Expert Systems With Applications 71 (2017) 479–492

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback

Tabitha L. James*, Eduardo D. Villacis Calderon, Deborah F. Cook

Department of Business Information Technology, Virginia Tech, 1007 Pamplin Hall, Blacksburg, VA 24061, United States

ARTICLE INFO

Article history:
Received 21 July 2016
Revised 2 November 2016
Accepted 2 November 2016
Available online 9 November 2016

Keywords:
Service quality
Text mining
Healthcare

ABSTRACT

Mechanisms for collecting unstructured feedback (i.e., text comments) from patients of healthcare providers have become commonplace, but analysis techniques to examine such feedback have not been frequently applied in this domain. To fill this gap, we apply a text mining methodology to a large set of textual feedback of physicians by their patients and relate the textual commentary to their numeric ratings. While perceptions of healthcare service quality in the form of numeric ratings are easy to aggregate, freeform textual commentary presents more challenges to extracting useful information. Our methodology explores aggregation of the textual commentary using a topic analysis procedure (i.e., latent Dirichlet allocation) and a sentiment tool (i.e., Diction). We then explore how the extracted topic areas and expressed sentiments relate to the physicians' quantitative ratings of service quality from both patients and other physicians. We analyze 23,537 numeric ratings plus textual feedback provided by patients of 3,712 physicians who have also been recommended by other physicians, and determine process quality satisfaction is an important driver of patient perceived quality, whereas clinical quality better reflects physician perceived quality. Our findings lead us to suggest that to maximize the usefulness of online reviews of physicians, potential patients should parse them for particular quality elements they wish to assess and interpret them within the scope of those quality elements.

1. Introduction

The determination of quality in healthcare presents certain challenges. First, there are two distinct elements of a healthcare experience described in the literature: clinical and process (Marley, Collier, & Meyer Goldstein, 2004). Clinical quality is defined as "the technical quality delivered and result[ing] from medical procedures," whereas process quality is defined as "the result of the service (nontechnical) delivery process engaged during and outside of medical procedures" (Marley et al., 2004 p. 350). In other words, clinical quality specifies the 'what' that includes; for example, the medical procedures, results of tests, diagnoses. The quality of such healthcare elements may be difficult for patients to gauge, because patients are likely to lack highly skilled medical training that may be necessary for the interpretation of these experiences. Process quality, on the other hand, revolves around the delivery of healthcare and it is often easier for patients to judge. For example, the physician's bedside manner, wait time to see the physician, and the kindness and aptitude of the physician's staff are all elements of process quality. Both clinical and process quality experiences may effect the customer's perception of the service, and, perception of service quality in healthcare is very important. Marley et al. (2004) state "U.S. hospitals continue to work to reduce costs while improving quality and efficiency, and have shifted toward a market-driven approach of using patient satisfaction as a measure of organizational performance" (p. 350). It has been suggested that patients in a healthcare setting tend to rely on functional attributes such as facilities, cleanliness, and personnel attitudes rather than technical attributes because they do not typically have technical medical expertise (Babakus & Mangold, 1992; Lanning & O'Connor, 1989; Lee, 2013). However, Marley et al. (2004) found both clinical and process quality to be significant drivers of patient satisfaction using a survey-based methodology. Thus, the healthcare domain as an application area provides a robust set of quality considerations that may benefit from a granular examination of textual feedback.

While collecting quality and customer perception data via survey or comment card has been the traditional approach to the analysis of service quality, online customer feedback mechanisms have become incredibly prevalent. In fact, one researcher found that there are more than 40 websites that allow patients to review their medical care (Lee, 2013). In another study, it was suggested that "physician evaluation websites have tremendous potential to

* Corresponding author. Fax.: (540) 231 3752.
E-mail addresses: tajames@vt.edu (T.L. James), villacis@vt.edu (E.D. Villacis Calderon), dcook@vt.edu (D.F. Cook).

http://dx.doi.org/10.1016/j.eswa.2016.11.004
0957-4174/

help doctors and patients" and that physicians should learn to use such websites to their advantage because they are becoming an unavoidable element of their practices (Merrell, Levy, & Johnson, 2013 p. 1676). However, it has been suggested that online feedback mechanisms present unique consideration and use challenges due to the technological delivery characteristics (Dellarocas, 2003). The importance of exploring the usability of online physician evaluation websites further motivates this study. The relevance, quality, and reliability of the data on such websites is still an open question, and a first step in examining such factors is to be able to explore the characteristics of a large amount of data at an aggregate level from such a website. Specifically, we explore tone (sentiment) and topic of the feedback regarding 3712 physicians, and the relation of these to numeric quality ratings assigned by both patients and other physicians.

Research on the usefulness of online customer reviews is varied, spans across multiple application areas, and has provided somewhat mixed findings (Gao, Greenwood, Agarwal, & McCullough, 2015). Some studies have found that certain forms of online feedback may lead to higher product sales. For example, Chen, Wu, and Yoon (2004) found that consumer ratings did not influence sales, but that the number of reviews did. Similarly, Duan, Gu, and Whinston (2008) found that volume rather than rating had an impact on movie box office revenues. J. Lee, Park, and Han (2008) found that while consumers' conformity to negative reviewer perspectives tends to increase with the proportion of negative feedback, it differs depending on quality for high versus low involvement consumers. Consumer feedback has also been demonstrated to have reputational impacts on sellers in online marketplaces (Pavlou & Dimoka, 2006). Quality of feedback mechanisms have also been suggested to be sensitive to context, such as the case where one seller offers many products of different quality (Dellarocas, 2002). Reviews of physicians have been shown to reflect offline quality ratings, but research has indicated low quality physicians are less likely to be rated online and physicians rated online are given higher marks than offline (Gao et al., 2015).

Research has also suggested that product type, along with review extremity and depth, impacts the perceived helpfulness of reviews (Mudambi & Schuff, 2010). One interesting consideration raised by Mudambi and Schuff (2010) is the distinction between evaluating search and experience goods. Search goods are those for which "it is relatively easy to obtain information on product quality prior to interaction with the product; key attributes are objective and easy to compare, and there is no strong need to use one's senses to evaluate quality" (Mudambi & Schuff, 2010 p. 187). Experience goods are those for which "it is relatively difficult and costly to obtain information on product quality prior to interaction with the product; key attributes are subjective or difficult to compare, and there is a need to use one's senses to evaluate quality" (Mudambi & Schuff, 2010 p. 187). Healthcare is often cited as an experience good, which suggests that the characteristics of a healthcare experience may not be easily determinable prior to consumption.

Healthcare is a complex service and it could be argued that some elements of process quality could be easily related between patients (e.g., timeliness, organization, cleanliness), whereas some may be more difficult (e.g., kindness, concern, bedside manner). Clinical quality should be less opinion-based in theory, but due to the highly technical nature of this element of healthcare, may not be something patients have the expertise to evaluate or are comfortable judging. Therefore, it is of interest to explore what elements are driving consumer perceptions of quality in healthcare to explore just how helpful online reviews may be in evaluating physician quality. In the current study, we examine what topics and tones (sentiments) are prevalent in patient feedback of physicians. Through our topic analysis of unstructured text reviews left by patients, we investigate what aspects of clinical or process quality healthcare consumers' reviews are describing. Then, we explore the tone of the textual feedback using a commercial sentiment package that provides a granular exploration of sentiment using refined dictionaries that provide more depth of tone than simply positive or negative. We then examine which topic and sentiment elements are driving the numeric ratings healthcare consumers assign to physicians. In other words, we seek to determine how the topics described and the sentiments expressed in the textual reviews may affect the numeric ratings assigned. We also explore if and how the topics and sentiments expressed in the patient reviews are related to another measure of perceived quality – physician recommendations of other physicians. This allows us to explore whether the patients' perception of quality is similarly reflected in physicians' evaluation of other physicians.

We seek to extend previous work on online reviews by explicitly examining the characteristics of the textual feedback in the context of service quality in healthcare to provide insight into the following questions.

- Will textual analysis of unstructured patient feedback reveal similar dimensions of service quality?
- Which service quality elements and consumer-expressed sentiments will most influence the numeric patient rating?
- Do the identified textual characteristics drive the patients' and physicians' quality perceptions in similar ways?

In Section 2, we first examine what researchers have explored in regard to online reviews. This discussion is followed by a description of what the literature tells us about dimensions of service quality in healthcare, which helps to guide our topic analysis approach. The framework we follow for our analysis is presented in Section 3 and accompanied by an overview of each analysis technique applied. Section 4 discusses our results and their implications for practice and we end with conclusions in Section 5.

2. Literature review

2.1. Online reviews: evaluating unsolicited and unstructured consumer feedback

Online service reviews enable two primary functions: to assist the decision-making of service consumers and to assist service providers in service quality improvement. For a potential service consumer, online reviews can be used to inform purchasing decisions since they relate other individuals' experience with that service. In other words, they provide others' perceptions of the quality of service obtained. However, Mudambi and Schuff (2010) found differences between the helpfulness of online reviews to other consumers for experience versus search goods. For the service provider, these reviews provide feedback that can help them improve the quality of service and lead to better reviews.

Online forums that are run by either the service provider or an independent party (e.g., Yelp, Angie's list, RateMDs) with the intent of capturing consumers' perceptions of product or service quality are relatively new. Researchers have only begun to examine the wealth of data available through such online review websites. Studies exploring data from online review websites have often bypassed extensive examination of the textual comment data, rather relying on the numeric ratings (e.g., customer numeric rating of the product or service or helpfulness rating assigned to a comment) and calculated variables such as the length of the review (i.e., number of words in review). The current study extends the explorations conducted in much of the literature below by focusing on the topics and tones of the textual feedback left by the patients and exploring their relationships to the numeric ratings. In addition, our study explores consumer reviews of a complex service

– healthcare, where a granular exploration of topic and sentiment over a large dataset can provide insight to complement previous studies exploring the compatibility of online quality perceptions to offline ones (Gao et al., 2015). Previous research provides useful findings to guide our exploration, providing both insight into the usefulness of online feedback and considerations for service quality feedback both offline and online. We will briefly review these findings below.

Mudambi and Schuff (2010) explored Amazon reviews of both search and experience goods and found that review extremity (i.e., numeric ratings at the extreme upper or lower end of the 5-point scale), review depth (i.e., length of review), and product type (i.e., search versus experience) all influenced the consumers' perception of helpfulness of a review (i.e., the helpfulness ranking for reviews indicated on Amazon). In particular, for experience goods, they discovered moderate rated reviews (i.e., closer to 3 on the numeric scale) were more helpful than those at the extremes (i.e., closer to 1 or 5 on the numeric scale). They also found review depth to be positively associated with perceived helpfulness and product type to moderate the relationship between review depth and perceived helpfulness.

Another study considered the textual feedback characteristics in more depth (i.e., basic, stylistic, and semantic) and found that textual feedback that is more extreme in nature received more helpfulness votes (Cao, Duan, & Gan, 2011). Cao et al. (2011) used latent semantic analysis (LSA) to explore the semantic features of reviews and relate the extracted semantic features to the number of helpfulness votes the review received. Abrahams, Fan, Wang, Zhang, and Jiao (2015) investigated vehicle quality using stylistic, social, and sentiment features of online textual feedback. Wallace, Paul, Sarkar, Trikalinos, and Dredze (2014) employed a method that allows topic and sentiment to be jointly explored in a guided manner to investigate correlations between these features and state-level measures of healthcare quality. These three studies are perhaps the closest to the current study in terms of method, because Abrahams et al. (2015) used principal component analysis (PCA), Cao et al. (2011) used LSA, and Wallace et al. (2014) used factorial latent Dirichlet allocation (f-LDA), all of which are topic analysis methods similar to the latent Dirichlet allocation (LDA) procedure employed in the current study. Similar to our study, Abrahams et al. (2015) also explored sentiment, but they measured only negativity, positivity, and subjectivity of the reviews, whereas the method used in this study provides more granularity in tone. Wallace et al. (2014) used f-LDA, which allowed them to explore sentiment and topic concurrently guided by text that was manually annotated by experts to specify topic/sentiment categories. Another key difference between two of these studies and ours is the domain studied. Abrahams et al. (2015) and Cao et al. (2011) examined reviews of products (i.e., automobiles, consumer electronics, and software), whereas our study explores reviews of a service (i.e., healthcare). Furthermore, Abrahams et al. (2015) relate defects to quantitative measures extracted from the textual feedback and Cao et al. (2011) explore the relationship between semantic characteristics and helpfulness of reviews. Wallace et al. (2014)'s study was conducted in the same domain as our study, but explored how guided topic/sentiment categories influenced state-level quality indicators such as return visits to a physician within two weeks of discharge. The current study differs from Wallace et al. (2014) in the following ways: we do not pre-define topic categories, we examine sentiment with more granularity, we adopt physician as the unit of analysis rather than state, and we examine relationships between topic, tone, and quantitative physician ratings from consumers and other physicians.

Other studies of online reviews have discussed the link between online consumer ratings and sales (Chen et al., 2004; Dellarocas, 2003; Duan et al., 2008). For example, Chen et al. (2004) found that product sales and consumer ratings are not related, but that the number of reviews is positively related to sales. Duan et al. (2008) found that the volume of online reviews was significantly related to box office sales. Lee et al. (2008) examined negative online reviews and found that both low and high involvement consumers would conform to negative opinions as the number of negative reviews grew, but that high-involvement consumers were more apt to consider the quality of the reviews. Another study found that the influence of reviews on sales weakens with time (Hu, Liu, & Zhang, 2008). Furthermore, Hu et al. (2008) found that the market responded to reviewers that had high reputations and exposure, and that consumers were considerate of these reviewer characteristics as well as review scores.

Other studies have examined online reviews, but with more of a focus on how they influence buyers, sellers, or product perceptions. Pavlou and Dimoka (2006) used content analysis of text reviews to explore trust, price premiums, and seller differentiation on eBay. Pavlou and Dimoka (2006)'s study used textual feedback, but employed traditional coding methods to classify comments into categories (specifically, benevolence and credibility categories). Vermeulen and Seegers (2009) examined review valence and found that positive reviews improved the consumers attitudes towards hotels. Vermeulen and Seegers (2009)'s study was conducted as an experiment to explore the influences of positive and negative reviews rather than using a text mining approach. Misopoulos, Mitic, Kapoulas, and Karapiperis (2014) examined service quality perceptions of airlines applying sentiment analysis tools on Twitter comments to identify areas of customer satisfaction, dissatisfaction, and delight. Sentiment analysis has also been used as a method to explore hotel service quality, where positive and negative service quality dimensions were found to influence the number of reviews (Duan, Cao, Yu, & Levy, 2013), and in healthcare where sentiment was used to explore dissatisfaction (Alemi, Torii, Clementz, & Aron, 2012).

Online reviews have been explored in the healthcare domain to examine two different quality outcomes. As previously mentioned, Wallace et al. (2014) explored the correlations of quality indicators in healthcare and state-level indicators of quality (e.g., likelihood of patient visiting physician within two weeks of discharge). Another study in the healthcare domain examining the star ratings from a physician review website against another measure of patients' evaluation of physician quality found that "online reviews [of physicians] reflect offline population opinion suggest[ing] that these reviews are informative" (Gao et al., 2015 p. 584). Gao et al. (2015) found that lower quality physicians (i.e., those rated lower through a traditional quality collection survey) were less likely to be rated online, online star ratings were more informative for average quality physicians, and that the highest quality physicians (i.e., those rated higher through a traditional quality collection survey) were harder to distinguish amongst using online star ratings. These findings are interesting in that they suggest the star ratings have been determined to be informative and in line with traditional measures of service quality, but may not be sufficiently granular to help evaluate service quality based solely on the star rating without considering the textual content.

Gao et al. (2015)'s study provides motivation for the current study by illustrating that online reviews of physicians may be informative, but that the star ratings may not make differences in quality easily apparent. The next step is to examine the textual content of the reviews in more detail to see if topic and tone can be extracted and aggregated in way that reveals quality indicators and to determine if these indicators relate to numeric quality ratings. Such relationships may indicate which topics and sentiments are driving the numeric ratings, and furthermore, provide some insight into the correlation between the numeric ratings and the textual feedback that would suggest whether the nu-

meric ratings provided similar informational value to the textual feedback.

2.2. Exploring dimensions of service quality in healthcare

The current study focuses on service quality in healthcare and employs automated methods to examine large amounts of unstructured textual feedback written by patients on a website dedicated to providing quality information about physicians. Quality for both products and services has long been a topic of interest, and thus, has been studied over the years using a variety of methodologies. In what follows, we briefly review some of the major methods to explore quality in healthcare and what these analyses revealed in terms of the dimensions of quality in healthcare. These findings guide our own analysis by indicating the number of overarching dimensions of quality in healthcare services typically identified in past research.

Perhaps one of the most influential instruments to measure service quality, SERVQUAL was originally suggested by Parasuraman, Zeithaml, and Berry (1985), Parasuraman, Zeithaml, and Berry, 1988. SERVQUAL is a survey instrument that was developed to measure perceived quality by collecting pre-service (expectations) and post-service (perceptions) data from consumers using survey methods. Ananthanarayanan Parasuraman et al. (1988) p. 143) define the difference between expectations and perceptions thusly "service quality, as perceived by consumers, stems from a comparison of what they feel service firms should offer (i.e., from their expectations) with their perceptions of the performance of firms providing the services". This is an important distinction in the development of SERVQUAL, because the instruments are used to measure and examine the 'gaps' between expectations and perceptions. SERVQUAL was later refined and replicated in Parasuraman, Berry, and Zeithaml (1991a), as well as tested across multiple service companies in Parasuraman, Berry, and Zeithaml (1991b).

The scale development process for SERVQUAL revealed five dimensions of perceived quality: tangibles, reliability, responsiveness, assurance, and empathy (Ananthanarayanan Parasuraman et al., 1988). The 'tangibles' dimension includes aspects of the service quality experience that have to do with the physical environment (e.g., equipment, facilities, personnel). The 'reliability' and 'assurance' dimensions reflect the consumers' perceptions of the capacities of the service provider to perform the service. The former includes the assessment of the service provider's abilities that relate dependability and accuracy and the latter with qualities such as knowledge and courtesy that may inspire trust and confidence in the service provider. The 'responsiveness' dimension focuses on the perceived helpfulness and promptness of the service provider. Finally, the 'empathy' dimension relates the perception of personalized, compassionate service. SERVQUAL has been suggested as part of a proposed framework to examine information technology (IT) service quality using unstructured text (Sperkova, Vencovsky, & Bruckner, 2015) and as a framework to explore hotel service quality using online reviews (Duan et al., 2013), as well as encapsulated in other decision support methods (e.g., data envelopment analysis) (Lee & Kim, 2014).

SERVQUAL models have been used to measure service quality in hospital and healthcare settings, primarily using survey-based methods. Butt and de Run (2010) developed and tested a SERVQUAL model for measuring private healthcare service quality in Malaysia. Wisniewski and Wisniewski (2005) adapted the SERVQUAL instrument and used it to measure service quality in a hospital colposcopy clinic. Rohini and Mahadevappa (2006) used SERVQUAL to study the perception of service quality in five Bangalore based hospitals and reported that it appears to be a reliable and consistent instrument to measure healthcare service quality. Ramsaran-Fowdar (2005) also described health care quality attributes using the five SERVQUAL dimensions as the basis. Other researchers that used SERVQUAL to evaluate healthcare related service quality include Anderson (1995), Babakus and Mangold (1992), Bowers, Swan, and Koehler (1994), Ramez (2012) and Youssef, Nel, and Bovaird (1995).

Some researchers have developed their own frameworks to conceptualize service quality in healthcare. Donabedian (1980) identified two dimensions of healthcare service quality: technical and functional. Technical quality relates to the technical accuracy of the medical diagnosis and procedures, where functional quality relates to the way that the care is delivered to the patient. This is distinction is similar in nature to the process and clinical quality categories suggested by Marley et al. (2004) and described above. Vandamme and Leunis (1993) identified tangibles, medical responsiveness, assurance, nursing staff quality, and personal values as the dimensions of hospital service quality. Padma at al. (2010) used dimensions that were specifically developed for healthcare services (infrastructure, personnel quality, process of clinical care, administrative procedures, safety indicators, hospital image, social responsibility, and trustworthiness of the hospital) to analyze the relationship between service quality and customer satisfaction in hospitals in India. One widely used instrument to measure patient satisfaction in healthcare (Venkatesh, Zhang, & Sykes, 2011) suggests it has six dimensions: technical quality, interpersonal aspects, communication, financial aspects, time spent with doctor, and access/availability/convenience (Hays, Davies, & Ware, 1987; Ware, Snyder, & Wright, 1976a, 1976b; Ware, Snyder, Wright, & Davies, 1983).

A qualitative study using online reviews of physicians conducted by López, Detz, Ratanawongsa, and Sarkar (2012) identified three specific factors in healthcare service: interpersonal manner, technical competence, and system issues. Technical competence is a view of clinical quality and López et al. (2012) indicated that it included commentary discussing the physicians' knowledgeableness, attention to detail, efficiency, along with descriptions of clinical skills, follow-ups, perceptions of decisions making, perceptions of successfulness of treatment, etc. Interpersonal manner is related to process quality, but of the physician rather than his or her staff. López et al. (2012) found that it included descriptions of the physicians' empathy, friendliness, helpfulness, trustworthiness, along with descriptions of their ability to put someone at ease, listen, explain, etc. System quality is related to process quality staff and management issues and López et al. (2012) found it to include descriptions of staff, appointment, wait times, environment, location, cost, etc. Wallace et al. (2014) used these three topics to guide their f-LDA models, but divided each topic into a negative and a positive sentiment resulting in six overall categories of quality: positive technical, negative technical, positive interpersonal, negative interpersonal, positive system, and negative system. Manually labeled data relating these six categories were used to incorporate prior information into the f-LDA approach of Wallace et al. (2014).

This review illustrates that quality in services, and particularly in healthcare, is normally found to include two to eight factors. The models most commonly used to explore quality in healthcare have five or six factors. Thus, we incorporate this information into our analysis and explore six factors of healthcare service quality, but unlike Wallace et al. (2014) we do not pre-specify the topic areas. In addition, we complement the LDA topic analysis with an exploration of sentiment in a richer form than simply positive or negative tone using a commercial sentiment analysis software package. In the current study, we examine unstructured textual feedback from an online physician review website. We suggest that all the comments for a physician taken together provide a 'view' of the service quality of that physician, and that by examining more than 20,000 such reviews for more than 3700 physicians we can determine if similar service quality or patient satisfaction dimensions to

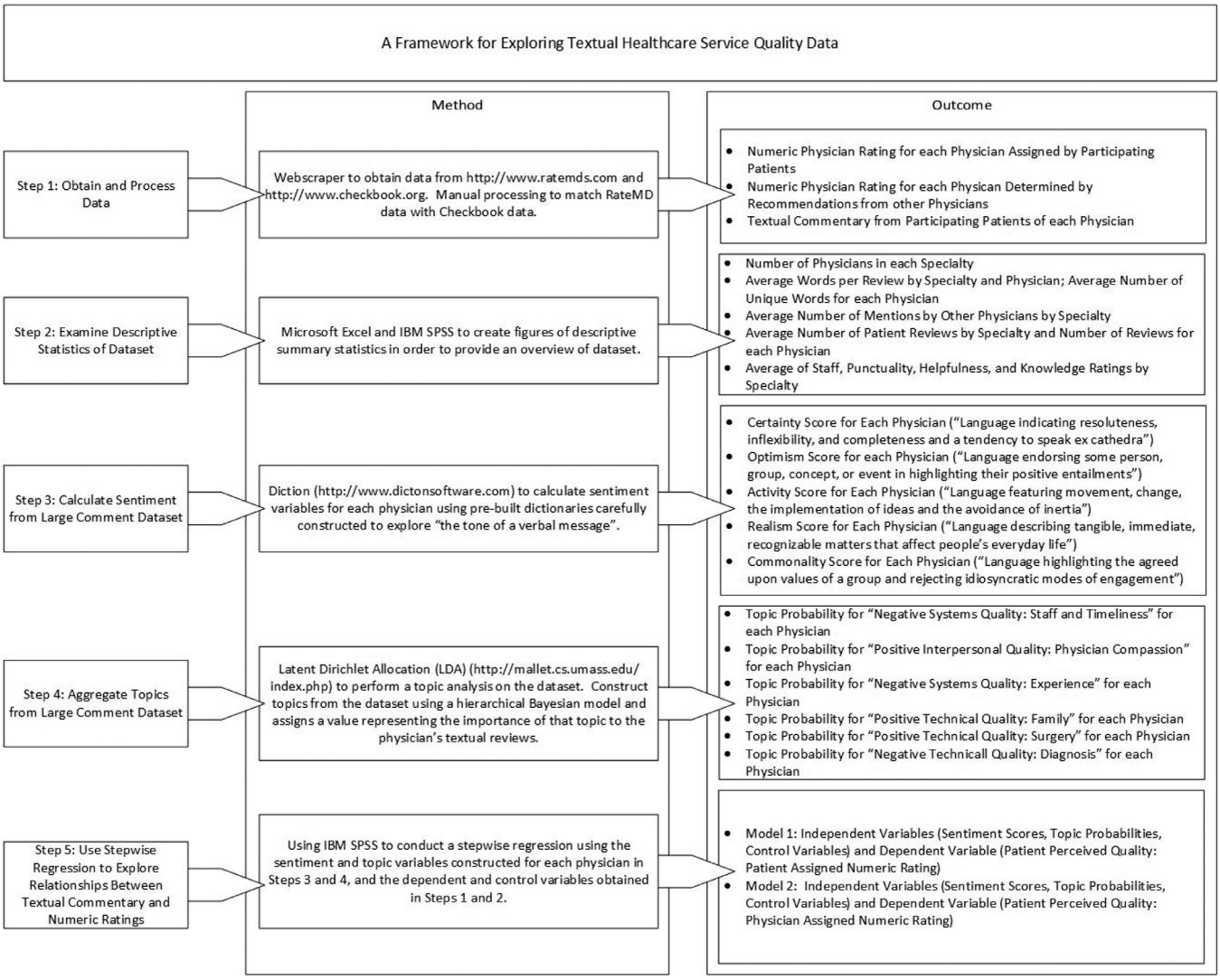

Fig. 1. Framework for exploring textual healthcare service quality data.

the ones described above arise from the unsolicited, unstructured feedback. This approach also allows us to explore the impact of the text-mined dimensions and sentiment on the overall numeric star (quality) rating of the physician, which is a useful step in determining the value of such comments to healthcare consumers or providers.

3. Methodology

To apply our text mining procedure to data collected from a service quality website for physicians, we followed the framework illustrated in Fig. 1. The process consists of five primary steps: obtaining and processing the data, examining the descriptive statistics, calculating sentiment variables for each physician, calculating topic probabilities for each physician, and exploring the relationships between the text-based variables and the numeric ratings of the physicians. The method applied for each step and the outcomes resulting from the application are detailed in Fig. 1.

3.1. Step 1: data collection

The patient review data was scraped from the website RateMDs (http://www.ratemds.com). The website allows an individual to select a star rating for a physician on four areas: staff, punctuality, helpfulness, and knowledge. The individual is also able to write an unstructured text review of his or her perception of or experience with the physician. The physician's name, location (city, state/province, and country), and specialty are also listed. The web scraper we used was coded in Python by one of the authors and was used to download all of these fields, as well as all of the reviews, for each of the physicians.

We pulled data for 106,852 physicians with a variety of specialties from a range of countries. To constrain the sample to English language comments and from the same country's medical system, we focused on physicians from the United States (U.S.). The scrape contained 75,336 U.S. physicians. In order to have a measure of physician quality that was not linked to patient perceptions, we purchased access to data from Consumers' Checkbook (http://www.checkbook.org) that provides data on U.S. physicians from major metropolitan areas. The data includes the names, specialties, geographic location, and a variable called the 'number of mentions' for each physician, which we will explain in the following section. We could locate data for 25,922 physicians on the Consumers' Checkbook website, but this data needed to be matched with the data from RateMDs. We automated the matching of the physicians from the two datasets by comparing their last names and geographic locations. Then, we cleaned the remaining dataset

All Physicians (3,712)

# Reviews: 23,537					
Avg. Words per Review: 65.1					
AM	AR	AS	AP	AH	AK
7.1	6.3	3.2	3.7	3.9	4.1

Legend
AM: Average Number of Mentions by Other Doctors
AR: Average Number of Reviews by Patients
AS: Average of Staff Ratings by Patients
AP: Average of Punctuality Ratings by Patients
AH: Average of Helpfulness Ratings by Patients
AK: Average of Knowledge Ratings by Patients

By Specialty: # Physicians in Speciality (Avg. Words per Review)

	AM	AR	AS	AP	AH	AK
Addiction Medicine: 2 (54.3)	3	2	4.5	4.8	4.3	4.8
Allergist/Immunologist: 83 (58.7)	11	4.8	3.3	3.9	4	4.2
Anesthesiologist: 15 (70.1)	5.3	5.5	3.1	3.3	3.3	3.4
Cardiologist: 188 (47.3)	6.4	3.1	3.7	4.3	4.5	4.6
Cardiothoracic Surgeon: 25 (53.8)	15	4.8	3	4.1	4.2	4.5
Colorectal/Proctologist: 1 (76)	13	7	4.4	3.6	4.3	4.4
Cosmetic/Plastic Surgeon: 102 (69.9)	7.6	8.5	3.7	4	3.9	4
Dermatologist: 132 (66.9)	7.1	7.9	2.9	3.3	3.5	3.8
Ear Nose and Throat (ENT): 135 (65.8)	7.3	6.9	3.2	3.8	3.9	4.1
Emergency/Critical: 13 (40.5)	5.9	1.7	3	3.8	3.8	4.1
Endocrinologist: 159 (74)	10.3	9.4	2.6	3.3	3.3	3.7
Family/G.P.: 176 (55.2)	4.1	6.3	3	3.8	4	4.2
Gastroenterologist: 183 (61.6)	6.7	5.9	3.4	3.8	3.9	4.2
Gynecologist (OBGYN): 327 (71.4)	4.7	10.2	3.2	3.7	4.1	4.3
Infectious Disease: 39 (54.7)	13.2	3.7	3.1	3.4	3.2	3.4
Internist/Geriatrician: 305 (54.1)	4.6	5.3	3.2	4	4.2	4.4
Nephrologist (kidney): 58 (39.4)	9.6	3.2	3.5	4	4.4	4.5
Neurologist: 173 (68.6)	7.8	6.1	3.1	3.6	3.5	3.8
Neurosurgeon: 123 (71.5)	13	8	3.4	3.8	4	4.1
Oncologist/Hematologist: 181 (58.3)	9.8	4.5	3.5	3.9	4.2	4.4
Ophthalmologist: 110 (57)	6.1	4	3.2	3.6	3.8	4.1
Optometrist: 2 (49.9)	3.3	5	2.8	2.9	3.1	3.2
Orthopedics/Sports: 257 (65.2)	5.6	6.3	3.5	3.7	3.9	4.2
Pain Mgmt/Physical Rehab: 64 (79.9)	5.6	7.9	3.1	3	3.1	3.4
Pathologist: 8 (61.6)	10.4	2.5	2.5	3.4	3.3	3.5
Pediatrician: 152 (63.5)	4.6	5.3	3	3.9	4.2	4.3
Preventive/Occupational: 1 (61.8)	5	6	4.7	4	3.7	4.3
Psychiatrist: 49 (67.1)	4.2	6.9	2.5	3.2	3.1	3.3
Pulmonologist (lungs): 78 (56.1)	8	3.3	3.2	3.8	3.9	4.1
Radiologist: 20 (56.5)	3.9	1.9	3.9	4.2	4.1	4.2
Reproductive Endocrinologist: 57 (86.6)	8.8	15.6	3.4	3.8	3.7	4
Rheumatologist: 111 (70.4)	10.1	7.7	3	3.4	3.7	3.9
Sleep Disorders: 2 (89.4)	5	3.5	3.9	3.1	2.9	3.6
Surgeon (general): 254 (56.6)	7.7	4.5	3.6	4	4	4.2
Urologist: 121 (57.4)	7.9	5.7	3.2	3.7	3.8	4.1
Vascular/Phlebologist: 6 (35.7)	9	2.3	3.2	3.6	3.6	3.9

Fig. 2. Descriptive statistics for physician data set.

by hand to remove entries where the first name and specialty did not match. The final dataset contains 3712 physicians.

3.2. Step 2: sample description

Fig. 2 provides the number of physicians in each specialty. For physicians within each specialty, the following statistics are presented: the average words per review, the average number of mentions by other physicians, the average number of patient reviews, and the average numeric rating for each of the four areas requested on RateMDs. A trend can be seen in the descriptive statistics. Of the four numeric ratings (staff, punctuality, helpfulness, and knowledge), the ratings for staff and punctuality are lower than those for helpfulness and knowledge. This indicates that in general the complaints the patients have about the physicians in our dataset are more likely due to process quality issues.

3.3. Step 2: overview of variables

In our analysis, we explored two dependent variables and three categories of independent variables: sentiment variables, topic variables, and controls. The three categories of independent variables provide us with a robust representation of the 23,537 freeform text comments left by patients regarding the healthcare provided by 3712 physicians in our dataset. An overview of our variables is given in Fig. 3.

We examine two dependent variables that give us two different measures of perceived quality: 'patient perceived quality' and 'physician perceived quality'. Patient perceived quality is obtained from the numeric star rating assigned to the physicians by the patients. When the patient rates a physician on RateMDs, they are asked to numerically rate four categories (staff, punctuality, helpfulness, and knowledge). We averaged these ratings over all patient reviews for each physician and then averaged the four categories to obtain an 'overall perceived quality', which corresponds to the overall star rating displayed on RateMDs for each physician. Thus, this is the overall star rating the information consumer would first see for a physician on the website, and the primary rating information the consumer would see if they did not click on the physician's name.

For a measure of physician perceived quality, we used the 'number of mentions' variable collected from Consumers' Checkbook. Consumers' Checkbook asked physicians in 53 major metropolitan areas to recommend other physicians in their community. The physicians were asked to recommend one or two other physicians from each specialty to whom they would send their loved ones for treatment. Number of mentions is the total number of times a physician was recommended by their peers. The number of mentions variable provides a measure of perceived

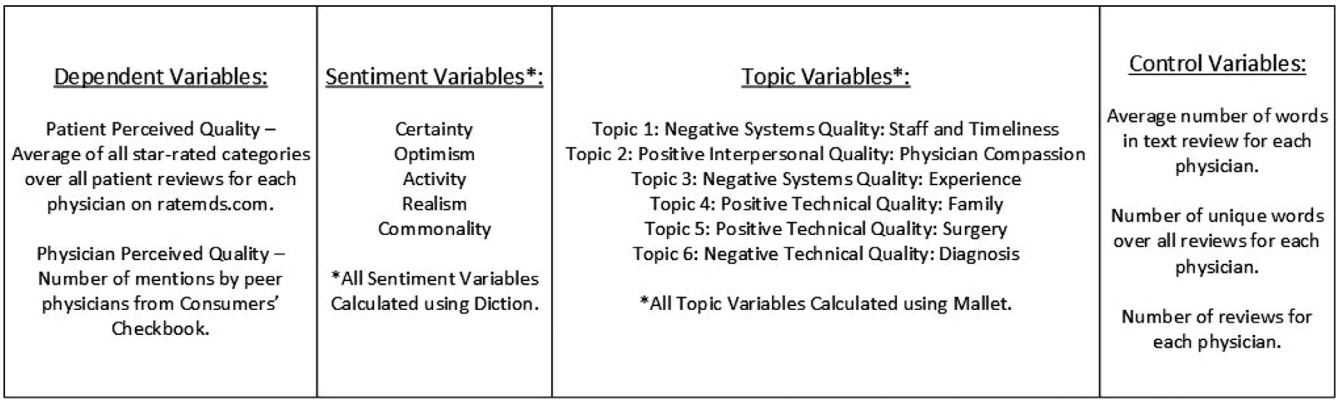

Fig. 3. Dependent and independent variables for analysis.

Table 1
Descriptive statistics.

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Patient Perceived Quality	3712	4.250	0.750	5.000	3.865	0.836
Physician Perceived Quality	3712	56.000	3.000	59.000	6.860	4.848
Certainty	3712	63.210	11.930	75.140	43.264	5.556
Optimism	3712	268.580	−72.290	196.290	54.045	10.582
Activity	3712	1072.600	−972.840	99.760	42.316	52.332
Realism	3712	108.530	−26.390	82.140	46.636	5.501
Commonality	3712	152.970	−66.600	86.370	49.129	3.274
Topic1	3712	1.000	0.000	1.000	0.162	0.125
Topic2	3712	1.000	0.000	1.000	0.212	0.160
Topic3	3712	1.000	0.000	1.000	0.149	0.104
Topic4	3712	1.000	0.000	1.000	0.152	0.135
Topic5	3712	1.000	0.000	1.000	0.151	0.155
Topic6	3712	1.000	0.000	1.000	0.156	0.116
Average Number of Words	3712	200.000	1.000	201.000	55.184	29.948
Number of Unique Words	3712	2502.000	0.000	2502.000	221.050	261.629
Number of Reviews	3712	47.000	1.000	48.000	6.370	6.163

quality from other physicians (i.e., peers) who are more likely to have the technical expertise to evaluate clinical quality better than most patients.

In the sections that follow, we will describe how the sentiment and topic variables in Fig. 3 were obtained. Table 1 provides descriptive statistics (i.e., range, minimum, maximum, mean, and standard deviation) for each one of the raw variables in our dataset described in Fig. 3. The spreads were large for the dependent variables, the sentiment variables, and the controls. The topic variables are calculated as percentages by the LDA procedure, and therefore each physician's comments are assigned a value between zero and one on each topic. To equal the spreads, we applied a transformation to the sentiment variables, control variables, and dependent variables (i.e., the natural log). We also calculated the Pearson correlation coefficient between each of the variables in our dataset in their final form, shown in Table 2, to detect any collinearity issues. We also report variance inflation factors (VIFs) with our regression models in a later section. The correlations and VIFs suggest collinearity is not an issue (Larose & Larose, 2015).

3.4. Step 3: sentiment calculation

To calculate sentiment variables for the text reviews for each physician, we used a software program called DICTION. DICTION is text analysis software used to determine "the tone of a verbal message" (http://www.dictionsoftware.com). The software has dictionaries of words defined for many themes and uses the count of the words from these themes to calculate overarching scores for five qualities, shown in Fig. 4. For example, 'optimism' is defined by DICTION as "language endorsing some person, group, concept, or event or highlighting their positive entailments" . It is calculated by comparing the input text against six dictionaries of words associated with the themes of 'praise', 'satisfaction', 'inspiration', 'blame', 'hardship', and 'denial'. For each of these themes, DICTION searches the input text for the terms in the associated dictionary. For example, 'praise' is defined as "affirmations of some person, group, or abstract entity" and includes terms such as 'delightful', 'shrewd', 'reasonable', 'successful', 'conscientious', and 'good'. DICTION counts the words from the dictionary that appear in the input text. It then uses the counts from these various dictionaries and the formulas shown in Fig. 4 to calculate values for each of the overarching themes. Optimism, for example, is calculated by adding together the word counts for 'praise', 'satisfaction', and inspiration' and subtracting from that sum, the sum of the word counts for 'blame', 'hardship', and 'denial'. DICTION's formulas adjust for errors, such as homographs. The result is a score for each document. DICTION has been used recently in the management and marketing literature (Allison, McKenny, & Short, 2013; Short & Palmer, 2008; Short, Broberg, Cogliser, & Brigham, 2010; Zachary, McKenny, Short, Davis, & Wu, 2011).

For our analysis, we created a corpus of documents containing one file for each physician. Each physician's file contained the textual feedback for his or her reviews on RateMDs. We then used DICTION to calculate a score for each of the sentiment qualities in Fig. 4 for each physician. These scores are what we used for our five sentiment variables listed in Fig. 3.

3.5. Step 4: topic analysis

The sentiment scores convey the tone of the textual feedback, but do not explicitly convey the topics being discussed. Therefore, we employed a second tool to examine the textual feedback for

Table 2
Correlations.

	ln Optimism	ln Certainty	ln Realism	ln Commonality	ln Activity	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	ln Unique Words	ln Avg Of Number ofWords	ln Number of Reviews	ln No of Mentions	ln Avg Overall
ln Optimism	1.000	−0.055	−0.155	0.014	−0.520	−0.124	0.445	−0.377	0.074	−0.047	−0.153	−0.282	−0.314	−0.196	−0.014	0.366
ln Certainty	−0.055	1.000	0.163	0.017	0.093	0.064	0.016	−0.052	0.065	−0.062	−0.050	0.094	−0.028	0.264	−0.009	0.097
ln Realism	−0.155	0.163	1.000	0.046	0.232	0.116	−0.010	0.043	0.013	−0.131	−0.003	0.189	0.231	0.121	−0.014	−0.046
ln Commonality	0.014	0.017	0.046	1.000	−0.002	0.005	0.033	−0.004	0.014	−0.063	0.022	−0.028	−0.030	−0.020	−0.057	0.005
ln Activity	−0.520	0.093	0.232	−0.002	1.000	−0.049	−0.083	0.068	0.052	0.009	0.023	0.180	0.228	0.107	0.003	−0.041
Topic 1	−0.124	0.064	0.116	0.005	−0.049	1.000	−0.296	0.060	−0.217	−0.278	−0.091	0.064	0.048	0.064	−0.080	−0.281
Topic 2	0.445	0.016	−0.010	0.033	−0.083	−0.296	1.000	−0.474	−0.106	−0.250	−0.164	−0.325	−0.323	−0.244	−0.043	0.385
Topic 3	−0.377	−0.052	0.043	−0.004	0.068	0.060	−0.474	1.000	−0.182	−0.109	0.055	0.233	0.230	0.175	0.051	−0.438
Topic 4	0.074	0.065	0.013	0.014	0.052	−0.217	−0.106	−0.182	1.000	−0.251	−0.291	0.011	0.052	0.047	−0.117	0.174
Topic 5	−0.047	−0.062	−0.131	−0.063	0.009	−0.278	−0.250	−0.109	−0.251	1.000	−0.289	0.033	0.052	0.005	0.138	0.077
Topic 6	−0.153	−0.050	−0.003	0.022	0.023	−0.091	−0.164	0.055	−0.291	−0.289	1.000	0.069	0.107	0.010	0.054	−0.129
ln Unique Words	−0.282	0.094	0.189	−0.028	0.180	0.064	−0.325	0.233	0.011	0.033	0.069	1.000	0.858	0.856	0.085	−0.170
ln Avg Of Number of Words	−0.314	−0.028	0.231	−0.030	0.228	0.048	−0.323	0.230	−0.026	0.052	0.107	0.858	1.000	0.481	0.069	−0.178
ln Number of Reviews	−0.196	0.264	0.121	−0.020	0.107	0.064	−0.244	0.175	0.047	0.005	0.010	0.856	0.481	1.000	0.078	−0.112
ln No of Mentions	−0.014	−0.009	−0.014	−0.057	0.003	−0.080	−0.043	0.051	−0.117	0.138	0.054	0.085	0.069	0.078	1.000	0.014
ln Avg Overall	0.366	0.097	−0.046	0.005	−0.041	−0.281	0.385	−0.438	0.174	0.077	−0.129	−0.170	−0.178	−0.112	0.014	1.000

topics being discussed by the patients. There are several natural language processing tools for performing topic or content analysis over a corpus of text documents. Such tools include latent semantic indexing/analysis (LSI/LSA) (Landauer, Foltz, & Laham, 1998), principle components analysis (PCA), and latent Dirichlet allocation (LDA). LDA stems from work on LSI and probabilistic LSI (Blei & Lafferty, 2009). These techniques work on word occurrence matrices, sometimes called bag-of-words. LDA (Blei & Lafferty, 2009; Blei, Ng, & Jordan, 2003) "is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topic probabilities" (Blei et al., 2003 p. 993). Like PCA, LDA can be used for dimension reduction. That is, a set of topics can be defined and a value assigned to each document (in our case, each physician's reviews) representing the importance of that topic to the document. In other words, "in the context of text modeling, the topic probabilities provide an explicit representation of a document" (Blei et al., 2003 p. 993). This allows us to use the topics probabilities (i.e., feature representations) as variables in our regression models. A similar approach was employed using PCA in Abrahams et al. (2015) to reduce a large vocabulary of words into manageable semantic elements or features. Similarly, Cao et al. (2011) used LSA, and Wallace et al. (2014) used factorial latent Dirichlet allocation (f-LDA). The f-LDA approach of Wallace et al. (2014) was guided, that is, not only did they specify the number of topics considered (i.e., six) as we do, they also used manually annotated data. Our approach allows for a set number of topics to be derived from the text without the intensive preprocessing to annotate sample text. In our discussion, we contrast our findings to those from Wallace et al. (2014).

We employed the MALLET (MAchine Learning for LanguagE Toolkit) implementation of LDA (http://mallet.cs.umass.edu/index.php) to extract our topics from a corpus of patient reviews (one document for each physician, containing all of his or her textual feedback). MALLET automatically removes stop words (i.e., words that do not add meaning to the text such as prepositions and conjunctions, e.g., a, to, by, and, but, in) using its own built in stop word list prior to executing the LDA algorithm on the corpus.

Our literature review suggested that there are typically between two general and eight more specific quality or satisfaction categories for healthcare services, most commonly five or six. Notably, the methods applied to data most similar to ours used three and six categories. Thus, we used LDA to determine six primary topics from our data. An example of how MALLET works is provided in Fig. 5. In Fig. 5, we show only a couple of reviews for a physician and have removed names and identifiers. This particular physician had comments that related most highly to Topic 1 (51%) from the LDA model. Topic 1 can be seen in Fig. 5 to relate to negative systems quality (i.e., poor performance on process quality indicators such as staff and timeliness). One physician's comments do not have to be assigned exclusively to one topic, and as can be seen in Fig. 5, Topic 6 was modeled as 16% important to this physician's comments. Topic 6 in our analysis is described as negative technical quality (i.e., poor performance on technical quality elements such as diagnosis). Thus, once the topic model is determined by LDA, each physician is assigned a percentage for each topic from zero to one based on the textual content of his or her comments from the online review website.

3.6. Step 5: regression models

To test the influence of the sentiment, topic, and control variables on both patient and physician perceived quality (see Fig. 3), we ran two stepwise regressions in IBM SPSS 23. The resulting models will be presented and discussed in the next section.

Certainty	Optimism
[Tenacity + Leveling + Collectives + Insistence] – [Numerical Terms + Ambivalence + Self Reference + Variety]	[Praise + Satisfaction + Inspiration] – [Blame + Hardship + Denial]
Language indicating resoluteness, inflexibility, and completeness and a tendency to speak ex cathedra.	Language endorsing some person, group, concept, or event or highlighting their positive entailments.
Activity	**Realism**
[Aggression + Accomplishment + Communication + Motion] – [Cognitive Terms + Passivity + Embellishment]	[Familiarity + Spatial Awareness + Temporal Awareness + Present Concern + Human Interest + Concreteness] – [Past Concern + Complexity]
Language featuring movement, change, the implementation of ideas and the avoidance of inertia.	Language describing tangible, immediate, recognizable matters that affect people's everyday life.
Commonality	
[Centrality + Cooperation + Rapport] – [Diversity + Exclusion + Liberation]	
Language highlighting the agreed upon values of a group and rejecting idiosyncratic modes of engagement.	

Fig. 4. Definitions and formulas for diction 7.1 sentiment qualities (Formulas and definitions obtained from diction 7.1 help file).

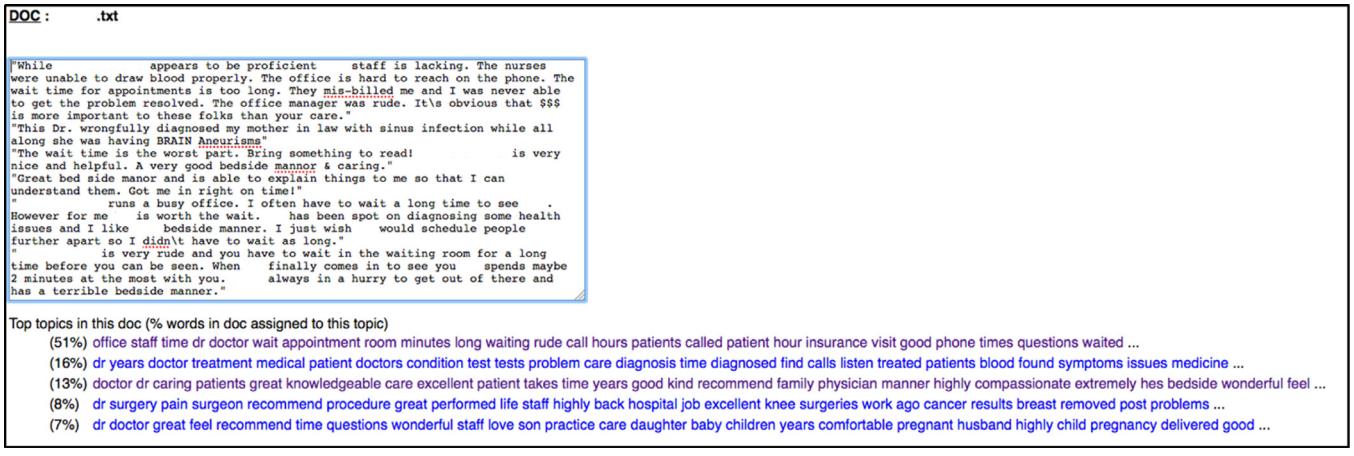

Fig. 5. Truncated example of LDA topic analysis for one physician.

4. Results and discussion

The intent of our analysis was to explore the following three questions: 1) would the topics extracted from the textual feedback reveal similar quality dimensions to those commonly described in the service/healthcare quality literature, 2) could we employ granular expressions of sentiment and quality topic features extracted from the textual feedback to explore the influence of tone and topic on a numeric patient perceived quality measure (i.e., which sentiment and topic elements were driving the numeric quality measure assigned?), and 3) did similar relationships between such textual elements and quality ratings hold for both patient and physician perceived quality?

In regard to the first question, the topics extracted using LDA indeed reveal similar quality dimensions to those described in the literature on healthcare service quality. This finding suggests that it is possible to automate the topic analysis portion of our framework to reveal similar healthcare service quality dimensions without the need to manually annotate text. The six topics resulting from the application of MALLET's LDA to the physician textual feedback are shown in Fig. 6. For each topic, the lists of words that make up that topic are provided below the topic name. Algorithms such as LDA perform a topic analysis by mathematically extracting prevalent words, but the interpretation is left to the researchers. Therefore, we assigned the names to each topic based upon examination of the word lists in consideration of the dimensions of healthcare service quality described in our literature review.

We found that our topics most closely resembled the categorization of López et al. (2012) (i.e., systems, interpersonal, and technical), but using the polarity adopted for these three categories by Wallace et al. (2014) (i.e., positive systems, negative systems, positive interpersonal, negative interpersonal, positive technical, and negative technical). There were some notable differences in our topics compared to the forced topics of Wallace et al. (2014) that provide some interesting insight.

System issues are process quality staff and management issues, which López et al. (2012) described as focusing on staff, appointment, wait times, environment, location, cost, etc. Topic 1, which

Topic 1 Negative Systems Quality: Staff and Timeliness	Topic 2 Positive Interpersonal Quality: Physician Compassion	Topic 3 Negative Systems Quality: Experience	Topic 4 Positive Technical Quality: Family	Topic 5 Positive Technical Quality: Surgery	Topic 6 Negative Technical Quality: Diagnosis
Office	Doctor	Told	Dr	Dr	Dr
Staff	Dr	Back	Doctor	Surgery	Years
Time	Caring	Didn't	Great	Pain	Doctor
Dr	Patients	Don't	Feel	Surgeon	Treatment
Doctor	Great	Doctor	Recommend	Recommend	Medical
Wait	Knowledgeable	Asked	Time	Procedure	Patient
Appointment	Care	People	Questions	Great	Doctors
Room	Excellent	Months	Wonderful	Performed	Condition
Minutes	Patient	Im	Staff	Life	Test
Long	Takes	Found	Love	Staff	Tests
Waiting	Time	Bad	Son	Highly	Problem
Rude	Years	Amp	Practice	Back	Care
Call	Good	Money	Care	Hospital	Diagnosis
Hours	Kind	Good	Daughter	Job	Time
Patients	Recommend	Wanted	Baby	Excellent	Diagnosed
Called	Family	Gave	Children	Knee	Find
Patients	Physician	Make	Years	Surgeries	Calls
Hour	Manner	Called	Comfortable	Work	Listen
Insurance	Highly	Call	Pregnant	Ago	Treated
Visit	Compassionate	Work	Husband	Cancer	Patients
Good	Extremely	Wrong	Highly	Results	Blood
Phone	Hes	Rude	Child	Breast	Found
Times	Bedside	Care	Pregnancy	Removed	Symptoms
Questions	Wonderful	Experience	Delivered	Post	Issues
Waited	Feel	Made	Good	Problems	Medicine

Fig. 6. Topics extracted through the application of LDA to physician feedback data.

we call negative systems quality: staff and timeliness, is similar to the negative systems category from Wallace et al. (2014) in that it relates the negative process quality elements referencing staff and management issues such as office staff, appointments, calling, visiting, waiting and time elements such as hours and minutes, insurance, etc., but suggests problems with this element of the service such as rude, waited, long. There are three adjectives in this topic that suggest connotation: rude, long, and good. Given the words rude and long, combined with multiple of forms of the word wait (wait, waiting, waited), we label this topic as negative in connotation. However, while there are weak indications of a possible negative undertone, Topic 1 may not have a strong connotation. Interestingly, Topic 3 (negative systems quality: experience) also relates systems process quality concerns. Topic 3 contains negative words, including: wrong, rude, bad, didn't, and don't, along with one positive word, good. However, there are few words to clearly indicate one overarching topic, but there are several that indicate this is also a systems process quality topic (months, money, called, call, experience). Finding two negative systems quality topics in our dataset suggests that systems quality concerns (e.g., long wait times, communication issues, payment and insurance issues, staff problems) are prevalent in our dataset.

López et al. (2012) described interpersonal manner as relating the physicians' empathy, friendliness, helpfulness, trustworthiness, along with descriptions of their ability to put someone at ease, listen, explain, etc. Therefore, this dimension has more to do with the physician's behavior than the functioning of his or her office. The words for Topic 2 (positive interpersonal quality: physician compassion) revolve around the manner or demeanor of the physician. For example, the words compassionate, wonderful, great, excellent, takes, time, kind, manner, and bedside are all included in Topic 2. Thus, Topic 2 also has an obvious positive connotation referencing the kind and compassionate physician. Interestingly, our analysis revealed no negative interpersonal quality dimension. Gao et al. (2015) found that lower quality physicians (i.e., those rated lower through a traditional quality collection survey) were less likely to be rated online, and our findings may add nuance to this finding. Our data indicates, that in terms of process quality, physicians that are considered to be compassionate and kind are often rated online, whereas physicians with poor bedside manner may not be as frequently discussed. However, when it comes to the physician's practices (i.e., systems process quality) the opposite seems to be true – problems with wait times, billing, rude staff are frequently indicated and good systems service quality experiences are less frequently described. This may suggest that for physicians to improve their online ratings, they should take care to carefully manage the systems service quality elements.

Thus, three of the identified topics pertain to healthcare process quality and reflect poor systems process quality (e.g., timeliness, staff, and financial aspects of the healthcare experience), but

good interpersonal process quality (e.g., interpersonal, empathic dimensions of the healthcare experience). Therefore, prominent over all 23,537 comments were three process quality elements relating timeliness, staff, financial, interpersonal and empathic aspects of healthcare service quality, where the one relating the doctors' demeanor or interpersonal skills was positive and the two relating management issues were negative. These findings are reinforced by the overall star ratings for the different dimensions shown in Fig. 2 and suggest that many of the comments on RateMDs describe the physician's demeanor or interpersonal skills as positive, but are unhappy with the practical management elements of the business (i.e., staff, wait times, payment).

Technical competence is a view of clinical quality that López et al. (2012) suggests includes commentary discussing the physicians' knowledgeableness, attention to detail, efficiency, along with descriptions of clinical skills, follow-ups, perceptions of decisions making, perceptions of successfulness of treatment, etc. Topics 5 and 6 contain specific healthcare terminology rather than process quality elements (do not describe demeanor/empathy or management concerns). Topic 5 (positive technical quality: surgery) contains words such as: surgery, pain, surgeon, procedure, life, knee, cancer, and breast. Topic 5 also has words that suggest a positive connotation: recommend, great, highly, excellent. Topic 6 (negative technical quality: diagnosis) contains words such as: treatment, medical, tests, diagnosis, diagnosed, treated, blood, symptoms, medicine, and condition. Topic 6 has weak indications of a possible negative connotation, but with words that could also reflect the health condition, such as issues, time, and problem. So, we label this one as negative, but it may not have a strong connotation. These findings suggest that patients are relating clinical quality concerns, both positive and negative on online review websites for physicians.

Topic 4 (positive technical quality: family) is less straightforward because it appears to revolve around perceptions of care of others, containing words such as: children, baby, daughter, son, husband, pregnant, child. Topic 4 also has a positive connotation, containing words such as: wonderful, recommend, great, comfortable, good. Some elements of this topic suggest it could be categorized under positive systems or interpersonal quality, but Wallace et al. (2014) classified similar words under positive technical quality, so we adopted a similar approach. We would suggest that what is important to recognize about this topic is that it centers on care provided to family members. This suggests that the physicians are being evaluated not only by the primary patient, but also by members of patients' family.

To examine our second two questions, we used the scores for tone calculated by DICTION, the topics probabilities obtained through application of LDA, and three control variables that provide descriptive statistics for the text as independent variables in two stepwise regressions. Since there is little theory to suggest the relationships we might discover between topics, tones, and the quantitative ratings, our analysis is exploratory in nature, which suggests stepwise regression is an appropriate method to explore our second two questions.

Table. 3 presents the results of the best model found by the stepwise regression procedure using the sentiment, topic, and control variables (shown in Fig. 3 and described in the preceding sections) as independent variables and patient perceived quality (i.e., the natural log of the overall numeric rating for each physician) as the dependent variable. We examine the regression output to explore which elements of tone and service quality are driving the overall numeric star rating assigned to the physicians. The best model discovered by the stepwise regression had an R-Squared of 0.327 with variables lnOptimism, lnCertainty, lnActivity, Topic1, Topic2, Topic4, Topic5, and Topic6 entering the model.

Examining the coefficients in Table 3, we see that two of the process (negative systems quality: staff and timeliness and positive interpersonal quality: physician compassion) quality topics – Topic1 (0.137, $p < 0.001$), Topic2 (0.525, $p < 0.001$), and the positive technical quality: family topic – Topic4 (0.417, $p < 0.001$) were positively correlated with patient perceived quality. In other words, the more important negative systems quality: staff and timeliness, positive interpersonal quality: physician compassion, and positive technical quality: family reviews are to the physician's overall description, the higher the overall rating. Topic 2 and Topic 4 have two of the biggest coefficients. Topic 2 relates aspects of physician demeanor (i.e., whether the physician is perceived to be kind and makes the patient feel comfortable). Topic 4 relates aspects of care centered on other people. This suggests that positive perceptions of the physician's demeanor, and how they are perceived to treat others are crucial to the overall numeric rating assigned online. The interpretation for Topic 1 is that negative staff, timeliness, and financial aspect descriptions are also driving the patients' ratings of physicians.

The clinical quality topics related to the patients' care (i.e., rather than care of family) are also positively correlated with patient perceived quality – Topic5 (0.441, $p < 0.001$) and Topic6 (0.256, $p < 0.001$). These indicate that the more important clinical quality (positive technical quality: surgery and negative technical quality: diagnosis) reviews are to the physician's overall description, the higher the overall rating. Interestingly, only Topic 3 drops out of the stepwise regression model for patient perceived quality. Topic 3 is the only topic with a strong negative connotation. This could be an artifact of the finding by Gao et al. (2015) that lower quality physicians were less likely to be rated online. However, the fact that our topic analysis revealed negative process quality topics indicates that there are a significant number of patients with process quality complaints on the website.

One of the sentiment themes is positively correlated (0.229, $p < 0.001$) with patient perceived quality – optimism. The interpretation is the more positive the tone with which the physician is described, the higher the overall numeric rating. Thus, physicians' that are described with positive, optimistic sentiment receive higher ratings. This is reinforced by a positive relationship between Topic 2 (positive interpersonal quality: physician compassion) and patient perceived quality (0.525, $p < 0.01$). Taken together, these findings indicate that physicians perceived as kind or having good bedside manners are rated more highly. The sentiment themes certainty and activity are also positively related (0.098, $p < 0.001$; 0.088, $p < 0.001$) to patient perceived quality. This indicates that patients speaking with certainty and activity in their reviews are more likely to give the physician a higher overall rating.

In summary, these findings indicate that text analytics can be used to granularly examine online unstructured feedback. More importantly, they indicate that patient perceived quality is driven by emotional process quality concerns (i.e., positive interpersonal quality experiences and optimistic tone) and consideration of care for family members, as seen in the importance of Topic2, Topic4, and lnOptimism. That is, whether or not the physician is rated highly has a lot to do with whether they were perceived as kind and to treat others well. One practical implication of these findings is that online patient feedback may be a good indicator of interpersonal process quality elements of healthcare.

While the previous results indicate what elements of tone and topic are driving the patients' perception of healthcare quality, we also want to explore our third question of how these elements correspond to a measure of physician perception of quality. Thus, we ran another stepwise regressions of the same independent variables using the natural log of the number of mentions of each physician by other physicians as the dependent variable. Table 4 presents the statistics for the best model for the same set of in-

Table 3
Stepwise regression results for patient perceived quality.

Model 1 (Stepwise Linear Regression)::
Dependent Variable – Patient Perceived Quality
R = 0.572:: R Square = 0.327:: Adj. R Square = 0.325:: Std. Error of the Estimate:: 0.211

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics
	B	Std. Error	Beta	t	Sig.	VIF
(Constant)	−72.920	5.717		−12.755	0.000	
lnOptimism	5.869	0.478	0.229	12.278	0.000	1.917
lnCertainty	4.624	0.646	0.098	7.158	0.000	1.025
lnActivity	0.132	0.025	0.088	5.350	0.000	1.472
Topic1	0.282	0.046	0.137	6.170	0.000	2.718
Topic2	0.846	0.038	0.525	22.111	0.000	3.104
Topic4	0.792	0.043	0.417	18.619	0.000	2.759
Topic5	0.732	0.041	0.441	17.997	0.000	3.299
Topic6	0.566	0.047	0.256	11.928	0.000	2.533

Table 4
Stepwise regression results for physician perceived quality.

Model 2 (Stepwise Linear Regression)::
Dependent Variable – Physician Perceived Quality
R = 0.211:: R Square = 0.045:: Adj. R Square = 0.043:: Std. Error of the Estimate:: 0.532

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics
	B	Std. Error	Beta	t	Sig.	VIF
(Constant)	57.280	18.715		3.061	0.002	
lnCommonality	−8.018	2.698	−0.048	−2.972	0.003	1.005
Topic1	−0.296	0.083	−0.068	−3.582	0.000	1.394
Topic3	0.188	0.089	0.036	2.103	0.036	1.119
Topic4	−0.337	0.083	−0.084	−4.698	0.000	1.656
Topic5	0.401	0.073	0.114	5.484	0.000	1.678
Topic6	0.258	0.092	0.055	2.811	0.005	1.492
lnNoReviews	0.046	0.010	0.078	4.698	0.000	1.060

dependent variables, but using physician perceived quality as the dependent variable.

Our regressions revealed much different models for the patient quality perception measure and the physician quality perception measure. Examining the results in Table 4, we see that the R-Squared values for the model with the physician perceived quality dependent variable are low (0.045). This indicates that patients' perceptions of healthcare service quality do not explain much of the variance in the physicians' quality perceptions, as measured by recommendations. The sentiment or tone of the patients' reviews was not related to physician perceived quality, with one exception. Commonality had a weak but significant negative relationship (-0.048, $p < 0.05$) with physician perceived quality.

Topics 5 and 6 (positive technical quality: surgery and diagnosis) both have significant positive relationships (0.114, $p < 0.001$ and 0.055, $p < 0.01$, respectfully) with physician perceived quality. This indicates that physicians whose patient reviews include descriptions of clinical quality (surgery or diagnosis) tend to be recommended more frequently by other physicians. Topic 1 (negative systems quality: staff and timeliness) is negatively related to physician perceived quality (-0.068, $p < 0.001$), lending weak support for patient process quality complaints being predictive of physicians being recommended less by their peers. However, Topic 3 (negative systems quality: experience) has a significant, weak, and positive relationship with physician perceived quality (0.089, $p < 0.05$), which is unexpected. Topic 4 (positive technical quality: family) has a negative relationship (-0.084, $p < 0.001$) with physician perceived quality and Topic 2 (positive interpersonal quality: physician compassion) has no significant relationships with physician perceived quality. The findings for Topic 2, 3, and 4 may seem somewhat counterintuitive, but if one assumes peer physician recommendations are more likely to reflect clinical quality, then confounding findings for the process quality topics and perceptions of care of others are not terribly surprising. This is especially true given the strongest relationship was with the positive clinical quality topic. Considering these findings together, we would suggest the possibility that patient feedback is a good indicator of positive process quality and that the peer physician reviews may be more reflective of positive clinical quality. The number of reviews also had a significant positive relationship with physician perceived quality (0.078, $p < 0.001$), indicating weak support for better physicians, as defined by overall numeric patient rating, being recommended more frequently by other physicians.

Taken as a whole, our findings indicate that the peer recommendations from other physicians are likely reflecting clinical quality, whereas the online patient reviews tend to reflect emotion and process (systems and interpersonal) quality observations. Implications of these results suggest that if a consumer is searching for a compassionate physician with a well-managed practice, the online reviews may be helpful. Of interest, our findings suggest that if patient reviews indicate positive clinical quality, this may be inline with the opinions of other physicians since this topic was the most strongly related to physician perceived quality. Furthermore, process quality complaints by patients may also reflect a lower physician perceived quality. The overall implication being that while patient reviews containing process quality complaints and clinical quality assessments may not be as influential towards the overall star rating, for a consumer, mentions of either of these items may provide feedback inline with physician perceptions of the quality of their peers.

For online reviews to really be useful, our results suggest that just looking at the star rating of the physician may not be very informative. Healthcare consumers should read the textual comments left by other patients and parse them for particular discussions of process and clinical quality elements that are of concern to them regarding a potential physician. For example, if a review explicitly mentions clinical quality (e.g., details of a particular surgery) or a process quality concern (e.g., lengthy wait time), then

our findings indicate this may be information worth considering in choosing a physician based upon specific process or clinical desires, in part because this level of detail was somewhat reflective of physician perceived quality. Our results indicate two very distinct quality considerations: process and clinical, and furthermore subdivides process quality into elements of a physician's practice (systems) and elements of the physician's demeanor (interpersonal). Thus, online physician reviews may also be prone to misinterpretation if a potential patient interprets optimistic tone and positive interpersonal process quality as indicative of clinical healthcare quality. In other words, to locate a physician with a disposition that people tend to like, the star rating or a cursory examination of the comments may be useful. To gauge quality at a more granular level, the comments need to be more carefully parsed, and perhaps supplemented with additional sources. Therefore, we suggest that online reviews of physicians may be helpful depending on what quality elements a potential patient wishes to assess and if he or she interprets them within the scope of those quality elements.

5. Conclusion

In our analysis we examined unstructured textual feedback of physicians in order to determine how the extracted sentiment and topics compared to traditional identified dimensions of service quality in healthcare, what tone and topic elements were driving patients' service quality ratings, and to examine whether the relationships also held with regard to physician service quality perceptions. Specifically, we performed a topic analysis using LDA to extract six quality topics from our dataset of more than 20,000 patient reviews of more than 3700 physicians. We also applied a dictionary based text analysis technique to determine five elements of tone in our physicians' reviews. We then explored the relationships between the variables derived from our text analysis and a measure of patient perceived quality and physician perceived quality.

Our findings reflect the fact that healthcare service quality is a complex and multi-faceted area. We find that optimistic tone and elements of process quality, both systems (e.g., staff, timeliness, payment) and interpersonal (physician interpersonal skills and empathy) are important drivers of the numeric ratings assigned to physicians in online reviews. However, we find that while overall what patients say about physicians online does not relate well with peer physicians' quality assessments, patient descriptions of clinical quality are the strongest correlates to physician perceived quality. This leads us to suggest that online review forums can be useful, particularly in assessing process quality, but should be carefully parsed and perhaps used only as a secondary source, if evaluations of clinical quality or details on underperforming physicians are desired.

Exploration of unstructured textual feedback is still a rather novel technique, and is not without its limitations. Future research could evaluate variations of sentiment and topic analysis techniques not employed in the current study. It should be noted LDA employs optimization algorithms that require the user to specify the number of features desired before creating a model based upon the vocabulary of the corpus. Thus, it is possible to extract a different number of topics and to obtain different model results by rerunning the algorithm with different parameters. We based our choice of number of topics from our literature review and suggest that the topic areas extracted are useful in evaluating the unstructured physician feedback. Future research could include more extensive evaluations of topic models in this context. Other topic analysis approaches have been used in the literature (LSA, f-LDA, PCA) and several sentiment approaches are available (Tang, Tan, & Cheng, 2009). Furthermore, we used stepwise regressions to explore relationships in our dataset, which is appropriate for an exploratory data analysis like ours. Machine learning techniques (i.e., data mining algorithms) could also be employed in future studies. Future research could examine the impacts of making method substitutions in our framework. It may also be of interest in future work to explore different units of analysis; for example, models exploring characteristics aggregated at the specialty-, service- or state-levels. In summary, our study showed that analysis of unstructured quality feedback using automated text-mining procedures can provide interesting insights for patients and physicians regarding service quality.

References

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management, 24*(6), 975–990.

Alemi, F., Torii, M., Clementz, L., & Aron, D. C. (2012). Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Quality Management in Healthcare, 21*(1), 9–19.

Allison, T. H., McKenny, A. F., & Short, J. C. (2013). The effect of entrepreneurial rhetoric on microlending investment: An examination of the warm-glow effect. *Journal of Business Venturing, 28*(6), 690–707.

Anderson, E. A. (1995). Measuring service quality at a university health clinic. *International Journal of Health Care Quality Assurance, 8*(2), 32–37.

Babakus, E., & Mangold, W. G. (1992). Adapting the SERVQUAL scale to hospital services: An empirical investigation. *Health services research, 26*(6), 767–786.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava, & M. Sahami (Eds.). In *Text mining: Classification, clustering, and applications: Vol. 10* (p. 34). Boca Raton, FL: CRC Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bowers, M. R., Swan, J. E., & Koehler, W. F. (1994). What attributes determine quality and satisfaction with health care delivery? *Health Care Management Review, 19*(4), 49–55.

Butt, M. M., & de Run, E. C. (2010). Private healthcare quality: Applying a SERVQUAL model. *International Journal of Health Care Quality Assurance, 23*(7), 658–673.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems, 50*(2), 511–521.

Chen, P.-Y., Wu, S.-y., & Yoon, J. (2004). The impact of online recommendations and consumer feedback on sales. Paper presented at the ICIS 2004, Washington, DC.

Dellarocas, C. (2002). *Goodwill hunting: an economically efficient online feedback mechanism for environments with variable product quality* agent-mediated electronic commerce IV. Designing mechanisms and systems (pp. 238–252). Springer.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*(10), 1407–1424.

Donabedian, A. (1980). The definition of quality and approaches to its assessment. *Explorations in Quality Assessment and Monitoring*: vol. 1. Ann Arbor, MI: Health Administration Press.

Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013). Mining online user-generated content: Using sentiment analysis technique to study hotel service quality. Paper presented at the. *2013 46th Hawaii international conference on system sciences (HICSS), Hawaii*.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems, 45*(4), 1007–1016.

Gao, G. G., Greenwood, B. N., Agarwal, R., & McCullough, J. (2015). Vocal minority and silent majority: How do online ratings reflect population perceptions of quality? *Management Information Systems Quarterly, 39*(3), 565–589.

Hays, R. D., Davies, A. R., & Ware, J. E. (1987). *Scoring the medical outcomes study patient satisfaction questionnaire: PSQ III. MOS memorandum*. Retrieved from CA: Santa Monica.

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology and Management, 9*(3), 201–214.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259–284.

Lanning, J. A., & O'Connor, S. (1989). The health care quality quagmire: Some signposts. *Hospital and Health Services Administration, 35*(1), 39–54.

Larose, D. T., & Larose, C. D. (2015). *Data mining and predictive analytics* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Lee, H., & Kim, C. (2014). Benchmarking of service quality with data envelopment analysis. *Expert Systems with Applications, 41*(8), 3761–3768.

Lee, J., Park, D.-H., & Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic Commerce Research and Applications, 7*(3), 341–352.

Lee, S. D. (2013). I hate my doctor: Reputation, defamation, and physician-review websites. *Health Matrix, 23*(573).

López, A., Detz, A., Ratanawongsa, N., & Sarkar, U. (2012). What patients say about their doctors online: A qualitative content analysis. *Journal of General Internal Medicine, 27*(6), 685–692.

Marley, K. A., Collier, D. A., & Meyer Goldstein, S. (2004). The role of clinical and process quality in achieving patient satisfaction in hospitals. *Decision Sciences, 35*(3), 349–369.
Merrell, J. G., Levy, B. H., & Johnson, D. A. (2013). Patient assessments and online ratings of quality care: A "wake-up call" for providers. *The American Journal of gastroenterology, 108*(11), 1676–1685.
Misopoulos, F., Mitic, M., Kapoulas, A., & Karapiperis, C. (2014). Uncovering customer service experiences with Twitter: The case of airline industry. *Management Decision, 52*(4), 705–723.
Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon. com. *Management Information Systems Quarterly, 34*(1), 11.
Padma, P., Rajendran, C., & Lokachari, P. S. (2010). Service quality and its impact on customer satisfaction in Indian hospitals: Perspectives of patients and their attendants. *Benchmarking: An International Journal, 17*(6), 807–841.
Parasuraman, A., Berry, L. L., & Zeithaml, V. A. (1991). Perceived service quality as a customer-based performance measure: An empirical examination of organizational barriers using an extended service quality model. *Human Resource Management, 30*(3), 335–364.
Parasuraman, A., Berry, L. L., & Zeithaml, V. A. (1991). Refinement and reassessment of the SERVQUAL scale. *Journal of Retailing, 67*(4), 420–450.
Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *The Journal of Marketing, 49*(4), 41–50.
Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–37.
Pavlou, P. A., & Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research, 17*(4), 392–414.
Ramez, W. S. (2012). Patients' perception of health care quality, satisfaction and behavioral intention: An empirical study in Bahrain. *International Journal of Business and Social Science, 3*(18), 131–141.
Ramsaran-Fowdar, R. R. (2005). Identifying health care quality attributes. *Journal of Health and Human Services Administration, 27*(4), 428–443.
Rohini, R., & Mahadevappa, B. (2006). Service quality in Bangalore hospitals - An empirical study. *Journal of Services Research, 6*(1), 59–82.
Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. C. (2010). Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation. *Organizational Research Methods, 13*(2), 320–347.
Short, J. C., & Palmer, T. B. (2008). The application of DICTION to content analysis research in strategic management. *Organizational Research Methods, 11*(4), 727–752.
Sperkova, L., Vencovsky, F., & Bruckner, T. (2015). How to measure quality of service using unstructured data analysis: A general method design. *Journal of Systems Integration, 6*(4), 3.
Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications, 36*(7), 10760–10773.
Vandamme, R., & Leunis, J. (1993). Development of a multiple-item scale for measuring hospital service quality. *International Journal of Service Industry Management, 4*(3), 30–49.
Venkatesh, V., Zhang, X., & Sykes, T. A. (2011). Doctors do too little technology: A longitudinal field study of an electronic healthcare system implementation. *Information Systems Research, 22*(3), 523–546.
Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management, 30*(1), 123–127.
Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association, 21*(6), 1098–1103.
Ware, J. E., Snyder, M. K., & Wright, W. R. (1976a). *Development and validation of scales to measure patient satisfaction with medical care services, Vol. I, Part A: Review of literature, overview of methods and results regarding construction of scales*. Retrieved from Springfield, VA.
Ware, J. E., Snyder, M. K., & Wright, W. R. (1976b). *Development and validation of scales to measure patient satisfaction with medical care services, Vol. I, Part B: Results regarding scales constructed from the patient satisfaction questionnaire and measure of other health care perceptions*. Retrieved from Springfield, VA.
Ware, J. E., Snyder, M. K., Wright, W. R., & Davies, A. R. (1983). Defining and measuring patient satisfaction with medical care. *Evaluation and program planning, 6*(3), 247–263.
Wisniewski, M., & Wisniewski, H. (2005). Measuring service quality in a hospital colposcopy clinic. *International Journal of Health Care Quality Assurance, 18*(2/3), 217–228.
Youssef, F., Nel, D., & Bovaird, T. (1995). Service quality in NHS hospitals. *Journal of Management in Medicine, 9*(1), 66–74.
Zachary, M. A., McKenny, A. F., Short, J. C., Davis, K. M., & Wu, D. (2011). Franchise branding: An organizational identity perspective. *Journal of the Academy of Marketing Science, 39*(4), 629–645.