitals and found hospital quality (or proxies thereof) is a key consideration. Tay (2003) uses Medicare claims data from patients over the age of 65 with a heart attack to estimate a discrete choice model and predict patient flow to hospitals. This work provides evidence of the importance of quality differentiation in healthcare markets. Razzouk et al. (2004) show that quality of care and accessibility are the two most relevant factors in patient choices of primary care physicians. Varkevisser et al. (2012) study the relationship between hospital quality measured by publicly available ratings and patients' hospital choices. Luca and Vats (2014) investigate the impact of online ratings on physician demand and show overall ratings significantly affect the demand. Wang et al. (2019) look into the quality gap among 35 hospitals in New York State and identify the major obstacles preventing patients from choosing the best-quality care. Dong et al. (2019) investigate the impact of delay announcements on patient choice and show patients take delay information into account when choosing emergency service providers. In outpatient care, Liu et al. (2018) conduct behavioral experiments to identify heterogeneous patient choices with several "operational" attributes, such as appointment delay and flexibility. Our work may be viewed as complementing theirs in that it shows some results consistent with theirs, such as the impact of waiting time on demand. The key difference is that our work is based on user-generated data rather than through a controlled experiment. Thus, we are able to derive service-quality proxies that affect patient demand through a purely data-driven approach.

A second stream of literature in marketing and IS studies how product reviews affect sales, especially in the online retail industry, such as Godes and Mayzlin (2004), Chevalier and Mayzlin (2006), Forman et al. (2008), Archak et al. (2011), and Ghose et al. (2012). In addition, existing papers have started to discuss different perspectives of the online healthcare reviews, such as the credibility of online reviews and the impact on patient satisfaction; see Harris and Buntin (2008), Gao et al. (2015), and Mein Goh et al. (2016). However, to our knowledge, none of these papers tries to leverage service-quality proxies derived from text reviews to predict doctor demand. Luca and Vats (2014) is the only paper we are aware of that considers the impact of ratings on physician demand, the details of which we mentioned in our introduction.

3. Data

Our data set consists of 17,440 observations from 872 physicians in the United States from one of the leading online appointment-booking platforms. Each observation contains the doctor's daily information on rating, review, booking, and so on. We began with a list of 7,607 active physicians;[3] but after combining the physician demographic information with their review and appointment-book data, only 2,369 physicians were left. This shrinkage in the pool of physicians is a result of the fact that over our 30-day period of data crawling, more than 25% of the time, many physicians were not offering any appointments. Therefore, we rule out this portion of physicians with incomplete information. Furthermore, we delete all the markets (city-specialty-week combination) in which the total number of physicians in our data set was no more than three. If we include all these data, the market shares of physicians from these markets would be very large, which may lead to overestimation. Following this data cleaning, we are left with a sample of 872 physicians with 18 specialty types across 31 U.S. cities. For these selected physicians, we collect 145 days of physician appointment data from November 27, 2014 to April 12, 2015 that contain the number of appointment slots offered over a 30-day window and the total number of appointments booked per day.[4] The daily demand is the total number of slots that have been booked during that day regardless of the date of the specific slot. Note the bookings measured in our paper are slots that disappear in the online appointment book,[5] which contain both online and offline bookings. Our main research question concerns the relationship between service quality and physician demand. Because service quality is unobservable, we derive service-quality proxies from online reviews. The supply intensity of the appointment book is computed as the total number of slots available in a 30-day window on a daily basis. To further reduce the dimension of our structural model, we merge the daily data into weekly data with 20 weeks in total. Finally, we obtain a total of 17,440 weekly observations over the 872 physicians.

We now discuss the details of the four types of data we collect: (i) appointment-slots supply and demand, (ii) physicians' profile information, (iii) city-level patient health status and income data, and (iv) numeric and textual reviews. We show the detailed summary of our data categories in Appendix D of the online supplement.[6] We write our web-crawler program in Python and automatically crawl each physician's daily appointment supply and demand data as well as profile data. The details of physicians' profile information could be found in Appendix B of our online supplement. We download the city-level patient health status data from the Henry J. Kaiser Family Foundation's statistics and income data from the U.S. census. We construct the review data set with topic extraction and sentiment analysis. Table 1 provides detailed summary statistics of our data set.[7] In this section, we start with the discussion on data selection. Then, we further derive the UGC data from text mining and sentiment analysis of the detailed reviews.

3.1. Physicians' Appointment and Rating Data

Out of the 872 physicians, 55.8% are male, 23% have more than one hospital affiliation, and 57.7% speak more than one language. For the appointment data, we choose a 30-day window to crawl physicians' appointments, because most of the physicians offer slots within this time window. Very few appointments are offered one month later; considering that some patients book appointments one or more weeks in advance, a 30-day window is a reasonable length. We further conduct a robustness check in Section 6.5 to verify if our results are robust across different time windows. We then collect both the date and time of the appointment slots. From Table 1, we can see the average number of offered slots in 30 days is 167, and hence the average offered slots per day is 5.57.

The physicians' rating data contain two parts: (i) the rounded average rating and (ii) the individual review ratings for three quality indicators, that is, overall, bedside manner, and waiting time, respectively. We collect both daily displayed rounded overall rating data and the three types of individual rating data. We then take the daily average of each type of rating data.

Table 1. Summary Statistics

Variable	Definition	N	Mean	SD	Min	Max
OVERALL	Overall reviewer rating	17,440	4.701	0.332	1.000	5.000
BEDSIDE	Bedside manner rating	17,440	4.744	0.275	3.000	5.000
WAITING	Waiting time rating	17,440	4.367	0.447	1.730	5.000
GENDER	Gender (male = 1, female = 0)	17,440	0.595	0.491	0.000	1.000
EXT_OFF	Number of extra offices	17,440	0.244	0.429	0.000	1.000
LAN_NUM	Number of languages spoken	17,440	2.132	1.489	1.000	12.00
AVAILABILITY	Number of slots available in 30 days	17,440	167.0	195.0	0.000	955.0
AVE_GAP	Average time gap between two slots	17,440	33.53	25.50	5.000	210.0
DEATH_RATE	Annual death rate	17,440	629.3	50.80	630.0	806.2
LIFE_EXP	Average life expectancy	17,440	79.64	1.010	77.00	80.80
PHY_ACT	Physical activities participation ratio	17,440	52.38	3.510	44.00	60.40
REVIEW_CNT	Total number of reviews	17,440	26.85	24.35	1.000	72.00
OVERALL_SQ	Square of overall reviewer rating	17,440	22.01	2.962	1.000	25.00
ID	Disclosure of reviewer identity	17,440	2.150	4.650	0.000	46.00
SMOG	SMOG index	17,440	3.200	4.100	0.000	17.00
WORD_CNT	Average sentence length	17,440	104.0	238.0	1.000	2,431
CHAR_CNT	Average number of characters	17,440	463.0	1056	0.000	10,776
SYLLABLE_CNT	Average number of syllables	17,440	143.0	325.0	0.000	3,357
BED_MANNER	Sentiment toward bedside manner	17,440	0.544	0.045	0.439	0.833
ACC_DIAG	Sentiment toward accuracy of diagnosis	17,440	0.511	0.024	0.404	0.694
WAITING_ TIME	Sentiment toward waiting time	17,440	0.512	0.025	0.376	0.953
SERVICE_TIME	Sentiment toward service time	17,440	0.513	0.032	0.329	0.886
EASE_INSUR	Sentiment toward insurance process	17,440	0.510	0.021	0.441	0.663
KNOWLEDGE	Sentiment toward physician's knowledge	17,440	0.533	0.037	0.377	0.823
OFFICE_ENVIR	Sentiment toward office environment	17,440	0.536	0.040	0.377	0.953

Note. SD, standard deviation.

From Table 1, we can see the average of overall, bedside-manner, and waiting-time ratings are 4.701, 4.744, and 4.367, respectively. Therefore, numeric ratings alone may not provide enough information on service quality because all three types of ratings are quite high (close to full), which is one of the key reasons we look into the text of reviews.

3.2. Statistics of Text Reviews

Now we focus on the quantification of the quality of reviews; in the next section, we discuss the extraction of patients' perspectives.

Consistent with prior work (Ghose and Ipeirotis 2011, Ghose et al. 2012), we control for the number of reviews and review ratings. In addition, we consider features related to "readability" and "complexity" of the text, which can also affect patient choice. Prior IS, marketing, and computer science (CS) literature has shown the quality of reviews (measured as review complexity and readability) would affect patients' online information acquisition and thus their perception of the online service provider. White (2006) shows easy-reading text improves comprehension and retention. Ghose et al. (2012) use the review-quality (readability and complexity) variables in demand estimation of product sales. Following Ghose et al. (2012), we measure the readability by the standard Simple Measure of Gobbledygook (SMOG) index, which can be computed in three steps (McLaughlin 1969): (i) count the number of sentences; (ii) in those sentences, count the polysyllables (words of three or more syllables); (iii) calculate the index

$$SMOG = 1.0430 \sqrt{\text{number of polysyllables} * \frac{30}{\text{number of sentences}}} + 3.1291.$$

In addition, based on Ghose et al. (2012), the complexity measurement of each review sentence is quantified by three variables: the total number of words, characters, and syllables. Our interpretation for why clarity of text reviews (readability and complexity) by itself matters is that although reviews can serve as proxies of service quality and service quality affects physicians' demand, when patients are reading these service-quality proxies from reviews, the textual quality and style of reviews will affect how they perceive these quality proxies and will thus affect the demand. Therefore, controlling for the readability and complexity of reviews is important because the textual review data are not standardized and thus will affect how patients interpret the service-quality information.

Moreover, following previous work (Forman et al. 2008), we compute the total number of reviewers who disclose their identities, that is, their names. To summarize, our review-quality analysis contains five types of information: (i) number of reviews, (ii) average review ratings, (iii) review readability level (SMOG index), (iv) review complexity level (the average number of characters, words, and syllables), and (v) the percentage disclosure of the reviewer's identity.

4. Service-Quality Measures from Texts

In this section, we discuss the details of the construction of our service-quality measures, which is decomposed into two parts: (i) service-quality-feature extraction and (ii) patient-opinion mining. In the first part, we follow the prior work of Hu and Liu (2004), Archak et al. (2011), and Ghose et al. (2012) to extract the frequently mentioned physician service features. In the second part, we extract the evaluation words (adjective and adverbs) corresponding to the identified service features (e.g., for the"bedside manner" feature, we extract evaluation words, such as "responsive," "professional," and "rush") and then assign "external" semantic scores to each service feature, measured by the probability of the attitude being positive. To summarize, our service-quality-measure construction could be decomposed into the following four steps:

1. Part-of-speech (POS) tagger for review, which marks each word of the review as a specific part of the speech.
2. Identify service features. For example, "the *waiting time* is too long" corresponds to the service feature of waiting time.
3. Identify opinion words with respect to each service feature. For instance, "the waiting time is too long" has the opinion "too long" that corresponds to the "waiting time" feature.
4. Determine the sentiment score of opinions, which we define as the probability of the attitude being positive.

Steps 1 and 2 belong to our first part, and steps 3 and 4 belong to our second part. We illustrate these four main steps in the following diagram (see Figure 1) referring to a specific example. We then discuss the details of our steps 1 and 2 in Section 4.1 and steps 3 and 4 in Section 4.2. Section 4.3 illustrates potential alternative sentiment scores.

4.1. Service-Quality-Feature Extraction

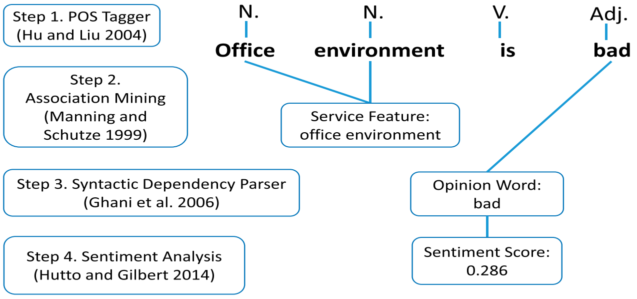
To extract service features from text, we first use a standard approach to derive explicit attributes. Given physician D and corresponding reviews $R_1, \ldots R_i \in \mathcal{R}_D$, where $\mathcal{R}_D$ is the set of all the reviews of physician D, our topic-extraction step aims to construct the (feature, opinion) sets. We first use the POS tagger to identify the part of speech of each word in a review and then derive the service features. Common candidates for product features are nouns and noun phrases; however, in our setting, verbs also construct some of the service features. For example, the verb "wait" in "I waited long" has a one-to-one mapping to the noun group "waiting time" in "waiting time was long." We then use WordNet (Fellbaum 1998) to group these phrases into a set and construct a mapping between the noun and noun phrases identified with their different verb tenses. We use an association-mining algorithm following Manning and Schutze (1999) to further cluster the candidate features. The association mining finds correlations among a set of data T_j (see definition in Algorithm 1). We summarize our algorithm in the following algorithm flow table.

Algorithm 1 (Topic Extraction)

```
Input: physician $D$, review set $\mathcal{R}_D$
for $R_i \in \mathcal{R}_D$ (i=1, ..., n) do
    for sentence $I_j \in R_i$ (j=1, ..., m) do
        POS tagger for each word in $I_j$
        Select nouns, verbs, noun phrases, and verb phrases in $I_j$ as candidates for service features defined as set $T_j$
        Association mining to group items in all the set $T_j$s that appear together frequently to find frequent features (item sets)
    end for
end for
Output: set of features $f$ of physician $D$, defined as $F_D$
```

Next, we deal with implicit views. For example, the phrases "pleasant dentist and staff," "clean office," "Dr. X took time to answer questions and discuss a plan," "made me feel comfortable," and "great experience overall" implicitly show the physician's bedside manner but do not actually mention the feature category to which they belong. To extract implicit features, we use pointwise mutual information (PMI) to link explicit features derived above with implicit features. We follow the paradigm of Ghani et al. (2006) to formulate the extraction of those implicit features as a classification problem. We execute the above algorithm for all the physicians and combine all the unique features in the feature sets F_Ds to get the combined feature set denoted as $\mathcal{F}$. Then, we count the frequency of each feature in the set $\mathcal{F}$ that shows up in all the physician reviews. We select the top M (i.e., $M = 7$ in our setting) most frequent features as our most frequently mentioned topics. After all these steps, we keep the seven most frequently mentioned topics: bedside manner, accuracy of diagnosis, waiting time, service time, ease of insurance process, physician's knowledge, and office cleanliness (the order is based on the frequency of appearance in the text, from highest to lowest); see Figure 2 for the service-quality proxies by frequency of occurrence. We do not include service-quality features with lower frequency; all of these other features have a frequency lower than 1%.

Figure 1. (Color online) Main Steps of Text Mining


Step 1. POS Tagger (Hu and Liu 2004)
Step 2. Association Mining (Manning and Schutze 1999)
Step 3. Syntactic Dependency Parser (Ghani et al. 2006)
Step 4. Sentiment Analysis (Hutto and Gilbert 2014)
N.
N.
V.
Adj.
Office
environment
is
bad
Service Feature: office environment
Opinion Word: bad
Sentiment Score: 0.286

To help interpret the seven frequently mentioned features, we illustrate them with some examples. Bedside manner means a doctor's attitudes toward a patient. For instance, one patient has posted a review stated "he explained everything in a really caring manner." Diagnosis accuracy characterizes a patient's perceived correctness of his or her diagnosis, for example, "excellent diagnosis with thorough tests and referrals as needed." Waiting time measures a patient's perceived wait in the clinic, for example, "the wait time is as long as an hour." Service time measures a patient's perceived time spent with the doctor in the clinic, for instance, "he spent extra time with his patients explaining pertinent information." Insurance process means how the clinic or doctor handles a patient's insurance plan, for example, "Dr. Lane was prompt in taking care of my insurance coverage." Physician knowledge captures a patient's feeling of whether the doctor is knowledgeable or not, for example, "very knowledgeable man and staff." An example of office environment could be "beautiful office." We further provide some sample reviews in Appendix D of our online supplement. We now have a set $\mathcal{F} = \{f_1, f_2, \ldots, f_M\}$, where f_i is each unique feature.

4.2. Patient-Opinion Mining

In this subsection, we describe the process of extraction of patient opinions toward each service feature. We start with the extraction of opinion words. We utilize a *syntactic dependency parser* (Archak et al. 2011) to choose the corresponding adjectives or adverbs of service features that we identified previously. Finally, we used the *Hidden Markov Model* (Ghani et al. 2006) in our negation-detection tool together with NegEx.[8] We show the details of opinion-words extraction in the following algorithm flow table.

Algorithm 2 (Opinion-Words Extraction)

```
Input: physician $D$, review set $R_D$
for $R_i \in R_D$ (i=1, ..., n) do
    for transaction $T_j \in R_i$ (j=1, ..., m) do
        if $T_j$ contains feature in $\mathcal{F}$ then
            Utilize a syntactic dependency parser to
                identify the adjectives corresponding
                to a service feature in each sentence $I_j$.
            Store all the negation words correspond-
                ing to the adjectives together as opinion
                words using NegEx
            Construct the opinion words set of feature
                $f$ as $\{(f, o_i, \ldots, o_j)\}$
        end if
    end for
    Take the union of sets $\{(f, o_i, \ldots, o_j)\}$ with same $f$
        for each $R_i$, so that each review only contains
        at most one (feature, opinions) set for
        each feature
end for
Output: (feature, opinions) set $\{(f, o_i, \ldots, o_j)\}$ for
    each $R_i \in R_D$
```

Now for each review of each physician, we have at most M (feature, opinions) sets $\{(f, o_i, \ldots, o_j)\}$, and our final step is to analyze patients' attitudes toward each set.

Figure 2. (Color online) Service-Quality Features by Frequency

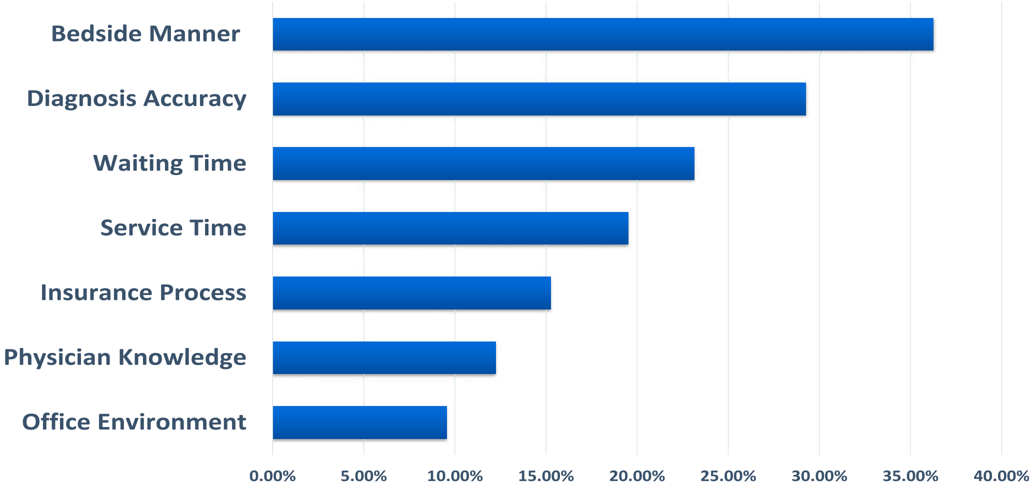


Table 2. Sentiment Strength Level

−4	−3	−2	−1	0	1	2	3	4
Very negative	. . .	. . .	Negative	Neutral	Positive	. . .	. . .	Very positive

The training data set we use is from the SentiWordNet (Esuli and Sebastiani 2007) which contains positive, neutral, and negative opinion words with different strength levels. For example, the opinion strength could go from −4 to 4, in Table 2.

We then use the *Naive Bayes analyzer* (Ghani et al. 2006) to classify our sets $\{(f, o_i, \dots, o_j)\}$ into different strength levels and use the VADER (Hutto and Gilbert 2014) sentiment analyzer from the Python Natural Language Toolkit (NLTK) to normalize the sentiment score into a positive probability between zero and one. The higher this probability is, the stronger the positive sentiment. We compute this score for each service feature of each review a physician has and then combine the scores of all the reviews into a daily average. We consider the following example as an illustration on how to compute the daily average scores: suppose a physician has 20 text reviews, and only one mentions bedside manner and is very negative. In this case, we consider the 19 other text reviews as neutral (with a sentiment score of 0.5) and take the average of all 20 reviews. The results are summarized in the last seven rows of Table 1. However, assigning a neutral sentiment score to reviews that do not mention a specific aspect may not always be appropriate. For example, suppose a customer commented that "I'm happy with everything, and especially appreciate how knowledgeable the physician was." In this case, assigning a neutral score to bedside manner, diagnosis, waiting time, service time, insurance process, and office environment may not be the best option. Therefore, to justify the estimation results based on our proposed approach of sentiment analysis, we come up with two alternative ways to compute the sentiment score of the text variables in the next subsection.

4.3. Alternative Sentiment Scores

In this subsection, we discuss the details of two alternative ways to compute the sentiment score of the text variables. First, instead of computing the probability of the attitude being positive, we define the sentiment score as the percentage of positive or negative reviews. We use two variables—"POS_PERCENT_X" and "NEG_PERCENT_X"—to replace the original sentiment score of service feature "X" (e.g., X can stand for bedside manner). These two variables are thus computed as

$$POS\_PER\_X = \frac{Number\ of\ Positive\ Reviews\ of\ Feature\ X}{Total\ Number\ of\ Reviews},$$

$$NEG\_PER\_X = \frac{Number\ of\ Negative\ Reviews\ of\ Feature\ X}{Total\ Number\ of\ Reviews}.$$

We can see the variables *POS_PER_X* have meanings similar to our original sentiment scores, which capture the likelihood of the reviews being positive. In addition, variables *NEG_PER_X* further measures the likelihood of the reviews being negative.

The second alternative sentiment scores intend to combine both approaches mentioned before and are computed as follows:

$$POS\_X = SEMATIC\_POS\_X * POS\_PERCENT\_X,$$
$$NEG\_X = SEMATIC\_NEG\_X * NEG\_PERCENT\_X,$$

where "SEMATIC_POS_X" ("SEMATIC_NEG_X") is computed as the average sentiment score (using the method discussed in Section 4.2) of X from all the positive (negative) reviews of X. Note that if no positive (negative) reviews or no reviews of X are recorded, then "POS_X" ("NEG_X") is zero. We can see again that the variables *POS_X* have meanings similar to our original sentiment scores that capture the likelihood of the reviews being positive, and variables *SEMATIC_NEG_X* further measure the likelihood of the reviews being negative.

5. Choice Model

In this section, we discuss our proposed BLP-type structural model. The reason we apply this model is the following. First, with the data set we collected, we can only observe the aggregated-level demand and not the individual choices. However, we believe heterogeneity exists among patients for physician choices (Liu et al. 2018). The key feature of the BLP-type structural model is that it introduces patient heterogeneity with random coefficients when econometricians cannot observe individual-specific choices. Second, several endogeneity issues exist in the model (see details in Section 5.3); for example, the literature has widely recognized the endogeneity between ratings and patient demand (Luca 2011, Luca and Vats 2014). Therefore, to reduce this endogeneity, we introduce instrumental variables (IVs). To estimate the random-coefficient choice model with IVs, we follow the estimation method used in the classic BLP model; see Berry et al. (1995). Finally, in the classic BLP model, the heterogeneity among patients is captured only through the random coefficients (i.e., the random preference shock). Here we want to go beyond the pure random shock and characterize more heterogeneity through accessible data. Therefore, in our patient-preference function, we follow Nevo (2001) and use the publicly available U.S. census data on patient income to capture part of the patients' heterogeneous choices caused by income differences among cities in the United States. Indeed, in our estimation results of Table 3, we find statistically significant preference heterogeneity among patients. Note we do not use a fixed-effects model as Nevo (2001) does to control for some unobserved time-invariant physician characteristics for the following reason. The inclusion of physician fixed effects would absorb a substantial amount of variation used in the analysis, and many of the quality proxies used in our model (e.g., rating and the seven service features) are unlikely to change substantially over the short span of our data set.

Now we discuss some details of our setup. Our data set could be constructed as a three-dimensional panel of demand and service characteristics for 18 physician specialties across 31 U.S. cities over 20 weeks. The estimation has two challenges: (i) the lack of patient individual-level choice and demographic information, which left us with the aggregated demand data, and (ii) the endogeneity problem caused by the correlation between service characteristics and specialty-city-week-specific demand shocks.

We deal with the first challenge by using the aggregated demand information from the physician side, which is captured as the market share, and then use the national census data to construct the empirical distribution of patient demographics to characterize the heterogeneity in patient preferences. Following Nevo (2001), we incorporate the income information into patients' heterogeneity preference function and try to partially relax the parametric assumptions used in the classic Berry et al. (1995) model. We can then model the interaction effects with income and the random preference shocks.

The endogeneity issues (discussed in Section 5.3) make our estimation difficult. We follow closely the discrete-choice literature, namely, McFadden (1978), Berry (1994), Berry et al. (1995), and Nevo (2001), to introduce a set of IVs in our estimation procedure to reduce potential endogeneity and construct the general method of moments (GMM) objective function.

5.1. Patient Choice

In this subsection, we discuss the patient choice structure. We follow the standard random coefficient choice model framework with market indexed by $t = 1, \dots, T$ ($T = 4580$), and each market has $I = 20$ patients[9] indexed by $i = 1, \dots, I$ and J_t physicians indexed by $j = 1, \dots, J_t$ that vary by market t. Furthermore, we assume each physician is captured by K ($K = 16$ in the basic model and $K = 23$ in the full model with sentiment analysis) characteristics. In typical choice models, a market is defined as a "city-time" combination; see Berry et al. (1995), Nevo (2001), and Ghose et al. (2012). However, we incorporate the special property of the healthcare market and consider a "market" as a "specialty-city-week" combination, because doctors with different specialties are not likely to substitute for each other. In our robustness check, we further group doctors with similar specialties and run the estimation again to confirm the validity of our results. The market share of one physician is then defined as the total number of slots booked for this physician divided by the size of the market that this physician belongs to, which is proportional to the number of available appointment slots in that market. We compute the number of slots offered in a 30-day window in each market via our data set. In our robustness check, we merge the available physicians from another leading online appointment-booking website and consider as our potential market the total number of slots offered by physicians from both websites.

The patient i's utility function u_{ijt} of choosing physician j in a market t consists of observed physician characteristics $\boldsymbol{x_{jt}}$ and unobserved (by the econometrician) physician characteristics ξ_{jt}, the overall rating, r_{jt}, and the market-specific characteristics, $\boldsymbol{z_t}$. In our model, the market-specific characteristics are important, because the demand in a healthcare market is likely to be driven by the local health characteristics, such as annual death rate, average life expectancy, and the number of regular physician visits. Therefore, we specifically include the market-specific characteristics, $\boldsymbol{z_t}$, in our patient utility function. Now we can write our utility function as follows:

$$u_{ijt} = \alpha_i r_{jt} + \boldsymbol{\beta_i x_{1jt}} + \boldsymbol{\gamma x_{2jt}} + \boldsymbol{\eta z_t} + \xi_{jt} + \epsilon_{ijt},$$
$$i = 1, \dots, I, \quad j = 1, \dots, J_t, \quad t = 1, \dots, T, \qquad (1)$$

with $\boldsymbol{x_{jt}} = (\boldsymbol{x_{1jt}}, \boldsymbol{x_{2jt}})$, and $\boldsymbol{x_{1jt}}$ denotes a Z-dimensional vector,[10] and Z is the number of physician characteristics that are subject to patient preference shocks. The econometric error ϵ_{ijt} is a mean-zero stochastic term. Denote by $\boldsymbol{\theta} = (\alpha_i, \boldsymbol{\beta_i}, \boldsymbol{\gamma}, \boldsymbol{\eta})$ the set of parameters to be estimated. In this setting, α_i and $\boldsymbol{\beta_i}$ are random coefficients that characterize patients' heterogeneous preferences toward the average rating and some of the physician's characteristics $\boldsymbol{x_{1jt}}$. We select the characteristics $\boldsymbol{x_{1jt}}$ by their practical interpretation as well as their statistical significance.

Next, we capture the patient-preference function. Note our approach mainly follows from Nevo (2001); however, we characterize the patient preference shocks over multiple variables, as compared with Nevo (2001), who only characterizes the preference shock over price. We denote patient i's demographic information as D_i and the pure random component as v_i. We expect a patient's income to affect his or her preferences (Nevo 2001, Ghose et al. 2012), so in our model, we also use the income distribution to characterize the patient demographics. We use data from the U.S. census to estimate the empirical distribution of income $\mathbb{P}_D(\cdot)$. Formally, the patient preference shocks in our paper are modeled as

$$\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} = \begin{pmatrix}\bar{\alpha}\\ \bar{\beta}\end{pmatrix} + \Pi D_i + \Sigma v_i, \quad v_i \sim \mathbb{P}_v(\cdot), \quad D_i \sim \mathbb{P}_D(\cdot), \tag{2}$$

where $\mathbb{P}_v(\cdot)$ is the standard multivariate normal distribution (Berry et al. 1995). Parameter Π is a $(Z+1)\times 1$ matrix, and Σ is a $(Z+1)\times(Z+1)$ matrix.

Recall that $\boldsymbol{\theta} = (\alpha_i, \boldsymbol{\beta_i}, \boldsymbol{\gamma}, \boldsymbol{\eta})$ is the set of parameters to be estimated. Now let $\boldsymbol{\theta_1} = (\bar{\alpha}, \bar{\beta})$ represent the linear parameters, and let $\boldsymbol{\theta_2} = (vec(\Pi), vec(\Sigma))$ be the nonlinear parameters. Therefore, $\boldsymbol{\theta}$ could now be written as $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$. Combining (1) and (2), we have

$$u_{ijt} = \delta_{jt}(r_{jt}, \boldsymbol{x_{jt}}, \boldsymbol{z_{jt}}, \xi_{jt}; \boldsymbol{\theta_1}) + \mu_{ijt}(r_{jt}, \boldsymbol{x_{1jt}}, v_i, D_i; \boldsymbol{\theta_2}) + \epsilon_{ijt}, \tag{3}$$

where $\delta_{jt} = \bar{\alpha} r_{jt} + \bar{\boldsymbol{\beta}}\boldsymbol{x_{jt}} + \xi_{jt} + \boldsymbol{\gamma}\boldsymbol{x_{2jt}} + \boldsymbol{\eta}\boldsymbol{z_t}$ is the mean utility and $\mu_{ijt} = [r_{jt}, \boldsymbol{x_{1jt}}]' \times (\Pi D_i + \Sigma v_i)$ is the deviation from the mean.

According to our model, a patient will choose the physician that maximizes his or her utility. Thus, we can construct the choice set of physician j in market t as

$$A_{jt} = \{(D_i, v_i, \epsilon_{ijt}) | u_{ijt} \geq u_{ilt}, \ \forall\ l = 0, 1, \ldots, J_t\}. \tag{4}$$

Note $l = 0$ refers to the choice of outside good. Assuming ties among utilities occur with probability zero, the market share of physician j is

$$s_{jt}(r, \boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \int_{A_{jt}} d\mathbb{P}(D, v, \epsilon) = \int_{A_{jt}} d\mathbb{P}(D)d\mathbb{P}(v)d\mathbb{P}(\epsilon), \tag{5}$$

which follows from the assumption of independence of D, v, and ϵ.

5.2. Estimation

Following the classical method of Berry et al. (1995) and Nevo (2001), we use the general method of moments estimation. Given a starting value for the deviation function $\boldsymbol{\theta_2^0}$, we search for δ that makes our estimated market share and the true market share equal. Note that once we add the IVs into a random-coefficients choice model, we cannot solve for δ analytically, so we have to solve for it numerically using the standard contraction-mapping method (Berry et al. 1995). For the convenience of readers, we highlight the standard contraction-mapping procedure. For any $(s, \boldsymbol{\theta}, r)$, the operator $T(s, \boldsymbol{\theta}, r) : R^{J_t} \to R^{J_t}$ for contraction mapping is defined pointwise by

$$T(s, \boldsymbol{\theta}, r)[\delta_j] = \delta_j + \ln(s_j) - \ln[s_j(r, \boldsymbol{x}, \delta; \boldsymbol{\theta})], \tag{6}$$

with its modulus less than one. The Equation (6) above implies we can solve for δ recursively. That is, we begin by evaluating the right-hand side of (6) at some initial value of δ, obtain a new δ' as the output of this calculation, substitute δ' back into the right-hand side of (6), and repeat this process until convergence. Note that compared with Nevo (2001), which has the same number of products in each market, we have a different number of physicians, so we need to introduce extra accounting vectors in each market to vary the choice space.

We construct our objective function by interacting the unobservable ξ with our IVs. The value of $\boldsymbol{\theta_2}$ will be updated with $\boldsymbol{\theta_2^1}$ through the Nelder-Mead simplex algorithm (Nelder and Mead 1965). We then repeat our procedure until the optimal values of $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are found.

5.3. Discussion of Empirical Identification and Endogeneity Issues

Although the focus of this paper is on introducing service-quality measures to predict doctor demand, discussing potential endogeneity issues in the estimation is still relevant to support and stimulate future research. In general, three potential endogeneity issues may be involved in the estimation. The first issue is the endogeneity between overall rating and demand, which Luca and Vats (2014) study. The second issue is the measurement error introduced by the computation of sentiment scores. The third issue is the potential other unobserved quality measures. In our work, we introduce IVs to address the first two issues; at the same time, we cannot fully address the third issue in this paper. We believe this issue could potentially be an interesting future research direction. In what follows, we present the details of these potential endogeneity issues.

5.3.1. Overall-Rating Endogeneity. Omitted variables are the primary cause of the endogeneity issue of the overall rating (e.g., the underlying doctor quality). To check the potential endogeneity, we run our estimation without any IV and show the results in Table 3 of the online supplement, where we can see that although the overall rating is still statistically significant, the rating square is no longer statistically significant, because of the biased estimation. Furthermore, the values of the GMM objective in Table 3 of the online supplement are higher than our main estimation results with IVs. The increasing GMM objective values indicate the explanatory power of current variables (without IVs) is lower than the original set of variables. Therefore, this check further validates the usefulness of introducing IVs in our study.

Thus, to reduce the rating endogeneity, we introduce four types of IVs. Following Hausman (1996), we compute the average ratings of the same specialty physicians in all other markets as our first IV. The underlying assumption follows from Hausman (1996); that is, controlling for specialty-specific means and demographics, market-specific valuations are independent across markets (but are allowed to be correlated within the same market). Hence, ratings for the same specialty physicians will be correlated among different markets because of the common specialty-related characteristics but are uncorrelated with other market-specific characteristics. Next, following Villas-Boas and Winer (1999), Archak et al. (2011), and Ghose et al. (2012), we use lagged ratings as instruments. To control for the systemic trends that might cause the demand to be correlated over time, we use search-volume data (Ghose et al. 2012) for different specialty-care services in each market, thus alleviating the concern to a certain extent. The third IV we use is the difference between the rounded rating (displayed online) and the average rating, calculated by taking the average of the overall ratings posted by all the reviewers. This IV captures the exogenous shock caused by the rounding effects of calculation, which is correlated with the ratings but uncorrelated with the demand. The last IV we use is the Medicare fee-for-service reimbursement-per-enrollee data for each borough in 2013 from Centers for Medicare and Medicaid Services.[11] We consider this variable as one of our instruments because (i) more than 95% of physicians on this booking website accept Medicare plans and (ii) starting from October 1, 2012, insurance companies have taken into account patient satisfaction as one of the factors that determine the Medicare reimbursement amount to the physicians.[12] We therefore consider this variable to be correlated with the service quality and thus the overall rating but uncorrelated with physician demand.

We further check the relevance and the exclusion-restriction conditions here. We start with the relevance condition. We conduct the first-stage regression with respect to the rating and the square of rating and show in Table 4 of the online supplement that all four IVs are statistically significant. We then check the F-statistics for the joint significance of our IVs for the first-stage regression; the value is over 10, which indicates our IV combination is not weak and satisfies the relevance condition (Staiger and Stock 1997). In addition, to account for the exogenous instruments, we conduct the Sargan test of overidentifying restrictions (Sargan and Desai 1988). The p-value here is 0.67, so we fail to reject the null hypothesis that our IVs are uncorrelated with the error terms. For the exclusion-restriction condition here, in addition to what we have already discussed in this section, we can see the difference in ratings and the insurance reimbursement ratio should affect demand only through the ratings, without directly affecting the patient's choice. Moreover, the Hausman-type IVs are popularly used in most of the BLP model applications. Nonetheless, so far, we do not have established statistical tools to test the exclusion-restriction assumption.

5.3.2. Service-Quality Measurement Error. The measurement errors underlying our text-mining variables become part of the error term in our econometric model, thus creating an endogeneity bias. As suggested by Wansbeek and Meijer (2000), two different ways to measure the variable under study will likely have different errors (each one is missing something else) and thus can help correct the measurement bias for each other. Accordingly, we compute sentiment scores for the seven service-quality measures based on an alternative algorithm. In particular, Brody and Elhadad (2010a) and Brody and Elhadad (2010b) propose a unique aspect-based sentiment analysis method, which could be applied to analyze healthcare-related text information. We then compute the alternative sentiment scores for the seven service-quality measures based on Brody and Elhadad (2010a) and Brody and Elhadad (2010b) and use them as IVs to correct the measurement error in the model. Following Wansbeek and Meijer (2000), these IVs from a second measurement approach are subject to another independent measurement error; they are correlated with service-quality measures because of the same underlying latent quality variables but are independent from each other's measurement errors.

5.3.3. Other Unobserved Quality Measures. The classic omitted variable problem causes a potential endogeneity issue; that is, the text-mining variables cannot capture all the quality features. Therefore, identifying the precise causal inference of our proposed seven quality measures is difficult. Because the goal of this paper is to show that deriving service-quality measures through text reviews can improve the predictive power of doctor demand, we leave the precise causal analysis of service-quality measures as a potential future research direction.

6. Empirical Analysis and Results

In this section, we discuss our main results. In Section 6.1, we present the estimation results from both the basic (without text variables) and full model (with text variables). In Section 6.2, we compare the basic model with the full model to illustrate the advantage of leveraging text information in predictive modeling. In Section 6.3, we run several subset analyses with our full model to study the heterogeneous effects of service features derived from texts. In Section 6.4, we present robustness checks.

6.1. Estimation Results

The estimation results of both the basic (without text variables) and full model (with text variables) based on two data sets are presented in Table 3. We define model I as the model estimation results with our main data set and model II as the results with both the main data set and a combined data set from another online platform. We find all the results to be consistent with each other. First, we show that the overall rating is positively correlated with the physician's demand; this positive correlation decreases as the rating increases. Moreover, we find the interaction between overall rating and income (from the interaction effects part) has a positive sign, which indicates patients from higher-income cities care more about the overall rating. However, we find the magnitude of the positive impact of the overall rating decreases compared with the basic model, because of the impact of service-quality proxies from the text information.

The four quality proxies derived from text that are statistically significant are bedside manner, accuracy of diagnosis, waiting time, and service time. Among these four proxies, bedside manner has the highest positive effect, followed by accuracy of diagnosis, waiting time, and service time. This result is consistent with the ongoing discussion in the healthcare industry on the importance of bedside manner;

Table 3. Results from Basic and Full Model

	Coefficients (Std. error)							
Variable	Basic I		Basic II		Full I		Full II	
OVERALL	0.3386^{***}	(0.0016)	0.3476^{***}	(0.0023)	0.2602^{***}	(0.0051)	0.2481^{***}	(0.0050)
GENDER	−0.0114	(0.0126)	−0.0109	(0.0129)	−0.0622	(0.0826)	−0.0602	(0.0588)
EXT_OFF	−0.1267	(0.1134)	−0.1273	(0.1339)	−0.0984	(0.1021)	−0.0922	(0.1002)
LAN_NUM	−0.0094	(0.0101)	−0.0099	(0.0113)	−0.0562	(0.0553)	−0.0523	(0.0590)
AVE_GAP	0.0579	(0.0663)	0.0557	(0.0561)	0.0455	(0.0912)	0.0401	(0.0902)
DEATH_RATE	-0.0640^{**}	(0.0213)	-0.0711^{**}	(0.0237)	−0.1124	(0.1255)	−0.1582	(0.1663)
LIFE_EXP	−0.2364	(0.2124)	−0.2179	(0.2266)	0.0971	(0.0998)	0.0931	(0.1001)
PHY_ACT	−0.1059	(0.1148)	−0.1063	(0.1127)	0.1533	(0.1356)	0.1489	(0.1050)
REVIEW_CNT	0.2543^{***}	(0.0115)	0.2579^{***}	(0.0108)	0.0778^{***}	(0.0070)	0.0695^{***}	(0.0073)
RATING_SQ	-0.0237^{***}	(0.0015)	-0.0236^{***}	(0.0017)	-0.0238^{***}	(0.0023)	-0.0229^{***}	(0.0010)
ID	0.2746^{***}	(0.0023)	0.2719^{***}	(0.0033)	0.2371^{***}	(0.0053)	0.2416^{***}	(0.0037)
SMOG	0.2531^{***}	(0.0055)	0.2579^{***}	(0.0088)	0.0295^{*}	(0.0134)	0.0276^{*}	(0.0125)
WORD_CNT	-0.1497^{***}	(0.0074)	-0.1501^{***}	(0.0077)	-0.0255^{*}	(0.0116)	-0.0235^{**}	(0.0077)
CHAR_CNT	-0.0248^{***}	(0.0012)	-0.0411^{***}	(0.0020)	−0.0080	(0.0092)	−0.0082	(0.0097)
SYLLABLE_CNT	-0.0155^{***}	(0.0011)	-0.0148^{***}	(0.0017)	−0.0101	(0.0099)	−0.0082	(0.0073)
BED_MANNER	–	–	–	–	0.3781^{***}	(0.0086)	0.3850^{***}	(0.0081)
ACC_DIAG	–	–	–	–	0.3032^{***}	(0.0089)	0.3114^{***}	(0.0056)
WAITING_TIME	–	–	–	–	0.2767^{**}	(0.0923)	0.2899^{**}	(0.0965)
SERVICE_TIME	–	–	–	–	0.1003^{**}	(0.0335)	0.1001^{**}	(0.0333)
EASE_INSUR	–	–	–	–	−0.0651	(0.5326)	−0.0602	(0.5216)
KNOWLEDGE	–	–	–	–	0.0679	(0.1655)	0.0690	(0.1450)
OFFICE_ENVIR	–	–	–	–	−0.0887	(0.1001)	−0.0916	(0.1023)
				Interaction effects				
OVERALL	0.3089^{***}	(0.0187)	0.3076^{***}	(0.0171)	0.2766^{***}	(0.0068)	0.2336^{***}	(0.0089)
BED_MANNER	–	–	–	–	0.1240^{**}	(0.0408)	0.1120^{*}	(0.0510)
ACC_DIAG	–	–	–	–	0.0969^{*}	(0.0442)	0.0979^{*}	(0.0446)
WAITING_TIME	–	–	–	–	0.0935^{*}	(0.0423)	0.0914^{*}	(0.0416)
				Standard deviations				
OVERALL	0.0748	(0.0756)	0.0801	(0.0799)	0.1006^{***}	(0.0036)	0.0998^{***}	(0.0056)
GENDER	0.0328^{**}	(0.0109)	0.0322^{**}	(0.0107)	0.0249	(0.0202)	0.0307	(0.0299)
BED_MANNER	–	–	–	–	0.1002^{*}	(0.0456)	0.1004^{*}	(0.0459)
ACC_DIAG	–	–	–	–	0.0922^{*}	(0.0419)	0.0976^{*}	(0.0422)
WAITING_TIME	–	–	–	−0000	0.0845^{**}	(0.0282)	0.0813^{*}	(0.0370)
SERVICE_TIME	–	–	–	–	0.0512^{*}	(0.0233)	0.0500^{*}	(0.0228)
GMM objective	6.324e-5		6.846e-5		1.351e-6		1.642e-6	
N	17,420		17,420		17,420		17,420	

Note. I, our focal platform data only; II, our focal platform together with another platform data; Std., standard.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

see Silverman (2012) and Renter (2015). Moreover, the importance of accuracy of diagnosis is to be expected in the healthcare industry. Finally, previous work has shown the waiting time is becoming more and more important in increasing patient satisfaction; see, for example, Pulia (2011). Beyond the direct relationship between these service-quality proxies and demand, we also find bedside manner, accuracy of diagnosis, and waiting time are statistically significantly correlated with income through the interaction term. Table 3 shows that the higher the income, the higher the correlation between physician demand and bedside manner, accuracy of diagnosis, and waiting time. Finally, we show our estimation results with the two alternative sentiment scores in Tables 5 and 6 of the online supplement and find the results are consistent with our main results.

We now assess the economic value of the important factors. A 10% increase in the total number of reviews or a 10% increase in the number of reviews with identity disclosure increase demand by 0.71% and 3.82%, respectively. A 10% increase in the SMOG index increases demand by 0.33%. A 10% increase in word count decreases demand by 0.30%. Finally, a 10% increase in the sentiment of bedside manner, diagnosis, waiting time, and service time changes demand by 7.19%, 6.47%, 5.48%, and 2.95%, respectively.

6.2. Predictive Modeling of Patient Choice

We have shown that text reviews have a significant effect on patient choices, and hence we can possibly learn patient preference for particular service features by relating changes in the doctor demand to the changes in the context of patient reviews. In this subsection, we aim to show the importance of introducing text reviews in predicting doctors' short-term future demand. In particular, we discuss the following three tasks: (i) the improvement of predictive power by introducing text variables; (ii) comparison between text variables and direct review scores on the same subcomponents; and (iii) predictive modeling with both text variables and direct review scores on subcomponents of bedside manner and waiting time.

To measure the predictive power, we conduct both in-sample and out-of-sample validations to compare the root-mean-square error (RMSE), mean squared error (MSE), and mean absolute deviation (MAD) for both models. In particular, the MSE measure is computed as follows:

$$MSE = \sum_{t=1}^{T}\sum_{j=1}^{J_t}\left(s_{jt} - \hat{s}_{jt}\right)^2,$$

where J_t is the total number of physicians in market t and T is the total number of markets as defined in Section 5.1. The RMSE measure is then computed as

$$RMSE = \sqrt{MSE},$$

and the MAD measure is computed as

$$MAD = \sum_{t=1}^{T}\sum_{j=1}^{J_t}\left|s_{jt} - \hat{s}_{jt}\right|.$$

We first compare our full model including text-mining variables with the basic model to show the significance of including text information. The in-sample comparison shows the full model including text mining has much lower RMSE, MSE, and MAD values (0.0894, 0.0080, and 0.0973, respectively) than the basic model (0.0959, 0.0091, and 0.1031, respectively); see Table 4.

Because the in-sample validation procedure is known to draw an overly optimistic conclusion regarding the model's forecasting capability, we conduct an out-of-sample validation as well. Following Steckel and Vanhonacker (1993), we first use repeated random subsampling to split our data set into 70% training and 30% validation subsets. For each round, we fit our model with the training data and then compute the RMSE, MSE and MAD using the validation data. We repeat this process 100 times and report the average results for all RMSE, MSE, and MAD in Table 4. The random subsampling comparison shows the full model including text mining has lower RMSE, MSE, and MAD values (0.1020, 0.0104, and 0.1089, respectively) than the basic model (0.1063, 0.0113, and 0.1132, respectively).

We further validate our results with a 10-fold cross validation (McLachlan et al. 2005) as our out-of-sample check. Our model again has higher predictive power with lower RMSE, MSE, and MAD values (0.1114, 0.0124, and 0.1151, respectively) than the basic model (0.1149, 0.0132, and 0.1176, respectively); see Table 4. To summarize, in all the in-sample and out-of-sample tests, we saw a 6%–12% increase in predictive power as measured by MSE, a 3%–6% increase as measured by RMSE, and a 2%–5% increase as measured by MAD.

Next, we compare the text variables with direct review scores on the same subcomponents of bedside manner and waiting time. To do so, we replace the bedside-manner and waiting-time measures from text analysis with the direct review scores on these subcomponents (i.e., bedside manner and waiting time) to run our estimation again. From the estimation results in Table 7 of the online supplement, we find the direct review scores on bedside manner and waiting time are not statistically significant, which might be caused by a potential multicollinearity issue. In fact, the correlation coefficient between the

Table 4. Model Comparison

Comparison	Measure	Basic model	Full model	Factor analysis
			(% improvement)	(% improvement)
In-sample	RMSE	0.0959	0.0894 (6.8%)	0.0883 (1.2%)
In-sample	MSE	0.0091	0.0080 (12.1%)	0.0078 (2.5%)
In-sample	MAD	0.1031	0.0973 (5.6%)	0.0962 (1.1%)
Out-of-sample (random subsampling)	RMSE	0.1063	0.1020 (4.0%)	0.1010 (1.0%)
Out-of-sample (random subsampling)	MSE	0.0113	0.0104 (8.3%)	0.0102 (1.6%)
Out-of-sample (random subsampling)	MAD	0.1132	0.1089 (3.8%)	0.1078 (1.0%)
Out-of-sample (10-fold cross validation)	RMSE	0.1149	0.1114 (3.0%)	0.1109 (0.4%)
Out-of-sample (10-fold cross validation)	MSE	0.0132	0.0124 (6.2%)	0.0123 (1.1%)
Out-of-sample (10-fold cross validation)	MAD	0.1176	0.1151 (2.1%)	0.1149 (0.2%)

Notes. Note that (% improvement) under full model compares full model with basic model. Note that (% improvement) under factor analysis compares factor analysis model with full model.

bedside-manner rating and overall rating is 0.82, and the correlation coefficient between the overall rating and waiting-time rating is 0.51. Both correlation coefficients are high, which can result in a multicollinearity issue. Multicollinearity can increase the standard errors of some coefficients. Increased standard errors can lead to cases in which some independent variables become statistically insignificant as shown in Table 7 of the online supplement. To be more specific, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant. By reducing multicollinearity (i.e., with lower standard errors), those coefficients might be significant; hence, we delete the overall rating and run our estimation again as shown in Table 8 of the online supplement. From Table 8 of the online supplement, we can see that both bedside-manner and waiting-time ratings are statistically significant; however, we then fail to capture the underlying information from the overall rating, thus this specification may not be the best analog of our main model. We therefore conclude the text analysis on particular components can potentially deliver better information than the numeric customer-given ratings on these same components. Previous studies from Ghose and Ipeirotis (2011), Archak et al. (2011), and Ghose et al. (2012) show the detailed emotional content extracted from the text reviews makes text analysis fertile ground for demand and sales forecasting and numeric ratings alone cannot achieve predictive power as good as that of text reviews. Moreover, Wu et al. (2015) mention that consumers can perceive the rating and text content of the same attribute as having different accuracy in reflecting the same experience, because of the difference between a numerical scale and contextual description. Our results thus provide further evidence on the importance of analyzing the unstructured text reviews and document that contextual description may better reflect users' perceived quality.

Finally, we consider predictive modeling with both text variables and direct review scores on subcomponents. To do so, we conduct a factor analysis that combines five variables (overall rating, bedside-manner rating, waiting-time rating, bedside-manner sentiment, and waiting-time sentiment) into three factors; see Table 5 for the loading results below. In particular, factor 1 has high loadings on three rating variables, factor 2 has high loading on the bedside-manner sentiment variable, and factor 3 has high loading on the waiting-time sentiment. We then run the BLP model again with the three new variables from factor analysis and show our estimation results in Table 7 of the online supplement, which show all three new variables are statistically significant and positively correlated with doctor demand. In addition, we compute the predictive power of this model in terms of RMSE, MSE, and MAD; see Table 4. In all the in-sample and out-of-sample tests, we see improvement in predictive power (i.e., a 1.1%–2.5% increase in predictive power as measured by MSE, a 0.4%–1.2% increase as measured by RMSE, and a 0.2%–1.1% increase as measured by MAD); however, the change is relatively marginal. Moreover, we observe similar GMM objective values in Table 7 of the online supplement as our main estimation. Therefore, we conclude that text variables alone can

Table 5. Loadings of Factor Analysis

Loadings	Factor 1	Factor 2	Factor 3
OVERALL	0.707	0.179	0.122
BED_RATING	0.611	0.293	0.106
WAIT_RATING	0.513	0.084	0.198
BED_MANNER	0.140	0.659	0.087
WAIT_TIME	0.083	0.165	0.691

achieve almost the same predictive performance as using both text variables and numerical ratings.

6.3. Heterogeneous Effects of Service Features

In this subsection, we conduct subset analyses to further understand the heterogeneous effects of service features derived from text reviews. In particular, we are interested in examining the interplay between service features derived from text content and the overall rating as well as the number of reviews.

We first conduct subset analysis of physicians based on different rating categories, namely, 3.5, 4, 4.5, and 5 stars; see Table 8 of the online supplement. From the subset estimation results, we find a "rating saturation" effect, which means quality proxies derived from the text information is more important for physicians close to the full rating (i.e., 4.5 or 5 star). For physicians with a 3.5 star rating, quality proxies from text information play very limited roles. Intuitively, for lower-rated physicians, our results suggest patients may not care too much about their detailed quality proxies. However, for higher-rated physicians, the rating provides limited additional information to the patients. Therefore, when choosing highly rated physicians, patients tend to study the details of quality proxies, such as the waiting time and service time, on which to base their decisions. In addition, we find that interaction effects are not statistically significant for lower-rated doctors (e.g., 3.5 star); but both the significance and magnitude increase as the rating increases. For higher-rated doctors, quality proxies (i.e., bedside manner, accuracy of diagnosis, and waiting time) matter more for high-income patients. These results suggest the predictive power of text reviews increases when the overall rating goes up.

Next, we conduct subset analysis for physicians based on the number of reviews. To be more specific, we consider the top 25%, 25%–50%, 50%–75%, and bottom 25% reviewed physicians. Table 11 of the online supplement presents our estimation results. We find the content of reviews matters more when the number of reviews increases. In particular, for the top 25% reviewed physicians, quality proxies from the text-review information are most important. In addition, we observe that interaction effects of the accuracy of diagnosis and waiting time are not statistically significant for less reviewed doctors, but both the significance and magnitude increase as the number of reviews increases. Intuitively, the number of reviews also reflects the popularity of the doctor; for example, Wu et al. (2015) show the economic values of service-quality features are more important for popular restaurants. Our results further indicate quality features derived from contextual reviews are more important for popular doctors.

6.4. Robustness Checks

In this section, we conduct robustness checks for our main estimation results. All the tables of estimation results of this subsection can be found in Appendix D of our online supplement.

First, the number of reviews in our main estimation only counts reviews with both a numerical rating and text context; hence, the mean of this variable is relatively low as compared with Luca and Vats (2014), namely, 26.85. Therefore, following Luca and Vats (2014), we use the total number of reviews (the mean is 51.30) to run our main model. We display our estimation results in Table 12 of the online supplement, and we find the results of the overall-rating and service-quality measures are consistent with our main results; that is, the service-quality features are statistically significant in explaining patient choices. Moreover, the interaction effects of overall rating, bedside manner, accuracy of diagnosis, and waiting time are all statistically significant.

Second, to account for the potential substitution across specialties, that is, certain services can be offered by multiple specialists, we conduct a robustness check that groups similar specialty physicians into one market, for example, the physical therapist and physiatrist, and we again find consistent results of overall rating and service-quality measures; see Table 13 of the online supplement. Moreover, the interaction effects of overall rating, bedside manner, accuracy of diagnosis, and waiting time are all consistent with the main estimation.

Third, we consider a 30-day appointment window in our main estimation. To verify that our results are robust across different time windows, we conduct robustness checks with varying appointment-window lengths (7-day, 14-day, and 60-day). Tables 14, 15, and 16 of the online supplement show our results remain consistent. In addition, from Tables 14 and 15, we can see that if we cut the 30-day window down to either a 7-day or 14-day window, the significance level as well as the magnitude of coefficients of the overall rating and sentiment variables decrease because we underestimate the demand change. Moreover, the interaction effects of the overall rating, bedside manner, accuracy of diagnosis, and waiting time are all statistically significant; but the magnitude of these effects are less for 7-day and 14-day windows.

However, the significance level as well as the magnitude of coefficients of the overall rating and sentiment variables in Table 16 (60-day window) are similar to our main estimation results. This result further confirms that a 30-day appointment window is appropriate.

Fourth, to account for potential autocorrelation, we conduct a robustness check with lagged demand. Our estimation results in Table 17 of the online supplement show consistent results of overall-rating and service-quality measures. In addition, the interaction effects of overall rating, bedside manner, accuracy of diagnosis, and waiting time are all consistent with the main estimation.

Fifth, postulating that for physicians with many reviews, patients only read a subset of their reviews, we conduct a robustness check for the top 50% reviewed physicians with part of their reviews; that is, reviews contained in the first page (on average in our data set is 23 reviews), the first 10 reviews, the first 5 reviews, and the first 3 reviews. Note that reviews on the first page of a physician's personal page can be seen by patients without clicking the "more" button. From our estimation results in Table 18 of the online supplement, we find our main results hold consistently. Moreover, when the number of reviews included in the analysis increases, quality proxies from text reviews become more important with higher magnitude. Therefore, we confirm that patients indeed read the reviews on the entire page to perceive service-quality proxies and choose physicians. We conduct this robustness check because, intuitively, people may think when a patient arrives to the platform, he or she might only read the first several reviews; however, our results show the service proxies derived from the entire page would affect physicians' demand.

Last, we use an RD design as a robustness check for the causal impact of numeric rating on demand. The overall rating displayed at the front page of the physician is the average over all the reviewer ratings, and the website rounds it to the nearest half-star. This change in rating is exogenous to physician quality but is correlated with the displayed ratings and thus provides us an ideal exogenous shock for an RD analysis. We present the details of the RD analysis in Appendix C of our online supplement. We find consistent results with Luca and Vats (2014), that is, a significant positive treatment effect of the overall rating; see Table 1 in Appendix C of our online supplement.

7. Implications, Limitations, and Conclusion

In this paper, we derive service-quality proxies from text reviews to predict physician demand, based on appointment-booking data of physicians in the United States, their demographics, service characteristics, and text reviews. Using text-mining techniques and sentiment analysis, we derive service-quality measures that are implicitly embedded in the text reviews. The ultimate goal of this paper is to establish a predictive model of patient choice, utilizing both structured and unstructured online information. Broadly speaking, with the rapid growth of online information and data, we anticipate that predictive modeling of patient choice will play a vital role in the entire healthcare ecosystem.

First, our empirical results show service-quality measures from online text reviews can improve the performance of predictive modeling of patient choice. We unveil the fact that patients indeed rely on text reviews to make their choices. Hence, our work calls for an immediate attention from doctors to carefully look into text reviews, if they want to have a better demand prediction.

Second, improving predictive modeling of patient choice and understanding the role of operational features in patient choices, for example, service time, waiting time, and so on, can help doctors better schedule their appointments and manage patient flow in their clinics. Our proposed predictive model can help manage patterns in patient flow to ensure optimal staffing levels while reducing waiting time.

Third, by mining the text reviews, we are able to identify detailed service-quality features that could affect patient satisfaction. Given the fact that patient relationship management has become a vital issue for both care providers and insurance companies, understanding the different dimensions of service-quality features can help promote healthy long-term relationships between patients and care providers.

Our work takes an initial step toward leveraging quality proxies derived from text reviews to predict physician future demand. However, our study still has some limitations, some of which might lead to future research directions.

7.1. Data

In our work, we collect the data ourselves from the online platform; therefore, the data set indeed has some limitations. First, we only observe the aggregate-level demand and not the individual choices or the patients' demographic information. Although we use a BLP-type structural model to incorporate patients' heterogeneous preference shocks, the ability to observe more granular-level data might help improve the predictive accuracy. A second limitation is related to this first one: the lack of granular-level data makes our demand information censored, in the sense that we do not know if a patient chooses physician A over B because of the online quality proxies or the lack of appointment availability of B. To address this issue, more detailed data, such as the click-through data, would be needed to construct patients' choice sets. Third, when appointments disappear from the appointment book, we consider them booked appointments and as demand for the physician. However, some bookings might disappear because of the doctor's change of schedule, which we cannot observe in our data set. However, as long as this disappearance is systematically uncorrelated with doctor and patient attributes, our results should hold. Last, all the quality features derived from the text reviews are patients' perceived healthcare quality. This information could be considered proxies for physician quality, but we still do not observe the true operations or quality of physicians. Potential selection issues related to the types of patients and experiences that trigger a patient review might exist (bad experiences are more likely to be present in the reviews). Therefore, future research might try to combine online information and offline operations information to address this issue.

7.2. Text Mining

The text-mining steps in our work are topic extraction and sentiment analysis. We consider a potential limitation in the sentiment-analysis step. Objectively measuring the different magnitudes of sentiment level can be difficult. The magnitudes of sentiment analysis measured in our work follow from SentiWordNet (Esuli and Sebastiani 2007). Nonetheless, the measurements depend considerably on these training data sets. Work on deriving a corpus of patient experience is growing (Rastegar-Mojarad et al. 2015), but text-mining research on healthcare reviews is still scant compared with Yelp reviews or Twitter data. Future work might look into better measurements of sentiment analysis and how new patients perceive the different levels of sentiment in online reviews. In addition, potential unobserved quality measures cannot be captured by the seven text-mining variables in our model, which can lead to biased estimation. Future work might come up with an innovative identification strategy to establish the precise causal relationship between quality proxies from text reviews and demand.

To our knowledge, this paper is the first to derive service-quality proxies from user-generated online information into a structural choice model to predict physician demand. Our interdisciplinary approach provides a framework of combining machine learning and structural modeling to predict future demand. In addition, our estimation results provide insights on how patients make outpatient physician choices when facing various sources of information.

Acknowledgments

The authors gratefully acknowledge the anonymous referees, associate editor, and department editor for detailed and thoughtful reviews and the helpful feedback from conference participants of the Workshop on Health IT and Economics, Production and Operations Management Society (POMS), and INFORMS Healthcare best student paper award presentations. The previous title is "The Effect of Online Reviews on Physician Demand: A Structural Model of Patient Choice."

Endnotes

[1] See https://www.softwareadvice.com/resources/how-patients-use-online-reviews/.

[2] See https://www.grandviewresearch.com/press-release/global-digital-health-market.

[3] We define active physicians as the ones who have offered appointments. The web page has around 21,000 physicians; however, many of them only have names listed on the website, so we do not include these physicians in our data set.

[4] Note that a physician may have more than one year's appointment calendar on this platform. However, we indeed observe that for the physicians in our data set, none of their patients booked slots more than 30 days in advance, which is the key reason we choose a 30-day appointment window. The word "offered" means the physician is shown as working in that specific time slot and is available for booking.

[5] We acknowledge a potential limitation of our data here is that some bookings that disappear might be due to the doctor's change of schedule; however, as long as this disappearance is systematically uncorrelated with doctor and patient attributes, our results should hold.

[6] From Luca and Vats (2014), we know most of the physicians accept the same insurance plans on this website and the variation in insurance plans among physicians is very small, so we do not include insurance as one of our control variables.

[7] Note the overall rating shown in Table 1 is the rounded average of individual ratings calculated by the website. The bedside-manner rating and waiting-time rating are the average of individual ratings we calculated without rounding.

[8] See http://code.google.com/p/negex.

[9] We follow Berry et al. (1995) and choose $I = 20$ as the total number of patients in each market. Note here the value for this I would not affect our estimation results, because we are using the aggregated demand, which is the market share, instead of individual choices.

[10] Note that $Z < K$ in our model, because the random coefficients of some physician characteristics go to zero in our estimation results.

[11] See https://www.cms.gov/medicare/health-plans/medicareadvtgspecratestats/ffs-data.html.

[12] See http://www.americannursetoday.com/patient-satisfaction-now-factors-into-medicare-reimbursement/.

References

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Berry S (1994) Estimating discrete-choice models of product differentiation. *RAND J. Econom.* 25(2):242–262.

Berry S, Pakes A (2007) The pure characteristics demand model. *Internat. Econom. Rev.* 48(4):1193–1225.

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. *Econometrica* 63(4):841–890.

Brody S, Elhadad N (2010a) Detecting salient aspects in online reviews of health providers. *AMIA Annual Sympos. Proc.*, vol. 2010 (American Medical Informatics Association, Bethesda, MD), 202–206.

Brody S, Elhadad N (2010b) An unsupervised aspect-sentiment model for online reviews. *Human Language Technologies: The 2010 Annual Conf. North Amer. Chapter Assoc. Comput. Linguistics* (Association for Computational Linguistics, Los Angeles, California), 804–812.

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

Cui R, Gallino S, Moreno A, Zhang DJ (2018) The operational value of social media information. *Production Oper. Management* 27(10): 1749–1769.

Dong J, Yom-Tov E, Yom-Tov GB (2019) The impact of delay announcements on hospital network coordination and waiting times. *Management Sci.* 65(5):1969–1994.

Esuli A, Sebastiani F (2007) Sentiwordnet: A high-coverage lexical resource for opinion mining. *Evaluation* (European Language Resources Association (ELRA), Genoa, Italy), 1–26.

Fellbaum C (1998) *Wordnet: An Electronic Lexical Database* (MIT Press, Cambridge, MA).

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.

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

Ghani R, Probst K, Liu Y, Krema M, Fano A (2006) Manipulation of the running variable in the regression discontinuity design: A density test. *SIGKDD Explorations* 1(8):41–48.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10):1498–1512.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. *Marketing Sci.* 31(3):493–520.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Harris K, Buntin M (2008) Choosing a healthcare provider: The role of quality information. *Policy* 1. 1(6):1–14.

Hausman JA (1996) Valuation of new goods under perfect and imperfect competition. *The economics of new goods*. (University of Chicago Press), 207–248.

Hu M, Liu B (2004) Mining and summarizing customer reviews. *Proc. 10th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining (KDD04)* (ACM, New York), 168–177.

Hutto CJ, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. *8th Internat. AAAI Conf. Weblogs Social Media*. (AAAI Publications, Ann Arbor, Michigan).

Liu N, Finkelstein SR, Kruk ME, Rosenthal D (2018) When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Sci.* 64(5):1975–1996.

Luca M (2011) Reviews, reputation, and revenue: The case of yelp.com. Working paper, Harvard Business School NOM Unit, Boston.

Luca M, Vats S (2014) *Digitizing Doctor Demand: The Impact of Online Reviews on Doctor Choice. Health & Healthcare in America: From Economics to Policy* (Ashecon).

Manning C, Schutze H (1999) *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA).

McFadden D (1978) Modelling the Choice of Residential Location. Spatial Interaction Theory and Planning Models, A. Karlgvist (ed.). (Amsterdam, North-Holland). Institute of Transportation Studies, University of California.

McLachlan G, Do K, Ambroise C (2005) *Analyzing Microarray Gene Expression Data*, vol. 422 (John Wiley & Sons, Hoboken, NJ).

McLaughlin GH (1969) Smog grading: A new readability formula. *J. Reading* 12(8):639–646.

Mein Goh J, Gao G, Agarwal R (2016) The creation of social value: Can an online health community reduce rural-urban health disparities. *MIS Quart.* 40(1):247–263.

Nelder JA, Mead R (1965) A simplex method for function minimization. *Comput. J.* 7(4):308–313.

Nevo A (2001) Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69(2):307–342.

Pulia M (2011) Simple tips to improve patient satisfaction. *Amer. Acad. Emergency Medicine* 18(1):18–19.

Rastegar-Mojarad M, Ye Z, Wall D, Murali N, Lin S (2015) Collecting and analyzing patient experiences of healthcare from social media. *JMIR Res. Protocols* 4(3):e78.

Razzouk N, Seitz V, Webb J (2004) What's important in choosing a primary care physician: An analysis of consumer response. *Internat. J. Health Care Quality Assurance* 17(4):205–211.

Renter E (2015) Why nice doctors are better doctors. *US News & World Reports* (April 20), https://health.usnews.com/health-news/patient-advice/articles/2015/04/20/why-nice-doctors-are-better-doctors.

Sargan D, Desai M (1988) *Lectures on Advanced Econometric Theory* (Blackwell, Portland, Oregon).

Silverman BD (2012) Physician behavior and bedside manners: The influence of William Osler and the Johns Hopkins School of Medicine. *Proc. Baylor University Medical Center* 25(1):58–61.

Song M (2011) A hybrid discrete choice model of differentiated product demand with an application to personal computers. *Internat. Econom. Rev.* 56(1):265–301.

Staiger D, Stock JH (1997) Instrumental variables regression with weak instruments. *Econometrica* 65(3):557–586.

Steckel J, Vanhonacker W (1993) Cross-validating regression models in marketing research. *Marketing Sci.* 12(4):415–427.

Tay A (2003) Assessing competition in hospital care markets: The importance of accounting for quality differentiation. *RAND J. Econom.* 34(4):786–814.

Varkevisser CM, Van der Geest SA, Schut FT (2012) Do patients choose hospitals with high quality ratings? Empirical evidence from the market for angioplasty in the Netherlands. *J. Health Econom.* 31(2):371–378.

Villas-Boas JM, Winer RS (1999) Endogeneity in brand choice models. *Management Sci.* 45(10):1324–1338.

Wang G, Li J, Hopp WJ, Fazzalari FL, Bolling S (2019) Using patient-specific quality information to unlock hidden healthcare capabilities. *Manufacturing Service Oper. Management* 21(3): 582–601.

Wansbeek J, Meijer E (2000) *Measurement Error and Latent Variables in Econometrics*, vol. 37 (North-Holland, Amsterdam, Netherlands).

White S (2006) Key concepts and features of the 2003 national assessment of adult literacy. Report, National Center for Education Statistics, U.S. Department of Education, Washington, DC.

Wu C, Che H, Chan T, Lu X (2015) The economic value of online reviews. *Marketing Sci.* 34(5):739–754.