What Can Online Doctor Reviews Tell Us? A Deep Learning Assisted Study of Telehealth Service

Haijing Hao
Bentley University
hhao@bentley.edu

Bin Zhang
Texas A&M University
bzhang@mays.tamu.edu

Yongcheng Zhan
Cal Poly
yozhan@calpoly.edu

Jiang Wu
Wuhan University
jiangw@whu.edu.cn

Abstract

The present study develops a novel deep learning method which assists text mining of online doctor reviews to extract underlying sentiment scores. Those scores can be used to estimate a healthcare service quality model to investigate how the online doctor reviews impact the online doctor consultation demand. Based on the data from the largest online health platforms in China, our model results show that the underlying sentiment scores have statistically significant impacts on the demand of online doctor consultation. Theoretically, the present study constructs an innovative deep learning algorithm with a better performance than four widely used text mining methods, which can be applied to text mining of many online forums or social media texts. Empirically, our model results show what factors impact the health service quality and online doctor consultation demand, and following those factors, healthcare professionals can improve their service.

Keywords: Online doctor consultation service, online doctor review, online consultation demand, text mining, deep learning, sentiment analysis.

1. Introduction

Online doctor consultation service (ODCS) has become popular in recent years, particularly since the COVID-19 pandemic, because of its convenience, safety of minimizing spread of COVID, cost effectiveness, and time efficiency (Jiménez-Rodríguez et al. 2020). In ODCS, patients consult physicians about their health problem over the internet or internet facilitated telephone consultations. Besides hospitals and clinics, many public platforms also contracted with physicians to provide ODCS service, which can empower patients to have more flexibilities to find matching doctors for their health concerns. Public ODSC is particularly popular in China because China's state-owned hospitals concentrate on off-line services only. In China, uneven distribution and development of the economy and medical facilities also cause most skilled physicians to work in the top ranked hospitals in a few metropolitan areas (Hao 2015). Hence, ODCS provides a new and unique channel for patients from non-metropolitan areas or rural areas to reach good doctors, avoiding traveling across a big country to see a doctor for a half an hour consultation in person.

Usually those ODCS platforms list a physician's basic information such as specialty area, degree, medical school, technical title, years of experience, as well as other online special information such as years on the platform, health related articles or information shared on the homepage, the number of virtual gifts that they received from patients, previous patients' quantitative ratings and qualitative reviews. Hao (2015) found that since patients could review their doctors online in China, a myriad of reviews have been accumulated online and become a rich asset of individual patients' experiences and sentiments about doctors' service. However, little is known about whether online doctor reviews have any influence on patients' decision about choosing a doctor for online consultation. Some literature showed that online doctor reviews can affect patients' choice of primary care doctors (Yaraghi et al. 2018), while another study also showed that many patients do not rely on online reviews (Lee 2016). A recent paper also found that online doctor reviews affect off-line doctor demand by using a structural model (Xu et al. 2021), and the affecting factors included bedside manner, diagnosis accuracy, waiting time, and service time. Online doctor reviews by patients provide a wealth of information and firsthand experience of a patient's interactions with their doctors, containing subtle sentiments of patients, which may influence other patients' choice of doctors for online consultation, and affect online doctor consultation demand (Piccoli and Pigni 2013).

URI: https://hdl.handle.net/10125/102896
978-0-9981331-6-4

Thus, the present study wants to investigate whether the underlying sentiments of online doctor reviews may affect patients' choice of online doctor consultation. To solve this problem, we compile a comprehensive dataset from the largest ODCS in China, Good Doctor Online (www.haodf.com), which was established in 2006 and there are more than 890,000 doctors with millions of reviews from over 10 thousand hospitals across China. We crawled a dataset within eight weeks from Good Doctor Online, then apply a novel deep learning algorithm to text mining of those online reviews. Then we utilize the text mining results to estimate a widely used health care service quality model, to explore the factors affecting healthcare service quality, thus influence patients' decision of choosing doctors for online consultation. Theoretically, we construct an innovative deep learning algorithm which can be applied to text mining of many online forums or social media posts. Empirically, our econometric model results show what factors may online doctor consultation demand, which is a proxy of the healthcare service quality, which may provide implications for online healthcare practitioners how to improve their online service.

The rest of the paper is organized as follows. In Section 2 we review the related literature. Section 3 introduces the research context and hypotheses development. Section 4 describes the data and methods. Section 5 is the results. Section 6 provides both theoretical and practical implications.

2. Literature review

2.1 Online doctor reviews

Online doctor reviews have been available in many countries since 2000 with Web 2.0 becoming popular, and prior research found that more and more physicians received online reviews worldwide, such as in the U.S. (Gao et al. 2012), U.K. (Greaves et al. 2012), Germany (Emmert 2013), and China (Hao 2015). More online reviews, particularly narrative reviews are available. These provide a new source for evaluating a doctor's healthcare service because the reviews are from many patients' personal experience, writing at home with ease and no pressure. Combining fast developing text mining technology, many studies have been focusing on investigating the rich information from the vast textual posts in online health communities, such as extracting the underlying topics from online consultation forum where patients posted their health question publicly on the forum and a doctor may volunteer to answer (Hao and Zhang 2015), or topic modeling online doctor reviews to examine what are patients talk about their doctors (Hao and Zhang 2016; Hao et al. 2017), or combining topic modeling and qualitative analysis of patient self-support online health community (Hao et al. 2020). However, the above text mining studies of online doctor reviews or online healthcare service were limited to the topic modeling method, which can only extract topics, not sentiments, from online textual posts. That is, we can know that many online reviews are about a doctor's bedside manner, but we cannot know it is a good manner or a bad manner.

2.2 Healthcare service quality model

Prior research showed that enabling patients to choose doctors can improve patient satisfaction (Kersnik 2001). Patients consider the quality of healthcare as the most important factor in doctor selection (Bornstein et al. 2000). Owing to the intangibility, inseparability and heterogeneity of healthcare service (Parasuraman and Berry 1985), it is difficult to directly measure a doctor's quality of service.

Among many healthcare service quality evaluation models, a healthcare quality assessment framework proposed by Institute of Medicine (IOM) is widely used (IOM 2020), which includes six dimensions: safety, effectiveness, patient-centeredness, timeliness, efficiency, and equity (SEPTEE). Safety means that patients should not be harmed by the healthcare service which is supposed to help them. Effectiveness means that the healthcare service should be evidence-based to determine what intervention, diagnostic test, or therapy should be used to treat the patient. Patient-centeredness means that the healthcare provided should be respectful of and responsive to each patient preferences, needs, and values, which guide all clinical decisions during the service. Timeliness means that when providing service, the healthcare system should reduce waiting time and harmful delays for both patients and providers. Efficiency means that a healthcare system should reduce quality waste and reduce administrative or product costs. Equity means that providing healthcare should not vary in quality because of patients' gender, ethnicity, geographic location, and socioeconomic status. Pillittere et al. (2003) also found that patients particularly cared about effectiveness, safety, and patient centeredness when evaluating the quality of healthcare service.

For the present study, we are interested in evaluating the quality of telehealth, ODCS, which has its own special perspectives because the service is entirely online and is different from regular health service. For example, timeliness would not be a big problem for telemedicine because patients seeing a doctor online usually is a direct meeting between a patient and their doctor at a given time, and no waiting time in waiting rooms or dealing with hospital staff procedure. Safety issues are minimized too because online consultation focusing on consultation only, no tests or medical procedures. Efficiency in a regular in-person health system is more about the system's administrative and quality controls, which do not apply to online consultation much.

3. Research hypotheses

The existing literature rarely studies doctor reviews' effect on patients' choosing doctors, which is an emerging area of consideration (Vennik et al. 2014). Patient's selection of doctor can be measured by doctor's consultation service reserved by patients. We follow NAM's healthcare quality assessment framework and define six dimensions – effectiveness, patient-centeredness, timelessness, efficiency, equity and safety, to study their effects on each doctor's consultation demand, measured by the count of consultation per week. We also control for doctor's online forum behaviors, such as his/her online tenure, and online posts, which may affect consultation demand. Figure 1 shows the conceptual model of this paper.

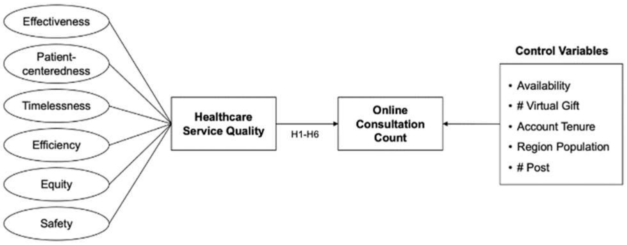

Figure 1: Conceptual Model

Patients tend to choose those doctors who they believe have greater medical competence and can provide high quality of service. We first test the impact of six dimensions of service quality from IOM's SEPTEE model on the demand of ODCS. Prior research showed that a positive review will increase purchase intention while a negative review may lower purchase intention (Zhu and Zhang 2013). We propose that the demand of ODSC is positively associated with the score about her or his service quality. The higher the sentiment score a patient reflects in his or her review, the higher the service quality the doctor gives to the patient. The three dimensions from SEPTEE model that fit the online context seamlessly are: effectiveness, patient-centeredness, timelessness. Therefore, we have the following three hypotheses first:

H1: A doctor's ODCS demand is positively correlated with the score of her or his effectiveness.

H2: A doctor's ODCS demand is positively correlated with the score of his or her patient-centeredness.

H3: A doctor's ODCS demand is positively correlated with the score of her or his timeliness.

The scores of each dimension of healthcare service quality are derived from online doctor reviews. With the help of Natural Language Processing (NLP), researchers are able obtain more information, such as topics and sentiments, from the user-generated content (Melumad et al. 2019), including online reviews. To extract the six dimensions of healthcare service quality and the correspondent sentiment from patients, we use NLP technologies including deep learning. We developed a novel deep learning method, D2VL, that can full leverage the strength of both Doc2Vec and long short-term memory (LSTM). The detail of our D2VL method is described in the next section.

In the IOM's SEPTEE model, there are also three dimensions that do not seem to fit the online context very well, and are likely not significant – efficiency, equity, and safety. For these three dimensions, we test their significance of effect on ODCS demand first. Therefore, we hypothesize that:

H4: A doctor's ODCS demand is NOT correlated with score representing the efficiency of the hospital where he or she works.

H5: A doctor's ODCS demand is NOT correlated with the score representing the equity of the platform.

H6: A doctor's ODCS demand is NOT correlated with the score representing the safety of her or his consultation service.

4. Data and model

4.1 Data collection

To test our research hypotheses, we collected data from Good Doctor Online for eight weeks. This dataset contains 79,802 online doctor reviews of 3,142 distinct doctors, and consists of reviews of 10 specific kinds of disease, including 5 high-risk diseases (leukemia, lung cancer, cirrhosis, coronary heart disease, and diabetes) and 5 low-risk diseases (hypertension, rheumatoid arthritis, gastritis, depression, and menoxenia). The risk intensity is defined according to the mortality for major diseases from the China Health Statistics Yearbook 2016 and prior study (Yang et al. 2015). A dummy variable is thus created to label the disease risk, with 1 representing high risk and 0 representing low risk.

All doctors' individual characteristics are scraped from their homepage on this online platform, which includes technical title, hospital tier, available time for online consultations, the number of health-related posts or articles shared on the doctor's homepage, the number of virtual gifts received on the platform, and quantitative ratings and qualitative reviews posted by their patients on this platform. More online reviews, particularly narrative reviews are available (see Table 1 for sample reviews). The doctor title reflects doctor's skill and working experience, which is categorized into four types, senior physician, associate senior physician, attending physician, and resident physician, from the highest skill to the lowest skill and (4 to 1 in our dataset), accordingly. The tier of hospital reflects a hospital's overall quality based on the hospital's capability in operation, infrastructure, and technical strength, according to the criteria published by Chinese government. There are three tiers, and the higher of the tier number, the higher overall service quality of the hospital.

Table 1. Example Reviews Translated into English

No.	Example Review
1	I was diagnosed with xxx (disease) in March 20xx. Thanks to the website, I was able to find Dr. (name). He is kind and patient. He asked me about my health history and medicines taken in very much detail. I knew that he has treated many patients before me. But he still maintains a very nice attitude. He explained to me the causes of my condition. Then he prescribed me some new medicine. He also gave me suggestion about change in meal. He provided wonderful treatment plan via consultation. I had been to multiple hospitals for treatment. But none of these treatments was successful. My condition is finally cured. I am very grateful to Dr. (name).
2	I have been on this website asking for two years. And have gotten consultation from many doctors on the platform. Dr. (name) is the best I've ever met. She really cares about her patients. She is patient, tender and soft. She always smiles to me. Her personality is very calming. Although my condition is chronic and may not have immediate cure, she still shows optimism and cheers me up. Her conversation reduces my anxiety. She persuaded me that my condition can be better. Due to her commitment, I strictly follow her treatment plan. Her plan finally complete cures my condition! I highly recommend her to other patients.

4.2 Deep learning methods

With the help of NLP technology, researchers are able to obtain more information efficiently from vast user-generated online content. To extract the six dimensions of SEPTEE healthcare service quality model and the corresponding sentiment scores from a myriad of online doctor reviews, we developed a novel deep learning assisted text mining method to analyze sentiments of online doctor reviews. In the following, we build and compare five different text mining methods, including our innovative deep learning assisted text mining method.

First, we build a classifier to label the topic embedded in the review text, then calculate the sentiment score following the same rule for each of the topics, by using widely accepted support vector machine (SVM) method.

Our second benchmark method is a supervised learning model measuring the sentiment directly. Instead of using the text to predict the topic of the review, the text is used to predict the sentiment score directly. We use a neural network regressor to build this method.

The third benchmark method, a recently advanced method of the vector representation of a document is Document-to-vector, or Doc2Vec, which represents a set of documents using numerical vectors, taking the distributional context information into consideration (Dai et al. 2015). The Doc2Vec method is developed upon the famous Work2Vec model which utilizes the surrounding context words to find the representation of the focal word (Mikolov et al. 2013). We use this trained document vector to represent the health service reviews and to predict the sentiment score of the six service quality dimensions.

The fourth benchmark method emphasizes, instead of a single document representation, the linguistic sense embedded in the link and connection of words (Hochreiter and Schmidhuber 1996). Thus the input of this method is not a single document vector, but a sequence of word vectors with their temporal dynamics. The length of the text in this problem is large, and thus requires the use of a deep learning model to handle the massive input of different word vectors. We used a long short-term memory (LSTM) network to cope with the problem of "memorizing useful information," or technically, the problem of vanishing gradient, in the implementation of the deep recurrent neural network. In this method, we first build a Word2Vec model and convert words to numerical vectors. Then, we feed the numerical vectors to a LSTM model to generate a representation of the document by using the output vector of the last

LSTM unit. Finally, we use this vector to predict the sentiment score of each review.

The last method, Doc2Vec-LSTM (D2VL), is our self-developed deep learning method for extracting multiple topics about service quality from each online review.

Sentiment analysis about online doctor review presents very hard challenges. Most of the NLP methods either infer a text's meaning at the word level, e.g. LSTM, or at the text level, e.g. Doc2Vec. The weakness of LSTM is that word with different meanings is still represented as one word. LSTM also does not consider the semantic at the level of document, so the context for words is lost. The weakness of Doc2Vec is that it does not consider the sequence of words or sentences. So documents with the same words but with very different sentence structure are interpreted as the same meaning. Our new method, D2VL, combines the advantages of both Doc2Vec and LSTM.

Combining the strengths of both LSTM and Word2Dec, we obtain both the embedding of the while document, which is learned from the Doc2Vec model, and the word embedding of the context relations, which is learned from the sequence of the LSTM network. The combination of the two embeddings of the text can bring both global and local linguistic sense, and thus capture more accurate features to learn the sentiment score of all service quality dimensions in a given online review. The architecture of this proposed model is shown in Figure 2.

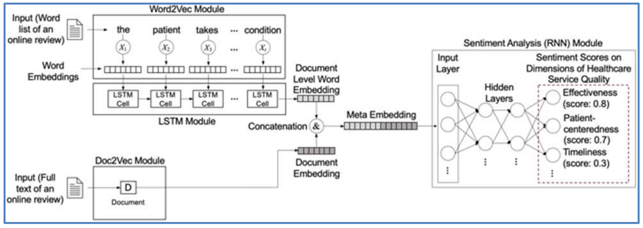
Word2Vec Module
Input (Word list of an online review)
the
patient
takes
···
condition
X_1
X_2
X_3
X_t
Word Embeddings
LSTM Cell
LSTM Cell
LSTM Cell
LSTM Cell
LSTM Module
Document Level Word Embedding
Concatenation
&
Meta Embedding
Document Embedding
Doc2Vec Module
Input (Full text of an online review)
D
Document
Sentiment Analysis (RNN) Module
Input Layer
Hidden Layers
Sentiment Scores on Dimensions of Healthcare Service Quality
Effectiveness (score: 0.8)
Patient-centeredness (score: 0.7)
Timeliness (score: 0.3)

Figure 2. Framework of Our Newly Designed Deep Learning Model – D2VL

4.3 Textual data processing

We drew a random sample of 1,020 reviews from the dataset and had them manually labeled for the six dimensions of SEPTEE model by two domain experts. The data labeling process took two steps. First, each of labeler thoroughly examined the text and labeled whether the text contained the discussion of the above mentioned six dimensions. Second, based on a predefined sentiment rule, the reviewer calculated a sentiment score for each of the six dimensions. The sentiment rule was developed based on the HowNet sentiment lexicon (Dong and Dong 2003) that includes lists of words with predefined emotional weight. In addition, adverbs and negation words are also included in the calculation of the sentiment score. The predefined sentiment rule is tested with 100 pre-labeled reviews and achieved 97% accuracy. After this step, we obtained a labeled dataset of 1,020 reviews with their topics and corresponding sentiments.

In the data preprocessing step, all the stop words and spaces from the review text are removed. Each review is labeled by doctor's unique identification in order to calculate the mean of sentiment score for the focal doctor. We also processed the Chinese text according to its language characteristics, and used the Python Jieba segmentation tool, including medical dictionaries from different diseases added into segmentation processing for more precise results.

Table 2. Performance Comparison for Learning the Sentiment Scores for Healthcare Service Quality

Compared Methods	μ_{MSE} (s.d.)
Benchmark 1: SVM	0.42 (0.04)
Benchmark 2: Neural Network	0.34 (0.02)
Benchmark 3: Doc2Vec	0.50 (0.03)
Benchmark 4: LSTM	0.36 (0.03)
Our Proposed Method: D2VL	**0.20** (0.03)

We randomly split our labeled sample dataset into a training set and a testing set, with an 80:20 rule. Then we tested performance of the above five text mining methods. Since the goal of NLP in this study is to generate a sentiment score for each of the six healthcare service quality dimensions which will be used to estimate online doctor demand, it is not a categorical classification task; instead, each score is numeric. Thus, we used the mean squared error (MSE) to measure the performance of the five methods, as listed in Table 2, and it shows that our proposed D2VL model achieves the best performance, or the lowest MSE value.

4.4 Descriptive Statistics

Table 3 and Table 4 show the descriptive statistics of our data, including both collected demographic data of online doctors and sentimental scores from our deep learning method. ODCS demand is defined as the number of ODCS appointments received by the focal doctor in a week on this platform. On average, a doctor may receive 1.6 online appointments and the popular doctor would receive 80 appointments; the distribution is right-skewed. We also can see that most doctors have the highest technical title, senior physician, and are affiliated with Tier 3 hospitals, the highest level hospital tier in China. In a week, the maximum number of posts a doctor shared is 44 and the minimum is 0 their homepage on this platform. The maximum number of virtual gifts a doctor received in a week in is 153 and the minimum is 0. We want to mention that Title and HospTier are associated with the Effectiveness dimension as well because they both can reflect a doctor's technical experience and quality, otherwise he or she would not be hired by a higher tier hospital or would not be promoted to a senior physician. The number of health related post on a doctor's homepage on this platform and the number of virtual gift a doctor received on this platform are highly correlated to a doctor's attitude towards their patients, showing how much they care about their patients. Hence these two measures are related to the Patient-centeredness dimension. The six sentiment scores are derived from online doctor reviews along the six dimensions of the SEPTEE model.

Table 3. Descriptive Statistics of Dependent Variable and Sentiment Scores

Variable	Description	Mean/ Percentage	SD	Min	Max
ODCS Demand	# of ODCS appointments received by a doctor in a week	1.61	3.50	0.00	80
Effectiveness	Doctor's service effectiveness score in a given week, derived from reviews	0.15	0.21	-1.31	1.48
Patient-centeredness	Doctor's score w.r.t being respectful to patient in a given week, derived from reviews	0.61	0.46	-0.06	1
Timeliness	Doctor's service promptness score in a given week, derived from reviews	0.83	0.23	-0.45	1.55
Efficiency	Efficiency of the institution in a given week, derived from reviews	0.59	0.45	-0.15	1.87
Equity	Score about the healthcare system being equal to everyone in a given week, derived from reviews	0.61	0.43	-0.25	1.21
Safety	Score about patient not being harmed by the ODCS in a given week, derived from reviews	0.23	0.19	0.00	0.43

In addition, we include a few control variables as Table 4 shows. A doctor who has been on this platform longer may be more experienced in online consultation and built up their reputation already, thus may have higher demand. Hence, Account Tenure shows that a doctor can be as long as nine years on this platform, and the average is 5.6 years. Available time for online appointment is another control variable because some doctors may be busy with off-line work and are not available for too many online appointments. A doctor may have maximum 12 appointments available in a week and the mean available time slots is 1.5. Population of the province where a doctor's hospital is located and the disease risk category are another two control variables to mitigate the cofounding factors of our analysis. For individual level data's descriptive statistics, such as Title, HospTier, Account Tenure, disease risk, and province population, the number of observations is 3,142 doctors and for all other panel data's descriptive statistics are based on the entire 23,638 observations.

We construct our econometric model based on the SEPTEE model as follows.

$$ODCS\_Dem_{i,t} = \beta_1 Title_i + \beta_2 HopsTier_i + \beta_3 \log(Post)_{t-1} + \beta_4 \log(VirtualGift)_{t-1} + \beta_5 Effectiveness_{t-1} + \beta_6 Patient\text{-}centeredness_{t-1} + \beta_7 Timeliness_{t-1} + \beta_8 Efficiency_{t-1} + \beta_9 Equity_{t-1} + \beta_{10} Safety_{t-1} + \beta_{11} AccountTenure_i + \beta_{12} AvailableTime_{t-1} + \beta_{13} \log(Population)_i + \gamma_i DeseaseDummy_i + \varepsilon \qquad (1)$$

The dependent variable, ODCS_Dem, is the number of online consultation a doctor *i* received at week *t*. The individual level characteristics include a doctor's title, a doctor's hospital tier, the number of health-related posts that a doctor shared on their homepage on this platform in week (t-1), and the number of virtual gifts a doctor received in week (t-1). The major predictor variables are the six sentiment scores derived from each doctor's online reviews about service quality in each week: effectiveness, patient-centeredness, timeliness, efficiency, equity, and safety. We also control for a doctor's available time for online consultation, a doctor's tenure time on the platform, the population of the province where a doctor works, and whether the disease is high risky or not. We take the natural logs of the highly skewed quantitative variables: number of post, number of virtual gifts, and the provincial population. We use the negative binomial regression to estimate the model because the response variable is a count value and its mean and variance are not similar, so the assumption of Poisson distribution does not satisfy.

Table 4 Descriptive Statistics of Control Variables

Control Variables	Description	Mean/ Percentage	SD	Min	Max
Post	Total # of health related posts a doctor shared on their homepage on the platform up to a given week	0.07	0.63	0	44
VirtualGift	Total # of virtual gifts a doctor received up to a given week	0.91	3.02	0	153
Title = 1	Resident Physician	0.1%			
Title = 2	Attending Physician	11%			
Title = 3	Associate Senior Physician	39%			
Title = 4	Senior Physician	49%			
HospTier = 1	Level 1	0.2%			
HospTier = 2	Level 2	0.4%			
HospTier = 3	Level 3	99.4%			
Online Tenure	# of years a doctor registered on the platform	5.57	2.68	0	9
Available Time	Doctor's available time for appointment in a given week	1.47	1.11	0	12
Population	Population of the province where the doctor works	4931	3179	593	10,999
Disease risk high	Whether the disease has a high mortality rate	62.5%			

5. Results

5.1 Empirical Results

Table 5 presents the estimation results for the three models. Model 1 contains a few predictors that can be observed by patients directly from the platform, no sentiment scores or the disease risk dummy variable. A doctor's title and a doctor's hospital tier reflect their effectiveness dimension in SEPTEE model. Both the title and hospital tier measures, which signal a doctor's technical skills and quality of the treatment, have statistically significant positive effects on the ODCS appointments received by a focal doctor. The number of health-related posts that a doctor posted on their homepage and the number of virtual gift that this doctor received on the platform reflects the interaction or how much a doctor centers on their patients, which are associated with the patient-centeredness dimension in SEPTEE model. The number of virtual gifts from patients shows statistically significant effects on the appointment demand, which suggests that patient-centeredness has an influence on the ODCS demand.

Table 5. Regression Results of Our Empirical Models

	Model 1	Model 2	Model 3
Title	0.28*** (0.031)	0.25*** (0.031)	0.28*** (0.031)
HospTier	0.46*** (0.11)	0.47*** (0.11)	0.46*** (0.11)
#Post	0.039 (0.045)	0.045 (0.045)	0.047 (0.045)
#VirtualGift	0.052* (0.023)	0.054* (0.023)	0.059** (0.023)
Effectiveness			0.76*** (0.10)
Patient-centeredness			0.18** (0.063)
Timeliness			0.22* (0.10)
Efficiency			0.012 (0.076)
Equity			0.36 (0.22)
Safety			0.24 (0.18)
Control Variables			
Account Tenure	-0.038*** (0.0088)	-0.040*** (0.0086)	-0.041*** (0.0093)
Available Time	0.036*** (0.0073)	0.035*** (0.0069)	0.033*** (0.0066)
Population	-0.065* (0.032)	-0.062 (0.032)	-0.061 (0.032)
Disease Dummy	No	Yes	Yes
Pseudo R-squared	0.31	0.35	0.72
*** p<0.001, ** p<0.01, * p<0.05. *N*=23,638			

Model 2 includes all the independent variables that Model 1 has but additional with the dummy variable of disease risk. Model 2's results are consistent with those of Model 1, confirming the results about job title, hospital tier, and the number of virtual gift are robust, even after diseases are controlled.

Model 3 expands Model 2 by adding sentiment scores of the six dimensions of SEPTEE model extracted from online qualitative reviews. In regard to these six dimensions of service quality, the results revealed that sentiment scores of three dimensions, effectiveness, patient-centeredness, and timeliness, have statistically significant impacts on the ODCS demand of focal doctors. The results indicate that a doctor's service quality, embedded in online reviews, affects the demand of a doctor's online consultation. Effectiveness' two direct measure, job title and hospital tier, are both significant in Model 3 again. Thus, hypothesis H1 is supported that effectiveness of a doctor's service has a positive impact on ODCS demand. One direct measure of Patient-centeredness dimension, the number of virtual gift received, is also statistically significant in Model 3, together with the two indirect measures extracted from online reviews, patient-centeredness score and timeliness score, are positive and statistically significant, respectively. Hence hypotheses H2 and H3 are supported too. On the contrary, three other dimensions of healthcare service quality do not have statistically significant influences on the ODCS appointment demand, i.e., efficiency, equity, and safety. Therefore, hypotheses H4, H5, and H6 are also confirmed that these three dimensions are about the policy, organization, and attribute of hospitals, which are not significantly reflected in online doctor reviews.

Finally, the average value of VIF in the three models are 1.96, 1.91, and 2.244 respectively, and all of them are less than 5, showing that no multicollinearity existing in all three Models.

5.2 Robustness Checks

As in many studies attempting to identify causality, endogeneity might occur in our regression model. One potential problem is that doctors receiving high online review sentiment scores might be systematically different from their counterparts. In this section, we use regression method combined with the propensity score matching (PSM) method to address this potential problem. In the PSM, we define the treatment group as doctors with high effectiveness sentiment scores in the 75th percentile or higher. We then create the control group using PSM to choose doctors from the rest. We ensure that the doctors in both control and treatment groups are comparable in terms of observable characteristics, also called confounders, including account tenure on the platform, the province where the hospital is located etc. We then re-estimate Equation (1). The estimation results about the predictors which are estimated by using the treatment group based on effectiveness sentiment score are shown in the "Effectiveness As Treatment" column of Table 6. As we can see from the table, all the estimated parameters are very similar to those of Model 3 in Table 5. Similarly, we also create control and treatment groups based on patient-centeredness, and timeliness, and re-run the regression model on the balanced control-treatment groups data and obtain the results in the "Patient-centeredness As Treatment" and "Timeliness As Treatment" columns in Table 4. All the results are very similar to those of Model 3 in Table 5. All the three models' outcomes after PSM show our results are robust.

Table 6. Regression Results after Propensity Score Matching

	Effectiveness As Treatment	**Patient-centeredness As Treatment**	**Timeliness As Treatment**
Title	0.279*** (0.031)	0.282*** (0.031)	0.285*** (0.031)
HospLevel	0.451*** (0.114)	0.461*** (0.115)	0.462*** (0.114)
Effectiveness	0.350*** (0.045)	0.654*** (0.097)	0.762*** (0.110)
Patient-centeredness	0.160* (0.066)	0.142* (0.064)	0.238** (0.055)
Timeliness	0.205* (0.102)	0.372*** (0.095)	0.266*** (0.084)
Efficiency	0.015 (0.082)	0.023 (0.047)	0.017 (0.074)
Equity	0.39 (0.21)	0.46 (0.27)	0.34 (0.22)
Safety	0.22 (0.17)	0.28 (0.18)	0.26 (0.16)
*** p<0.001, ** p<0.01, * p<0.05. *N*=23,638			

6. Implications

This research has the following implications. First, we utilize sentiment scores from online doctor reviews to estimate a healthcare service quality model to investigate how the online doctor reviews impact the online doctor consultation demand. To the best of our knowledge, our study is one of the first using online doctor reviews to estimate healthcare service quality model and the demand of ODCS. Second, we also develop a novel deep learning algorithm to improve the text mining approach to extract the underlying sentiment scores of patients based on online doctor reviews. Our new method, D2VL, combines the advantages of both Doc2Vec and LSTM which can interpret a word's meaning from both the words that are close it and from the entire document. Our new method is the first in the domain. This deep learning assisted text mining method can be applied to analyze sentiments or opinion mining of online forums or social media posts. Third, our empirical model finds that three dimensions of the SEPTEE model, effectiveness, patient-centeredness, and timeliness, have statistically significant effects on ODCS demand, which can help doctors improve their service and to design a better strategy of online consultation service.

This research has limitations. First, we do not have patients' data which may bring endogeneity to our model for patients' choice of online consultation.

Second, we just study one ODCS platform so the generalizability may be questioned. The future research should focus on including more data from the patient side and also utilizing more data from different ODCS platforms. Third, our theoretical model SEPTEE was derived from a traditional healthcare service quality evaluation model in the U.S. This model may have its own limitations to online service, and to the data outside of the U.S. or from another culture. At last, we did not consider health insurance's impact on patients' choice in our model. In China, health insurance does not cover online doctor consultation thus it would not be a problem for the empirical model of the present study, however, the future study should control for health insurance's impact.

7. References

Bornstein, B., Marcus, D., & Cassidy, W. 2000. "Choosing a doctor: an exploratory study of factors influencing patients' choice of a primary care doctor." Journal of Evaluation in Clinical Practice (6:3), pp. 255-262.

Dai, A.M., Olah, C. and Le, Q.V., 2015. Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998.

Dong, Z., and Dong, Q. 2003. "HowNet - A Hybrid Language and Knowledge Resource." Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China.

Emmert M, Meier F. 2013. An analysis of online evaluations on a physician rating website: evidence from a German public reporting instrument. J Med Internet Res 2013;15(8):e157

Gao, G., Greenwood, B.N., Agarwal, R., and McCullough, J. 2015. "Vocal Minority and Silent Majority: How Do Online Ratings Reflect Population Perceptions of Quality?" Mis Quarterly (39:3), pp. 565-589.

Greaves F, Pape UJ, Lee H, Smith DM, Darzi A, Majeed A, et al. 2012. Patients' ratings of family physician practices on the internet: usage and associations with conventional measures of quality in the English National Health Service. J Med Internet Res 2012;14(5): e146

Hao H. 2015. The Development of Online Doctor Reviews in China: An Analysis of the Largest Online Doctor Review Website in China J Med Internet Res 2015;17(6):e134. URL:

Hao Haijing and Zhang Kunpeng. 2016. The Voice of Chinese Health Consumers: A Text Mining Approach to Web-Based Physician Reviews. J Med Internet Res 2016;18(5):e108.

Hao, Haijing and Zhang, Kunpeng. 2015. "Text Mining Patient-Doctor Online Forum Data from the Largest Online Health Community in China" (2015). AMCIS 2015 Proceedings. 20.

Hao, Haijing, Kunpeng Zhang, Weiguang Wang, Gordon Gao. 2017. A tale of two countries: International comparison of online doctor reviews between China and the United States. International Journal of Medical Informatics, Volume 99, 2017, Pages 37-44.

Hao, Haijing, Sue Levkoff, Weiguang Wang, Qiyi Zhang, Hongtu Chen, Dan Zhu. 2020. Studying Online Support for Caregivers of Patients With Alzheimer's Disease in China: A Text-Mining Approach to Online Forum in China. International Journal of Healthcare Information Systems and Informatics. Vol. 15 (4): 1-17.

Hochreiter, S. and Schmidhuber, J., 1996. LSTM can solve hard long time lag problems. Advances in neural information processing systems, 9.

IOM. 2020. "Understanding Quality Measurement." Agency for Healthcare Research and Quality. https://www.ahrq.gov/patient-safety/quality-resources/tools/chtoolbx/understand/index.html.

Jiménez-Rodríguez, D., Santillán García, A., Montoro Robles, J., Rodríguez Salvador, M., Muñoz Ronda, F. J., and Arrogante, O. 2020. "Increase in Video Consultations During the COVID-19 Pandemic: Healthcare Professionals' Perceptions about Their Implementation and Adequate Management." International Journal of Environmental Research and Public Health 17 (14): 5112.

Kersnik, J. 2001. "Determinants of customer satisfaction with the health care system, with the possibility to choose a personal physician and with a family doctor in a transition country." Health Policy (57:2), pp. 155-164.

Lee, S. V. 2016. "Why Doctors Shouldn't Be Afraid of Online Reviews," Harvard Business Review, https://hbr.org/2016/03/why-doctors-shouldnt-be-afraid-of-online-reviews.

Melumad, S., Inman, J. J., and Pham, M. T. 2019. "Selectively emotional: How smartphone use changes user-generated content." Journal of Marketing Research 56 (2): 259-275.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Parasuraman, A., Zeithaml, V.A., and Berry, L. 1985. "A Conceptual Model of Service Quality and Its Implications for Future Research." Journal of Marketing (49:4), pp. 41-50.

Piccoli, G., and Pigni, F. 2013. "Harvesting External Data: The Potential of Digital Data Streams." MIS Quarterly Executive 12 (1): 53-64.

Pillittere, D., Bigley, M. B., Hibbard, J., and Pawlson, G. 2003. Exploring Consumer Perspectives on Good Physician Care: A Summary of Focus Group Results. The Commonwealth Fund (New York, NY).

Vennik, F. D., Adams, S. A., Faber, M. J., and Putters, K. 2014. "Expert and experiential knowledge in the same place: patients' experiences with online communities connecting patients and health professionals. " Patient Education & Counseling (95:2), pp. 265-270.

Xu, Yuqian, Mor Armony, Anindya Ghose. (2021) The Interplay Between Online Reviews and Physician Demand: An Empirical Investigation. Management Science 67(12):7344-7361.

Yang, H., Guo, X. and Wu, T. 2015. "Exploring the influence of the online physician service delivery process on patient satisfaction." Decision Support Systems 78(C): 113-121.

Yaraghi N, Wang W, Gao G, Agarwal R. 2018. How Online Quality Ratings Influence Patients’ Choice of Medical Providers: Controlled Experimental Survey Study. J Med Internet Res 2018;20(3): e99.

Zhu, F., and Zhang, X. 2013. "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics." Journal of Marketing (74:2), pp. 133-148.