SERVQUAL is a popular model used to examine customers' perceptions of service quality. In this research, we text mine a website that allows users to leave comments regarding their experience with physicians. By applying analytical methods to the text generated by users on the feedback site, we explore the dimensions defined by the SERVQUAL model from freeform, non-solicited post-service perceptions. We determine which dimensions of service appear most frequently and the sentiment related to those appearances for physicians. This analysis allows us to determine the influence of the SERVQUAL dimensions on service quality perceptions of physicians using a novel approach.

KEYWORDS: SERVQUAL, service quality, text mining, healthcare

INTRODUCTION

The Oxford Dictionary defines healthcare as “the maintenance and improvement of physical and mental health, especially through the provision of medical services [http://www.oxforddictionaries.com/us/definition/american_english/healthcare].” A more robust definition may be “the diagnosis, treatment, and prevention of disease, illness, injury, and other physical and mental impairments in humans [http://en.wikipedia.org/wiki/Health_care].” It has been stated that the United States (U.S.) healthcare system is “the world's largest service industry” (Kenagy et al., 1999). The World Bank lists the U.S. as maintaining the highest health expenditure at 17.9% of gross domestic product (GDP) in 2012 [http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS]. Canada follows at 15th with 10.9% of GDP, New Zealand at 10.3% of GDP, and the United Kingdom at 9.4% of GDP. These figures represent numerically what we already know intuitively, that health care is one of the most important service industries both at an economic and at a personal level.
It is the individualistic nature of and emotionality tied to the healthcare service, coupled with the knowledge-heavy technical demands of its practice that makes service quality in this industry distinct. Kenagy et al. (1999) state that there exists both a technical and a service dimension to healthcare, “by service, we mean the myriad characteristics that shape the experience of care for patients and their loved ones other than the technical quality of diagnostic and therapeutic procedures.” Furthermore, the authors go on to suggest “most patients do not feel qualified to judge technical quality but assess their healthcare by other dimensions that reflect what they personally value.” Similarly, Naidu (2009) suggests that “patients may be unable to assess medical service technical quality accurately; hence, functional quality is usually the primary determinant.” Additionally, Naidu (2009) observes, “healthcare quality is more difficult to define than other services…mainly because it is the customer himself/herself and the quality of his/her life being evaluated.” These statements suggest that defining service quality in healthcare is a complex task.

Previous attempts to explore healthcare service quality have focused on traditional methodologies such as literature surveys (for example: Naidu, 2009) or survey-based techniques that rely heavily on the design of the data collection instrument (for example: Butt and de Run, 2010). There are many benefits to these methodological approaches and they often use well-researched theory (SERVQUAL). The current study applies a different analytical approach, couched in the same service quality theory, that allows for the use of a different type of data. Specifically, we use data that was not specifically solicited for the study from a website that serves as a forum for comments about physicians. The site collects some qualitative data (scale ratings for certain factors of service), but most importantly allows for free form customer feedback. Guided by the dimensions of a well-known service quality model (SERVQUAL) and a survey of the literature regarding service quality in healthcare, we apply a text mining methodology to the collected data composed of 763,966 reviews of 98,698 physicians in order to explore the importance (defined as frequency and sentiment or use) of the service quality dimensions in post-service perceptions of healthcare.

LITERATURE REVIEW

SERVQUAL is an instrument for assessing service quality that was originally developed by Parasuraman et al. (1985, 1988) and later refined by the same authors (Parasuraman et al., 1991a, 1991b, 1994). The construct of service quality, as measured by SERVQUAL, involves perceived quality, which is the consumer’s judgment of an entity’s overall excellence or superiority (Zeithaml, 1987) and is a form of attitude that results from a comparison of expectations with perceptions of quality (Parasuraman et al., 1988). Oliver (1981) defined attitude as “the consumer’s relatively enduring affective orientation for a product, store, or process (e.g. customer service)”.

The initial study focusing on service quality by Parasuraman et al. (1985) developed a conceptual model of service quality where customers’ perceptions of service quality performance are influenced by four gaps (Table 1) that can impede service. Service quality is then defined as the difference between customer expectations and perceptions and is identified as Gap 5. It is expected that Gaps 1 – 4 will influence Gap 5.

<table>
<thead>
<tr>
<th>Gap 1</th>
<th>Difference between customer expectations and management perceptions of customer expectations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gap 2</td>
<td>Difference between management perceptions of customer expectations and service</td>
</tr>
</tbody>
</table>
Villacis Calderon et al.  
Exploring the SERVQUAL Dimensions in Healthcare

<table>
<thead>
<tr>
<th>Gap 3</th>
<th>Difference between service quality specifications and the service actually delivered.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gap 4</td>
<td>Difference between service delivery and what is communicated about the service to the customers.</td>
</tr>
</tbody>
</table>

Parasuraman et al. (1988) next developed a procedure for quantifying customers’ assessment of service quality performance and created a multiple-item SERVQUAL instrument with a scale consisting of 22 items representing 5 dimensions. The five dimensions are **Tangibles**, **Reliability**, **Responsiveness**, **Assurance**, and **Empathy** (Table 2).

<table>
<thead>
<tr>
<th>Table 2: SERVQUAL dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
</tr>
<tr>
<td>Tangibles</td>
</tr>
<tr>
<td>Reliability</td>
</tr>
<tr>
<td>Responsiveness</td>
</tr>
<tr>
<td>Assurance</td>
</tr>
<tr>
<td>Empathy</td>
</tr>
</tbody>
</table>

The SERVQUAL model has been well studied and widely adopted (Asubonteg et al., 1996; Buttle, 1996; Lam and Woo, 1997). It has been reported to be the preferred model for measuring service quality across various sectors (Butt and de Run, 2010). The model, however, is not without its critics, who believe that a performance-based measure is more appropriate for measuring service quality (Sureshchandar et al., 2001). However, patients in a health care setting tend to rely on functional attributes such as facilities, cleanliness, and personnel attitudes rather than technical attributes because they do not typically have the technical medical expertise (Babakus and Boller, 1981; Lanning and O’Connor, 1989). Consequently, SERVQUAL is the appropriate choice for our research as we have access to the customer’s (or patient’s) comments and consequently their attitude, based on functional attributes, about their service quality.

SERVQUAL has been used to measure service quality in healthcare, banking, hotel, fast food, retail chains, information systems, and libraries (Ladhari, 2009). Butt and de Run (2010) developed and tested a SERVQUAL model for measuring private healthcare service quality in Malaysia. They created a composite variable based on a priori dimensions and identified SERVQUAL as a robust instrument for measuring Malaysian healthcare service providers. Wisniewski and Wisniewski (2005) adapted the SERVQUAL instrument and used it to measure service quality in a hospital colposcopy clinic. They concluded that SERVQUAL has a useful role in monitoring and assessing their service quality. Rohini and Mahadevappa (2006) used SERVQUAL to study the perception of service quality in five Bangalore based hospitals. They reported that SERVQUAL 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 health care related service quality include Anderson, 1995; Babakus and Mangold, 1992; Bowers et al., 1994; Ramez, 2012; Wisniewski and Wisniewski, 2005; and Youssef et al., 1995.
Based on our literature review, we developed an exploratory framework of words related to each SERVQUAL dimension (Figure 1). Many of the studies we reviewed rely on the five SERVQUAL dimensions, noting that these are the most crucial dimensions. In order to obtain our initial framework, we analyzed each survey-based SERVQUAL research article examining service quality in healthcare environments (Butler et al., 1996; Butt and de Run, 2010; Dagger et al., 2007; Haque et al., 2012; Headley and Miller, 1993; Jiang et al., 2002; Lytle and Mokwa, 1992; Ramsaran-Fowder, 2005; Sohail, 2003; Soita, 2012; Swan, 2003). From each article we extracted the dimensions of service quality proposed for the healthcare environment and the described characteristics of each SERVQUAL dimension for this context. This resulted in a collection of meaningful words relating service quality characteristics for each dimension extracted from the prior literature. The words were then assigned to the relevant dimension. That is, if the word was used to describe a particular dimension in the article, the word was assigned to that corresponding dimension in our framework. This framework will be used in the current study to explore general service quality trends in a large data set. The analysis conducted in the current study will help inform the refinement of this framework in future work. Using the initial framework of words shown in Figure 1, in the next section, we will explore the text of the patient reviews to determine if we could extract any dimension(s) of SERVQUAL that were embedded in the patient comments. We will then expand our analysis to determine if there is a noticeable difference in the tone of the comments extracted for each dimension. Finally, we will identify key words (including, but not limited to the words from the framework in Figure 1) for each dimension. The sentiment and key word extraction will help us refine our framework in future work, as well as inform our next round of analysis. Meanwhile, the results of the current study are valuable because the ability to tie patient comments to the SERVQUAL dimensions provides opportunities to identify high-level trends related to the perception of service quality in healthcare. The analysis also helps us identify the relative importance of particular service quality dimensions in the healthcare arena.

![Figure 1: Text Mining Framework from Literature Review](image-url)
To conduct the analysis, we used data scraped from the website: http://www.ratemds.com. Using a scraper written in Python by one of the authors, we collected 763,966 physician reviews from approximately 98,698 physicians. The data included reviews for physicians in approximately 50 specialties, including: Family/G.P., Gynecologist, Dentist, Dermatologist, Cardiologist, Cosmetic/Plastic Surgeon, Neurologist, Psychiatrist, Orthopedics/Sports. Figure 2 shows the locations of the physicians reviewed. It is notable that the data covered a sizable amount of both the United States and Canada. To provide an idea of the amount of text the sample includes, when the words from all the reviews are tabulated, there are 23,520,483 words. Of these 23 million words, 278,366 are unique.

Our first step was to determine the occurrence of the words identified for each dimension in the SERVQUAL framework for healthcare in Figure 1 that resulted from the literature review. The number of reviews containing words from each of the five dimensions was tallied and is shown in Table 3. Table 3 illustrates that approximately 56% of the total reviews were used in this analysis. In other words, at least one of the words identified as belonging to a SERVQUAL dimension in Figure 1 was found in 56% of all reviews (431,590 unique reviews). Words from the “Responsiveness” dimension occurred with the most frequency (in 39% of the data or in more than 298,000 of the 763,966 reviews). Words from the empathy and tangibles dimensions also appear in large percentages of the data (35.8% and 20.8% respectively). In this large amount of data, the trends emerging are quite interesting. The words relating to responsiveness revolve around dependability, timeliness, and availability. Empathy relates to bedside manner, the physician’s demeanor, courtesy, attention, etc. The more frequent appearance of the words from these dimensions when compared to those from assurance and reliability, which revolve around technical competence and expertise, seems to reinforce the suggestion in prior qualitative research that the soft skills are more crucial to the overall perception of service quality in healthcare than those reflecting the technical aspects.
Our next step was to examine each dimension more closely. Using the Naïve Bayesian classifier available for sentiment analysis in the Natural Language Toolkit (NLTK), we calculated the sentiment score. The Naïve Bayesian classifier is a supervised learning technique that requires labeled data to train. A set of label text from movie reviews is available for use to train the classifier. For this exploratory analysis, we trained the classifier using this publicly available labeled data. Therefore, our analysis is limited by the quality of the training set and the results should be interpreted under this context. Future work will include testing with other training sets and the creation of one from our own data. The Naïve Bayesian classifier utilized in this study examined each review and calculated a sentiment score for that review based upon the number of words labeled as “negative” and “positive” in the training set. For example, each positively labeled word received 1 point and each negatively labeled word received -1 points (unlabeled words were ignored). Hence, if for example, the review was “the doctor was nice and friendly, but the staff was unhelpful” and the training set had labeled “nice” and “friendly” as positive and “unhelpful” as negative, the sentiment score would have been 2+1 = 1. Figure 3 shows that indeed some reviews were overall negative in their sentiment (used more negatively labeled words), but a higher number were overall positive (used more positively labeled words). For clarity, the frequency of 0 (or neutral) reviews was omitted from the graph as many reviews were classified as 0.

Table 3: Reviews Per Dimension

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Total Reviews</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assurance</td>
<td>56538</td>
<td>7.40%</td>
</tr>
<tr>
<td>Empathy</td>
<td>273647</td>
<td>35.82%</td>
</tr>
<tr>
<td>Reliability</td>
<td>112802</td>
<td>14.77%</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>298539</td>
<td>39.08%</td>
</tr>
<tr>
<td>Tangibles</td>
<td>159192</td>
<td>20.84%</td>
</tr>
<tr>
<td>Unique Reviews</td>
<td>431590</td>
<td>56.49%</td>
</tr>
</tbody>
</table>

Figure 3: Distribution of Sentiment Scores (Positive Values Indicate that Positively Labeled Words Outnumber Negatively Labeled Words by that Amount)
Figure 4 shows the average sentiment score for each dimension. The graph indicates that on average the reviews were positive in the words used for all dimensions. The reviews that corresponded to the empathy and responsiveness categories used slightly less positive words on average, however, not by much. It is interesting that the reviews skewed so heavily towards the positive side of the sentiment score using this training set. The slightly lower positive ratings for empathy and responsiveness could be interpreted as reinforcing the importance of the soft skills as suggested in prior literature.

![Figure 4: Average Sentiment Scores per Dimension](image)

Last, we turned to word frequency to further explore our SERVQUAL dimensions. We had extracted the reviews for each dimension based upon the words in Figure 1. However, that did not mean that the words from Figure 1 were the most prevalent words in the set of reviews extracted for that dimension. In order to explore this correspondence, we extracted the most frequent words for each dimension. Interestingly, there was a very significant amount of overlap in the most frequent words for each dimension. Examining the 50 most frequent words in each dimension, there were only 16 words that were unique to only one dimension. Some of these were quite insightful. For example, trust, quality, correct, and highly were unique and frequent for the assurance dimension (Figure 5 shows a wordcloud for the top 50 words appearing in the assurance dimension). This suggests that technical quality is being talked about most in these reviews. For the empathy dimension, bedside and manner were unique and frequent. This suggests the importance of the physician’s personal approach to this dimension. Understand, information, and helpful were unique and frequent in the reliability dimension reviews suggesting helpfulness and understanding may be a focus. Hours was the only unique and frequent word in the responsiveness dimension, reinforcing the criticality of time. Diagnosis and pain were meaningful, unique and frequent in the tangibles dimension, which lends some credence to the idea of the importance of the deliverable to the healthcare service.
DISCUSSION AND CONCLUSIONS

Our initial preliminary analysis supports the possibility of extracting the SERVQUAL dimensions from freeform, non-solicited post-service perceptions in the form of client reviews to an online feedback forum. Our analysis suggests that the soft skills represented in the responsiveness and empathy dimensions are indeed important as suggested in prior literature. Furthermore our sentiment scores determined that the overall tone of the feedback was positive in nature, but less positive words were used in the reviews corresponding to these soft skills. Finally, a frequency analysis suggests that similar words are frequent in each dimension regardless of how the dimensions were extracted, but that particular meaningful words that closely corresponded to the dimension in question were present for each dimension.

There were several limitations of this study. First, it was exploratory in nature since it was our first attempt to define the SERVQUAL dimensions using words from the literature and the intent was to conduct a basic analysis to determine obvious trends that may be obtained from the proposed framework. The framework presented will need to be refined in subsequent analysis based upon the findings of this study, as well as utilizing text-mining tools such as synonym facilities. This first study was purposefully limited to single word frequencies; future work will expand to examine multi-word associations and more sophisticated analysis tools.

REFERENCES


