DECISION SCIENCES
A JOURNAL OF THE DECISION SCIENCES INSTITUTE

Decision Sciences
Volume 00 Number 0
xxx 2020

Facebook Hospital Reviews: Automated Service Quality Detection and Relationships with Patient Satisfaction

Nohel Zaman
Loyola Marymount University, 1 LMU Drive, Los Angeles, CA, 90045, e-mail: nohel.zaman@lmu.edu

David M. Goldberg[†]
San Diego State University, 5500 Campanile Drive, San Diego, CA, 92182, e-mail: dgoldberg@sdsu.edu

Alan S. Abrahams
Virginia Tech, 880 West Campus Drive, Blacksburg, VA, 24061, e-mail: abra@vt.edu

Richard A. Essig
Virginia Tech, 880 West Campus Drive, Blacksburg, VA, 24061, e-mail: ressig19@vt.edu

ABSTRACT

As patient satisfaction is heavily linked to their choice of provider and medical outcomes, hospital administrations routinely consider a bevy of factors to improve patient satisfaction. These considerations are complex, so targeting the most important areas for improvement is challenging. However, consumers' online reviews of their hospital experience provide a vital lens into the factors associated with their satisfaction. In this study, we use a large dataset of Facebook reviews to construct a taxonomy of potential service attributes that consumers discuss online. We find partial overlap between this taxonomy and prior works and more traditional survey measures; the specific mix of service attributes found in these reviews is unique. Next, we utilize regression modeling to determine which service attributes are most closely associated with star ratings, which we use to measure overall satisfaction. This study demonstrates that mentions of waiting times, treatment effectiveness, communication, diagnostic quality, environmental sanitation, and cost considerations tend to be most associated with patients' overall ratings. Finally, we construct text analyses to rapidly detect consumers' mentions of these service attributes in an automated manner. We derive a set of "smoke terms," or terms especially prevalent in posts that mention specific service attributes. We find that these are generally non-emotive terms, indicating limited utility of traditional sentiment analysis. Managerially, this information helps to prioritize the areas in greatest need of improvement. Additionally, generating smoke terms for each service attribute aids health care policy makers and providers in rapidly monitoring concerns and adjusting

[†] Corresponding author.

policies or resources to improve service. [Submitted: August 20, 2018. Revised: June 17, 2020. Accepted: June 22, 2020.]

Subject Areas: Decision Support, Hospitals, Online Reviews, Service Quality, and Text Analytics.

INTRODUCTION

The health care industry has increasingly become connected with online media, as patients now routinely rate both doctors and hospitals on social media sites (Greaves, Ramirez-Cano, Millett, Darzi, & Donaldson, 2013; Pai & Alathur, 2018). These ratings serve as major indicators of the quality of care at medical facilities. For example, Bardach, Asteria-Peñaloza, Boscardin, and Dudley (2012) found that Yelp.com ratings of hospitals are positively associated with traditional survey measures, the Hospital Consumer Assessment of Healthcare Providers and Systems, or HCAHPS. Despite the importance of utilizing this data to better patient care, many medical institutions have yet to take advantage of the feedback available online, possibly due in part to uncertainty about how best to parse and analyze the large and unstructured dataset of feedback (Yang et al., 2018).

Several studies have used online media to understand patient feedback in hospital settings; however, making sense of the large volume of online discussions has proved to be a challenging problem. For instance, as most online posts do not directly describe service quality, it can be difficult to detect the most pressing posts; Hawkins et al. (2015) found that only 10% of hospital-related tweets concerned service quality. Additionally, prior work has often found that these posts are difficult to classify, as Hawkins et al. (2015) found that 77.3% of the tweets to be classified as "general" feedback. In this article, we focus on addressing the following three primary research questions. First (RQ1), which service attributes are present in online reviews of hospitals, and how prevalent is each service attribute? Second (RQ2), which service attribute values are most strongly associated with the final overall star rating chosen by the patient? Third (RQ3), to what extent can text analytic methodologies be utilized to automatically extract reviews of interest to a particular service attribute?

In this article, our goal is to develop tools to shed light on service quality concerns by classifying each service attribute into specific values. Furthermore, we seek to understand relationships between these service attributes and patients' overall star ratings of hospitals. Understanding what drives these ratings has important managerial implications for the hospital industry because the winnowing out and prioritizing of key linguistic terms related to service attributes can help health care providers to understand perceptions of their services and to target their efforts to improve quality. We demonstrate and develop these tools by collecting the patient reviews from one of the most active and popular social media outlets, Facebook. As Lagu et al. (2016) comment, Facebook "opens doors" that provide relevant patient feedback (i.e., online comments), which hospitals can use in their efforts to improve quality of care.

The first major contribution of this study is in assessing the specific service attribute present in online hospital reviews as well as the relative prevalence of each service attribute. Although several papers have examined service attributes for different intentions, there is a substantial level of variability in researchers' choices of service attributes for analysis (Hawkins et al., 2015; Wagland et al., 2015; Hao & Zhang, 2016; Zhang et al., 2018). Thus, a holistic approach of coding thousands of reviews for each attribute can unify this literature and clarify the utility of these reviews.

Second, we extend the specific service attributes of the past studies (Hawkins et al., 2015) and examine their association with star ratings, which are generally indicative of consumers' perceptions of quality (Chevalier & Mayzlin, 2006; Hu, Koh, & Reddy, 2014). Significant relationships have also been observed between patients' star ratings of hospitals and some key terms, specifically key terms referencing mortality and infection rates (Greaves et al., 2012). Investigating patients' complaints and quality concerns provides an opportunity to evaluate properties of health care services. In the United States, 95% of hospitals have a Facebook page, each of which can collect thousands of online reviews, and the overall quantitative ratings provide new insight about patient experience (Rozenblum, Greaves, & Bates, 2017). Understanding the service attributes most associated with these ratings can better explain the mechanisms associated with perceptions of service quality.

Due to the large volume of online reviews on Facebook, it would be cumbersome to manually examine all reviews on hospitals' Facebook pages. Based on previous research related to topic classification for measuring quality of care in U.S. hospitals (Hawkins et al., 2015), our third major contribution is proposing a toolset for investigating different service attributes present in patient reviews of hospitals on Facebook. We seek to identify whether and the extent to which emotive valence (i.e., negative, positive, or neutral) plays a significant role in influencing hospital star ratings compared to the specific terms explicitly mentioned in the reviews. We aim to extract the most prevalent key terms for each service attribute value, which aid in rapid discovery and investigation of specific concerns of hospital services from unstructured online text. Greaves et al. (2013) suggest generating distinctive phrases for each attribute as opposed to single word-level analysis. Our study adds to the literature (Wallace, Paul, Sarkar, Trikalinos, & Dredze, 2014; Hawkins et al., 2015) not only by using different service attributes, but also by disambiguating by finding distinctive multiword phrases. This novel information could be routinely extracted, processed, and interpreted by health care providers and regulators to monitor performance.

LITERATURE REVIEW

Online Ratings and Patient Experiences

Consumers are increasingly writing reviews about their quality of care, rating both hospitals and physicians on social media sites (Greaves et al., 2013; Pai & Alathur, 2018). Such ratings are important differentiators among health care organizations and providers and may have major future effects on behavior and decisions of patients (Rozenblum & Bates, 2013). Research has demonstrated that online ratings of patient care are positively correlated with several traditional (conventional) metrics of service and quality. For example, Bardach et al. (2012) investigated the relationship between commercial website ratings of hospitals (Yelp.com) and traditional hospital performance aspects ("Hospital Consumer Assessment of Healthcare Providers and Systems," or HCAHPS), finding significant associations between the Yelp star scores and HCAHPS overall survey scores. Prior research has shown that although not all hospitals collect this data (Yang et al., 2018), those that do experience improved quality of care (Griffiths & Leaver, 2018). Greaves et al. (2012) reinforced this finding by showing significant associations between web-based patient ratings on the National Health Service (NHS) Choices website and traditional paper-based survey aspects of patients' experiences in hospitals. In addition to industry stakeholders, consumers also make use of online reviews, as surveys have found that 91% of consumers read online reviews whether of products or services, and 84% trust online feedback equivalently to personal recommendations (BrightLocal, 2016). Negative feedback is particularly concerning, as it tends to weigh more heavily in user opinions than positive feedback (Wolf & Muhanna, 2011). As such, online opinion can greatly sway the expectations of potential consumers and affect sales (Thirumalai & Sinha, 2009).

Apart from the numerical ratings, the textual content in patient reviews could be a valuable resource for health care providers to improve their services (Rastegar-Mojarad, Ye, Wall, Murali, & Lin, 2015). As an example, data on patient experiences is becoming an essential component in the value-based purchasing program proposed by the Centers for Medicare and Medicaid Services (CMS). The HCAHPS survey is now required nationally in the U.S., providing a standardized survey score for measuring patient satisfaction on hospital care (Campbell & Li, 2017). Beginning in 2017, the U.S. government ties Medicare reimbursements to patient responses on HCAHPS, with poor-scoring hospitals not receiving full reimbursements. However, due to chronic low response rates of about 30 percent, there is concern as to how well these datasets represent patient experiences. Other data sources such as Leapfrog and ProPublica record hospital ratings based on patient safety and surgeon complication rates respectively. As a result, patients may experience difficulty in choosing a hospital due to the challenge of collecting all the information from these multiple data sources (Rosen, 2016). However, contrary to the response rate of HCAHPS and the limitation of the other data sources, the patient reviews posted on social media websites deliver a unique and diverse perspective for patients and health care providers alike to understand patient satisfaction (Rozenblum & Bates, 2013). Although the overall response rates across all platforms are unknown, utilizing a combination of data sources ensures that more patient feedback is considered.

Rozenblum and Bates (2013) emphasize the growing importance of social media feedback/data in regards to patients' experiences and asserted that this information will complement traditional patient surveys and help to identify poor versus outstanding care. Additionally, Greaves et al. (2013) described the possibility of using the "cloud of patient experience" on the Internet for detection of poor quality care. With many possible sources of information (e.g., rating sites, patient forums, social networks), the authors suggested the value of comparing between conventional aspects of patient experience (e.g., HCAHPS) and information sourced from online media.

Text and Sentiment Analysis

Text classification refers to a process in which a set of textual documents is organized into categories based on the content of the documents. Text classification has become popular in medical informatics, such as applications to social media networks to design an early warning system for Adverse Drug Reactions (ADRs) (Yang, Kiang, & Shang, 2015). James, Calderon, and Cook (2017) used Latent Dirichlet Allocation (LDA), an unsupervised topic modeling methodology, to examine the primary topics of discussion in patient reviews. Several previous studies (López, Detz, Ratanawongsa, & Sarkar, 2012; Lagu et al., 2016; Ranard et al., 2016; Devarakonda et al., 2017) contributed in segmenting categories of quality issues in the health care domain. Doing-Harris, Mowery, Daniels, Chapman, and Conway (2016) used a topic-modeling approach to extract the most common topics (e.g., “appointment wait,” “empathy,” “friendliness,” etc.) within only negative comments. Hawkins et al. (2015) classify social media discussions of hospital service attributes, such as “waiting time,” “communication,” “food,” “general,” etc. However, as the majority of patient tweets fall into the “general” category (Hawkins et al., 2015), further research is required to build a better understanding of the dimensions of service quality mentioned in online media.

Sentiment analysis is a well-known text classification technique that seeks to quantify the direction and/or magnitude of the emotive content in a body of text. Sentiment analysis has been applied to a wide range of text analytic domains. For instance, Farhadloo, Patterson, and Rolland (2016) used sentiment analysis with online TripAdvisor reviews to analyze overall customer satisfaction. Likewise, in health care, sentiment analysis had been used to help understand the mood of patients in health care/hospital settings (Rodrigues, das Dores, Camilo-Junior, & Rosa, 2016). Traditional sentiment analysis alone, however, poses limitations, especially if the body of text contains opinions, categories, or domain-related information impertinent to positive/negative emotive valence. For example, Abrahams, Jiao, Wang, and Fan (2012) assess the effectiveness of sentiment analysis in uncovering vehicle defects discussed in online media, and the authors concluded that sentiment analysis was an insufficient tool when applied on its own. In fact, many of the terms associated with serious defects, such as the term “airbag,” are non-emotive terms that are germane only within the context of the industry of study.

Previous studies in text analytics created industry-specific dictionaries of distinctive terms for specific text categories (Abrahams et al., 2012; Abrahams, Fan, Wang, Zhang, & Jiao, 2015; Adams, Gruss, & Abrahams, 2017; Goldberg & Abrahams, 2018; Mummalaneni, Gruss, Goldberg, Ehsani, & Abrahams, 2018). These dictionaries are comprised of specific terms referred to as "smoke terms,” which are significantly more prevalent in text of interest (e.g., quality concerns). Importantly, as an alternative to sentiment analysis, smoke terms are tuned to the specific linguistic properties of the industry of interest, so they may include both emotive and non-emotive terms. Smoke terms may be single words (unigrams), but researchers have also extended the methodology to 2-word (bigram) and 3-word (trigram) phrases. Using these terms, researchers and practitioners can filter out the most pertinent bits of text for categories of interest. Each term is assigned a weight based on a prevalence metric (Fan, Gordon, & Pathak, 2005) such that new (unseen) text may be scored based on a total “smoke score,” where higher scores reflect greater concerns.

HYPOTHESIS DEVELOPMENT

Table 1 summarizes previous research using text analysis of the Internet and social media in various categories of service attributes in health care. For each study, Table 1 shows the data source, number of records (reviews) analyzed, service attributes analyzed, whether service key terms were analyzed, and whether associations with overall star ratings were analyzed. We classified the service attributes using the classification proposed by Hawkins et al. (2015) with the addition of the patient-reported *diagnosis* (Wagland et al., 2015) and *incident type* (specialty areas) (Hao & Zhang, 2016; Zhang et al., 2018) that were not assessed in that study.

Reviewing Table 1, we see a wide range of possible service attributes studied. Some service attributes are widely studied in many prior works, such as the hospital environment or communication, whereas other attributes are only studied in an isolated work, such as food or readmission rates. In some cases, differences in service attributes studied may be attributable to a difference in scope; for example, interpersonal manner, staff interactions, and dignity and respect may fall under the broad umbrella of communication. Based on our review of prior work, the most common set of service attributes studied include environment, communication, treatment, waiting time, cost, and technical competence. No single prior work studied the exact combination of these service attributes together, but based on a combination of their attributes of study, we expect that service attributes in online reviews can be categorized into this taxonomy. Thus, we posit Hypothesis 1.

H1: *Service attributes discussed in online hospital reviews can be placed in a taxonomy of communication, treatment, waiting time, cost, and technical competence.*

Next, our work examines the relationship between the service attributes observed in online reviews and star ratings, which consumers use as a holistic measure of their perception of service quality (Chevalier & Mayzlin, 2006; Hu et al., 2014). Patient satisfaction is multi-faceted and has been conceptualized in different forms in prior work. For instance, Marley, Collier, and Meyer Goldstein (2004) delineate between the clinical service dimension of patient satisfaction, concerned with the quality of outcomes produced by medical procedures (diagnoses, treatments, etc.), and the process service dimension of patient satisfaction, concerned with the manner in which medical procedures are delivered (communication, waiting times, etc.). Jonsson, Johansson, Ancarani, Di Mauro, and Giammanco (2011) empirically measure these in a medical ward, using a taxonomy derived from Parasuraman, Zeithaml, and Berry (1985) to assess satisfaction with tangibles (facilities, equipment, and personnel), reliability, responsiveness, assurance, and empathy. Given the multifaceted nature of patient satisfaction, we expect that no singular service attribute will explain the entirety of overall star ratings, but rather each service attribute will be associated with overall star ratings and will contribute a piece of our understanding. Each service attribute can be

Table 1: Survey of prior work.

			Service attribute classification																						Service key term(s)				
Study	Data source(s)	Number of records	Environment	Communication	Treatment	Diagnosis	Waiting time	Cost (money)	Incident types	Pain management	Food	General (overall)	System issues	Technical competence	Interpersonal manner	Effective care	Safe care	Medical instructions	Staff interaction	Recommendations	Readmission rates	Bedside manner	Service	Dignity and respect	Emotive words	Emotive phrases	Non-emotive words	Non-emotive phrases	Association with overall star ratings
Bardach et al. (2012)	Yelp	6,260 patient ratings	X	X						X								X											X
Bardach et al. (2015)	Yelp, Yahoo, and Google	244 reviews	X	X				X		X			X					X											
Brody and Elhadad (2010)	RateMDs	33,654 reviews			X		X	X						X					X	X									
Campbell and Li (2017)	Facebook and HCAHPS	102 hospitals with star ratings	X	X				X										X											X
Doing-Harris et al. (2016)	Press Ganey satisfaction survey	51,234 free-text responses	X	X			X					X													X	X	X	X	
Glover et al. (2015)	Facebook	679 hospitals with star ratings																			X								X
Greaves et al. (2012)	NHS Choices site	10,274 posts	X																	X									X

Continued

Table 1: Continued.

			Service attribute classification																						Service key term(s)				
Study	Data source(s)	Number of records	Environment	Communication	Treatment	Diagnosis	Waiting time	Cost (money)	Incident types	Pain management	Food	General (overall)	System issues	Technical competence	Interpersonal manner	Effective care	Safe care	Medical instructions	Staff interaction	Recommendations	Readmission rates	Bedside manner	Service	Dignity and respect	Emotive words	Emotive phrases	Non-emotive words	Non-emotive phrases	Association with overall star ratings
Greaves et al. (2013)	NHS Choices site	6,412 posts	X									X												X	X	X	X	X	
Greaves et al. (2014)	Twitter	1,000 tweets	X				X									X	X		X										
Hao and Zhang (2016)	Good Doctor site	731,264 reviews			X				X			X		X								X			X	X	X	X	
Hawkins et al. (2015)	Twitter	11,602 tweets	X	X	X		X	X		X	X	X						X											
Jung et al. (2015)	Local online groups for parents and caregivers	9,450 messages	X	X	X			X					X										X		X	X	X	X	
Lagu et al. (2016)	Facebook	47 comments	X	X			X							X															
Lopez et al. (2012)	RateMDs and Yelp	712 reviews											X	X	X														

Continued

Table 1: Continued.

			Service attribute classification																					Service key term(s)				
Study	Data source(s)	Number of records	Environment	Communication	Treatment	Diagnosis	Waiting time	Cost (money)	Incident types	Pain management	Food	General (overall)	System issues	Technical competence	Interpersonal manner	Effective care	Safe care	Medical instructions	Staff interaction	Recommendations	Readmission rates	Bedside manner	Service	Dignity and respect	Emotive words	Emotive phrases	Non-emotive words	Non-emotive phrases	Association with overall star ratings
Paul et al. (2013)	RateMDs	50,000 reviews											X	X	X										X		X		
Ranard et al. (2016)	Yelp	16,862 reviews		X	X		X	X																					X
Rastegar-Mojarad et al. (2015)	Yelp	6,914 reviews							X																X		X		
Wagland et al. (2015)	National colorectal PROM survey	400 free-text comments		X	X	X													X										
Wallace et al. (2014)	RateMDs	60,000 reviews											X	X	X										X		X		

mentioned in either a positive or a negative sense. For instance, a patient could remark that the waiting time was very quick (positive) or that it was very long (negative). For service attributes mentioned in a positive sense, we expect the association with star ratings to be positive, and for service attributes mentioned in a negative sense, we expect the association with star ratings to be negative.

H2: *Service attributes discussed in online hospital reviews in a positive sense will be positively associated with overall satisfaction, while service attributes mentioned in online hospital reviews in a negative sense will be negatively associated with overall satisfaction.*

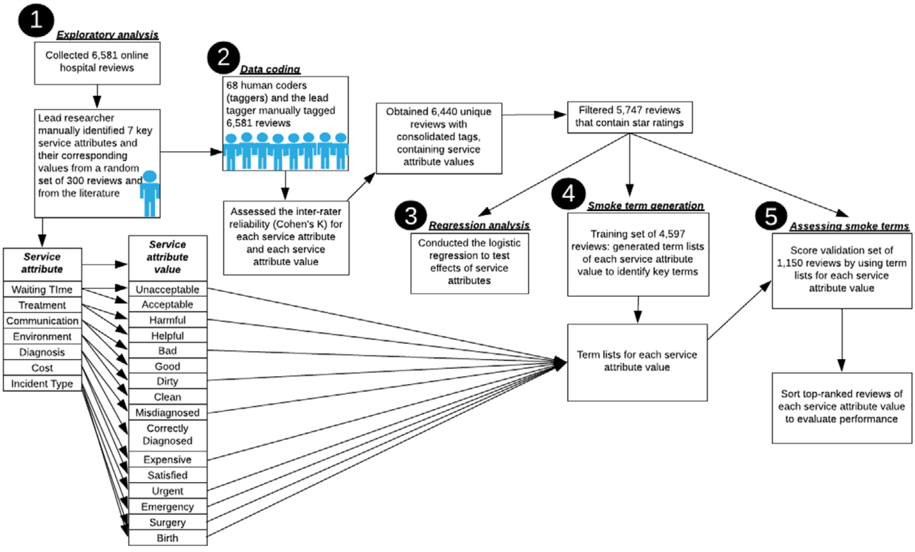
Finally, our work examines methods by which we may detect mentions of key service attributes using automated text analyses. Hawkins et al. (2015) indicate that this may be a particularly difficult problem as many online posts do not explicitly refer to any service attribute(s). Abrahams et al. (2012, 2015) indicate that the detection of unique attributes in online media can be highly context-specific, as patients tend to refer to their experiences in terms that relate to the domain of interest but do not invoke strong emotive sentiment. For example, the phrase “long wait” is not emotively strong, even though it indicates an element of a negative experience in the context of a hospital visit. Although sentiment analysis has been applied effectively in a wide range of applications (Farhadloo et al., 2016), we expect it to only be of limited utility for differentiating between specialized service attributes. “Smoke terms,” or unique words and phrases particularly prevalent in online posts that refer to specific attributes, have been shown to be effective for differentiating between more nuanced classifications (Abrahams et al., 2012, 2015; Adams et al., 2017; Goldberg & Abrahams, 2018; Mummalaneni et al., 2018). Thus, we expect that smoke terms will be more effective in this task than sentiment analyses, and we posit Hypothesis 3.

H3: *Smoke terms will provide superior performance in categorizing online hospital reviews by service attribute relative to sentiment analyses.*

METHODOLOGY

In this section, we describe the steps involved in the methodology employed in this study. In section “Exploratory Analysis,” we performed an exploratory analysis of our online reviews to discern which service attributes were most prevalent. Then, in section “Data Coding,” we coded a dataset of online reviews based on the service attributes detected. Next, in section “Regression Analysis,” we performed a regression analysis to determine the association between these service attributes and overall star ratings. In section “Smoke Term Generation,” we used text analytics methods to generate key “smoke” terms capable of rapidly filtering reviews according to these service attributes. Finally, in section “Assessing Smoke Terms,” we assessed the performance of these terms. Figure 1 displays a graphic overview of the methodology performed in this study.

Figure 1: Overview of methodology.


1
Exploratory analysis
Collected 6,581 online hospital reviews
Lead researcher manually identified 7 key service attributes and their corresponding values from a random set of 300 reviews and from the literature
2
Data coding
68 human coders (taggers) and the lead tagger manually tagged 6,581 reviews
Assessed the inter-rater reliability (Cohen's K) for each service attribute and each service attribute value
Obtained 6,440 unique reviews with consolidated tags, containing service attribute values
Filtered 5,747 reviews that contain star ratings
3
Regression analysis
Conducted the logistic regression to test effects of service attributes
4
Smoke term generation
Training set of 4,597 reviews: generated term lists of each service attribute value to identify key terms
Term lists for each service attribute value
5
Assessing smoke terms
Score validation set of 1,150 reviews by using term lists for each service attribute value
Sort top-ranked reviews of each service attribute value to evaluate performance
Service attribute
Waiting TIme
Treatment
Communication
Environment
Diagnosis
Cost
Incident Type
Service attribute value
Unacceptable
Acceptable
Harmful
Helpful
Bad
Good
Dirty
Clean
Misdiagnosed
Correctly Diagnosed
Expensive
Satisfied
Urgent
Emergency
Surgery
Birth

Exploratory Analysis

We chose to analyze user reviews from the top 100 U.S. hospitals identified by Becker's Hospital Review, 2016, reported on Facebook. Thus, these hospitals are widely utilized and ensure thorough geographic coverage of the United States. The top 100 hospitals together accounted for 106,541 beds in 2016. We collected all online hospital reviews available on the Facebook pages of these hospitals. In total, we collected 6,581 online hospital reviews dating from September 24, 2011, to March 09, 2016. Next, the lead researcher manually examined a set of 300 random hospital reviews to determine appropriate service attributes based on these reviews and using the groundwork from the previous studies (see Table 1). This process was initially structured such that the lead researcher searched for service attributes mentioned in prior work. It is important to note that the service attributes mentioned in prior work are not necessarily mutually exclusive. For example, the service attribute dignity and respect is very narrow, whereas communication is rather broad and likely encompasses dignity, respect, and other factors. Thus, the lead researcher was first tasked with identifying all service attributes mentioned; if substantial overlap existed between service attributes, or if some broad attributes contained narrower sub-attributes, then these issues could be reconciled in ensuing analyses. This analysis identified seven key service attributes to classify the online reviews along core dimensions. These were waiting time, incident type, treatment outcome, communication, hospital environment, diagnosis accuracy, and cost. Each of these attributes is acknowledged in the literature (Dobrzykowski, Callaway, & Vonderembse, 2015; Hawkins et al., 2015; Wagland et al., 2015; Hao & Zhang, 2016; Zhang et al., 2018). Communication was the only service attribute that contained subattributes, as we also observed instances of interpersonal manner, staff interaction, bedside manner, and dignity and respect. As these service attributes overlapped so substantially, we aggregated them into the communication service attribute. Interestingly, the analysis did not identify any prevalence of service attributes beyond this seven. For example, while analysis of pain management has been prevalent in the literature (Bardach et al., 2012, 2015; Hawkins et al., 2015), the lead researcher did not observe discussions of pain management in our dataset.

Data Coding

We randomly distributed the reviews to 68 coders, who were undergraduate students studying business at a major public university in the United States. We provided the coders with a detailed tagging protocol (see Appendix A) describing how they should label each review. A total of 61 of 68 coders tagged between 100 and 110 reviews; the remaining seven coders tagged ten or fewer reviews. Overall, we obtained 6,440 unique tags (i.e., 97.9% of total collected reviews). The lead researcher coded a further 300 random reviews for comparison to the students' tags (Abrahams et al., 2015; Goldberg & Abrahams, 2018). For each review, the coder determined whether a specific service attribute value was mentioned in the hospital review. If no code value was identified, the review was tagged as "not mentioned." We used Cohen's kappa (κ) to determine inter-rater reliability for each code value of each service attribute (specific values are detailed in Appendix B). The procedure we followed to calculate κ was to compare the judgments of Coder A versus Coder B as follows.

- Coder A was designated as the authority coder (the lead member of the research team).
- Coder B was the conservative combined opinion of all other coders. Though voting is often used to determine a final decision in the case of conflicting opinions, the cost of a false negative for concerns of service quality is high. Thus, we employed a conservative strategy, judging a service concern to exist whenever any coder indicated that a concern relating to that service attribute was mentioned (Goldberg & Abrahams, 2018).

We found that κ was consistently greater than 0.4, indicating "moderate" or better agreement (Landis & Koch, 1977) or "fair to good" or better agreement (Fleiss, 1971) between coders.

Regression Analysis

From the 6,440 reviews tagged in the previous step, we filtered our dataset to retain the 5,747 reviews (87.6% of total collected reviews) that contain a star rating for further analysis. In our study, we found that 65.6% of reviews gave five stars; 5.4% gave four stars; 2.7% gave three stars; 3.6% gave two stars; and 22.7% gave one star. This seemingly bimodal distribution is consistent with prior works, which have observed that users with extreme experiences are most likely to post about those experiences online (Hu, Pavlou, & Zhang, 2017). We constructed a model to test whether and the extent to which each service attribute value was associated with the online star rating of hospitals. We used ordinal logistic regression to test the extent to which each service attribute value was associated with high or low star ratings.

Each independent variable contains multiple values or levels, for example, “treatment” has three values: “helpful,” “harmful,” and “not mentioned.” Star ratings, our indicator of perceptions of service quality, were the dependent variable in our model. However, prior works have noted the difficulties in utilizing star ratings as a dependent variable, namely the potential for the use of a multinomial dependent variable to yield an overfitted and poorly performing model (Fürnkranz, 2002; Galar, Fernández, Barrenechea, & Herrera, 2014; King, 2015). However, treating star ratings as an ordinal scale as opposed to multinomial models mitigates against this issue. The model was run to demonstrate the association of each service attribute value with star ratings and ascertain the relative improvement or deterioration in star ratings with increasing mentions of service attributes.

We conducted a multicollinearity test on the variables, which demonstrated that the variables had very low correlations with one another (coefficient less than 0.2), thus confirming the adequacy of internal consistency reliability for all independent variables. In addition to the service attributes, we also utilized the negative word list in a popular conventional sentiment analysis dictionary – AFINN (Nielsen, 2011) – to test the effect of negative words on the star ratings. We first define θ_j as the probability of observing less than or equal to j stars divided by the probability of observing greater than j stars. For instance, if $j = 2$, then $\theta_j = \frac{p(1\ or\ 2\ stars)}{p(3\ or\ 4\ or\ 5\ stars)}$. We define α_j as a threshold value for each star rating, allowing us to differentiate amongst each possible value of the dependent variable. We display the ordinal logistic regression equation below in (1). Most of the variables were binary variables derived from our tagging of service attributes, in which case the variable was equal to 1 if that service concern was indicated and 0 otherwise. Two exceptions are the threshold, α_j, and the negative sentiment score assessed using the AFINN dictionary.

$$\begin{aligned} ln\ (\theta_j) = a_j - {} & \beta_1 waiting\ time\ unacceptable - \beta_2 waiting\ time\ acceptable \\ & -\beta_3 incidenttype\ birth - \beta_4 incident\ type\ emergency - \beta_5 incident\ type\ surgery \\ & -\beta_6 incident\ typeurgent - \beta_7 treatment\ helpful - \beta_8 treatment\ harmful \\ & -\beta_9 communication\ good - \beta_{10} communication\ bad \\ & -\beta_{11} environment\ dirty/unsanitary - \beta_{12} environmentclean/sanitary \\ & -\beta_{13} diagnosis\ correctly\ diagnosed - \beta_{14} diagnosis\ incorrectlydiagnosed \\ & -\beta_{15} diagnosis\ undetermined - \beta_{16} cost\ too\ expensive - \beta_{17} AFINN\ negative \end{aligned} \tag{1}$$

Smoke Term Generation

To identify the key words and phrases associated with each service attribute, we created a *training* set composed of a random 80% of the reviews (4,597) from the full set of 5,747 reviews. We computed the relative prevalence of each term (word or phrase) for each service attribute value using the correlation coefficient (CC) score (Fan et al., 2005) as the specific metric. This technique allowed us to generate lists of single words (unigrams) and two and three word phrases (bigrams and trigrams) specific to each service attribute using the process described in previous studies (Abrahams et al., 2012, 2015). We manually curated lists of distinctive or "smoke" terms for each service attribute value to improve generality; consistent with the prior studies (Abrahams et al., 2012; Winkler, Abrahams, Gruss, & Ehsani, 2016), we removed common English words (e.g., "a," "in," "for," "but," "with," "they," etc.). Appendix C contains the top 30 manually curated terms most prevalent in each service attribute.

Assessing Smoke Terms

We retained a *validation* set containing the remaining 20% of the random 5,747 reviews (1,150). This set of random reviews allowed us to conduct an objective assessment of the performance of automated discovery of each service attribute. We "scored" the validation set by using the term lists of each service attribute value; total "smoke scores" were computed as the sum of the frequency of each smoke term in a review multiplied by that term's prevalence score (Goldberg & Abrahams, 2018). After computing these scores for each review, we ranked all of the reviews in the validation set from highest score to lowest score, where higher scores indicate greater likelihoods of poor service quality. In order to assess the performance of each smoke scoring method, we computed the area under curve (AUC) statistic. Values of AUC that are close to 1 specify the best models, rarely containing false positives in the top-scored items and rarely containing false negatives in the bottom-scored items.

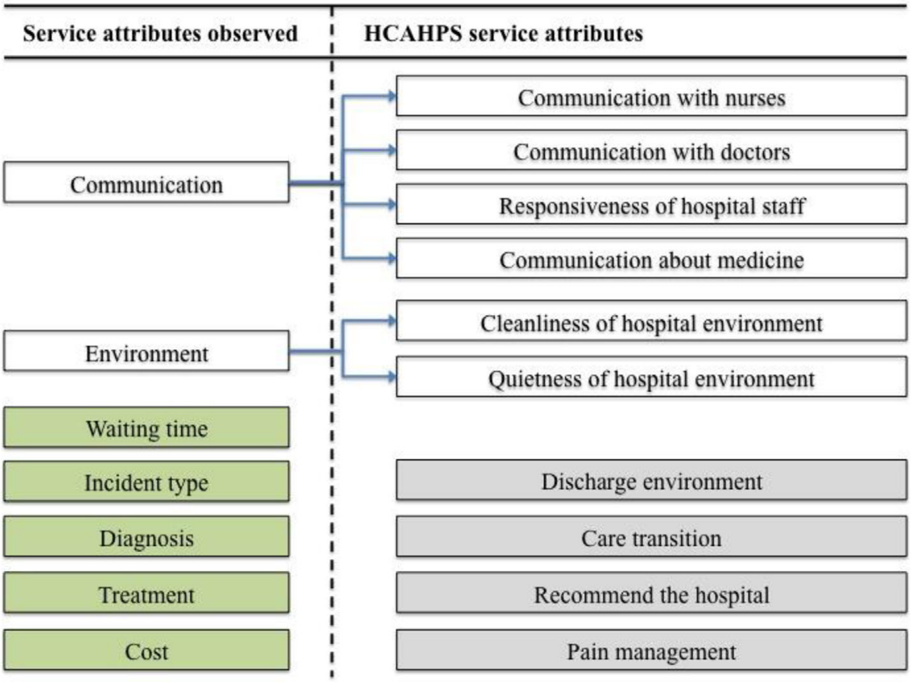
RESULTS AND DISCUSSION

Exploratory Analysis

Addressing RQ1, our initial screening of 300 random Facebook hospital reviews revealed seven major service attributes: communication (good, bad, or not mentioned); environment (clean/sanitary, dirty/unsanitary, or not mentioned); waiting time (acceptable, unacceptable, or not mentioned); incident type (urgent, emergency, surgery, birth, or not mentioned); diagnosis (correct, misdiagnosed, undetermined, or not mentioned); treatment (helpful, harmful, or not mentioned); and cost (too expensive or satisfactory/not mentioned). Thus, H1 is partially supported, as communication, environment, waiting time, and cost were hypothesized, but incident type, diagnosis, and treatment were not hypothesized. Additionally, while technical competence was hypothesized, we did not observe mentions of it in this screening. The observed service attributes overlap somewhat with other studies (see Table 1), although this specific combination of service attributes appears unique to our study. Interestingly, our service attributes partially map to the HCAHPS survey, a conventional intelligence-gathering tool (see Figure 2). We observed communication and environment as two of our service categories, and these categories overlap with a range of HCAHPS service attributes. The HCAHPS survey does not cover waiting time,[1] incident type diagnosis, treatment, or cost. This study did not cover a range of the survey items (shaded in grey) such as discharge information, care transition, recommend the hospital, and pain management.

[1] Although HCAHPS does not consider waiting time, many hospitals track emergency room waiting time as a component of other measures such as National Emergency Department Overcrowding Scale (NEDOCS) (Davis, 2018).

Figure 2: Mapping of service attributes to HCAHPS.


Service attributes observed
HCAHPS service attributes
Communication
Environment
Waiting time
Incident type
Diagnosis
Treatment
Cost
Communication with nurses
Communication with doctors
Responsiveness of hospital staff
Communication about medicine
Cleanliness of hospital environment
Quietness of hospital environment
Discharge environment
Care transition
Recommend the hospital
Pain management

Tagging Analysis

Next, our team of 68 undergraduate students assessed a large dataset of 5,747 reviews using the tagging protocol described in Appendix A. Furthering our analysis of RQ1, this step allows us to assess the prevalence of each service attribute value in our dataset; these statistics are reported in Table 2. Our service attributes were mentioned very frequently in these reviews. The reviews mentioned 1.72 service attributes on average. 77.92% of reviews mentioned at least one service attribute, and 50.27% mentioned multiple service attributes. Most of the relationships between service attribute values and star ratings are intuitive, but several of the effects are substantial and noteworthy. For example, waiting time was noteworthy, as patients mentioning acceptable waiting times were more than 10 times more likely to post a high-star review, and patients mentioning unacceptable waiting times were more than 6 times more likely to post a low-star review. Additionally, regardless of the type of incident that the patient reported, patients that specifically mentioned some incident were more likely to post a high-star review. Generally, these patients may have come to the hospital with serious health concerns that they were glad to resolve. Communication seemed to be of particular importance to patients, as 60.23% of reviews mentioned communication. Patients that mentioned good communication were over 22 times more likely to post a high-star review,

Table 2: Prevalence of service attribute values.

		Prevalence (percentage of all reviews)		
Service attribute	Service attribute value	Low star ratings	High star ratings	All star ratings
Waiting time	Acceptable	0.28%	2.82%	3.10%
Waiting time	Unacceptable	7.81%	1.17%	8.98%
Waiting time	Not mentioned	20.95%	66.97%	87.92%
Incident type	Urgent	2.61%	3.90%	6.51%
Incident type	Emergency	4.87%	5.19%	10.06%
Incident type	Surgery	2.12%	8.44%	10.56%
Incident type	Birth	0.96%	4.82%	5.78%
Incident type	Not mentioned	18.48%	48.62%	67.10%
Diagnosis	Correct	1.44%	12.22%	13.66%
Diagnosis	Misdiagnosed	2.51%	0.30%	2.80%
Diagnosis	Undetermined	1.18%	0.42%	1.60%
Diagnosis	Not mentioned	23.91%	58.03%	81.94%
Treatment	Helpful	1.74%	30.36%	32.10%
Treatment	Harmful	6.37%	0.38%	6.75%
Treatment	Not mentioned	20.93%	40.21%	61.14%
Communication	Good	1.84%	41.22%	43.07%
Communication	Bad	15.97%	1.18%	17.16%
Communication	Not mentioned	11.22%	28.55%	39.78%
Environment	Clean/sanitary	0.44%	4.18%	4.61%
Environment	Dirty/unsanitary	2.02%	0.54%	2.56%
Environment	Not mentioned	26.59%	66.24%	92.83%
Cost	Too expensive	2.30%	0.78%	3.08%
Cost	Not mentioned/satisfactory	26.74%	70.18%	96.92%

Table 3: Tukey-Kramer analysis of AFINN sentiment scores at each star rating.

Star rating	Connecting letters report					Mean AFINN sentiment score
5	A					6.35
4		B				4.46
3			C			2.60
2				D		−0.04
1					E	−2.30

while patients that mentioned bad communication were over 13 times more likely to post a low-star review.

We found that traditional sentiment analysis (here measured using the AFINN negative sentiment dictionary (Nielsen, 2011)) was strongly associated with overall star ratings, yielding a correlation of 0.54. As a further analysis of star ratings, we also performed an ANOVA to examine the extent to which reviews of certain star ratings differed via pairwise comparisons. A Tukey-Kramer test revealed that the average AFINN sentiment score for reviews of a given star rating differed from the average AFINN sentiment score for reviews of all other star ratings. We display this analysis in Table 3 below. As expected, five-star reviews were associated with the most positive sentiment on average, followed by four-star, three-star, two-star, and one-star reviews.

Regression Analysis

Addressing RQ2, we performed an ordinal logistic regression analysis to examine relationships between independent variables (i.e., service attribute values and sentiment values) and overall star ratings. The results of this regression analysis are displayed in Table 4. Variables whose regression coefficients significantly differed from zero at the 0.05 level are bolded in Table 4.

Our model had an R^2 value of 0.38, indicating that the model explained 38% of the variance in the star ratings. We also assessed model fit using area under the curve (AUC), where values closer to 1 indicate better fitting models, and values closer to 0 indicate worse fitting models. Treating a star rating of five as the base level, we observe AUC values of 0.90, 0.95, 0.95, and 0.94 for four-star, three-star, two-star, and one-star ratings, indicating an extremely high degree of fit.[2] The ordinal logistic regression coefficients provide odds ratios, for example, the odds of being associated with a five-star review as opposed to a non-five-star review. For instance, a hospital review with the waiting time indicated as acceptable suggests that the review is 2.68 times as likely to be associated with a five-star rating (i.e., $e^{2.68}$). Even when controlling for the other service attributes, communication again seemed to be the variable with the greatest effect on overall star ratings, as patients mentioning good communication were over four times as likely to give the hospital a five-star rating than a non-five-star rating. Mentions of helpful treatments

[2] All variance inflation factor (VIF) scores were below 5, indicating that multicollinearity was not a substantial concern for our model. In fact, all VIF scores were below 3 except for "diagnosis undetermined" (3.38).

Table 4: Ordinal logistic regression analysis output.

Variable	β coefficient	e^{β}	Standard error	X^2 value	p-value
Waiting time unacceptable	**−1.23**	**0.29**	**0.13**	**95.68**	**<0.01**
Waiting time acceptable	**0.99**	**2.68**	**0.20**	**25.60**	**<0.01**
Incident type birth	0.24	1.27	0.15	2.60	0.11
Incident type emergency	−0.20	0.82	0.12	2.84	0.09
Incident type surgery	−0.08	0.92	0.12	0.40	0.53
Incident type urgent	0.13	1.13	0.14	0.81	0.37
Treatment helpful	**1.20**	**3.32**	**0.09**	**189.71**	**<0.01**
Treatment harmful	**−1.30**	**0.27**	**0.12**	**126.54**	**<0.01**
Communication good	**1.40**	**4.08**	**0.06**	**526.46**	**<0.01**
Communication bad	**−1.64**	**0.19**	**0.07**	**563.92**	**<0.01**
Environment dirty/unsanitary	**−0.38**	**0.68**	**0.16**	**5.65**	**0.02**
Environment clean/sanitary	0.26	1.29	0.15	2.83	0.09
Diagnosis correctly diagnosed	**0.38**	**1.46**	**0.14**	**6.96**	**<0.01**
Diagnosis misdiagnosed	−0.28	0.76	0.20	1.92	0.17
Diagnosis undetermined	0.03	1.03	0.22	0.02	0.88
Cost too expensive	**−0.71**	**0.49**	**0.10**	**50.42**	**<0.01**
AFINN negative	**−0.27**	**0.76**	**0.01**	**343.55**	**<0.01**

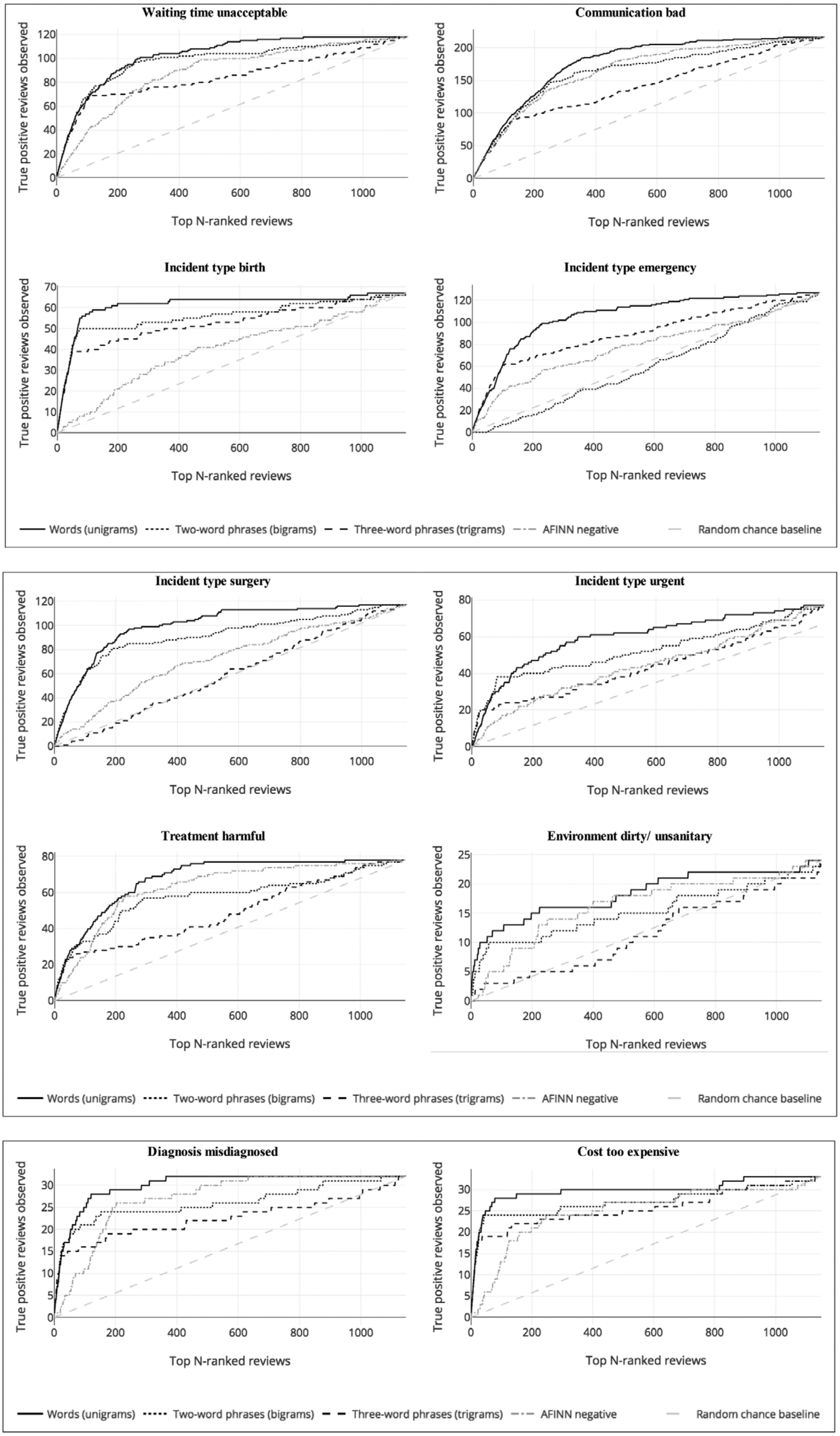
and correct diagnoses were also strongly associated with high star ratings. Interestingly, incident types did not seem to be strongly associated with star ratings, as all the values for these service attributes were statistically insignificant. These findings are generally consistent with H2, which suggested that positive mentions of service attributes would be positively associated with star ratings, while negative mentions of service attributes would be negatively associated with star ratings. The direction of the observed relationships is consistent with this hypothesis, although in some cases (e.g., environment clean/sanitary), the coefficient was not statistically significant. Our study adds to Hawkins et al. (2015)'s findings by demonstrating that specific service attributes that were mentioned in hospital-related Facebook posts were significantly associated with users' star ratings. Our study demonstrates that the text of online reviews provides critical content for uncovering the service attributes most associated with star ratings.

Text Analytic Analysis

As a final step in our analysis and to address RQ3, we generated specific terms predictive of each service attribute value. In doing so, we provide a means for hospital administrators to rapidly filter through online feedback and prioritize those reviews pertaining to particular types of feedback. As past studies (Greaves et al., 2013; Paul, Wallace, & Dredze, 2013; Wallace et al., 2014; Jung, Hur, Jung, & Kim, 2015; Rastegar-Mojarad et al., 2015; Doing-Harris et al., 2016; Hao & Zhang, 2016) mostly focused on single words, or unigrams, we examined the extent to which our terms overlapped with these prior studies. The top 30 most prevalent terms for each service attribute value had very little overlap with each other across each of the service attributes (see Appendix C). Additionally, the most prevalent terms for each service attribute value overlapping with those from prior studies were generally sentiment words. Contrary to the past studies, our study discovered more non-emotive than emotive terms for each service attribute value. Additionally, generating multi-word phrases reveals more details as to the content of specific reviews. For example, we found trigrams such as "for a bed" and "in the hallway" corresponding to unacceptable waiting times, whereas prior studies found only unigrams such as "hours," "waiting," "wait," "hour," etc. These trigrams provide specificity to forms of patient complaints, such as what the patients were waiting for (waiting to see a doctor versus waiting for a hospital bed) and where they were waiting (in a waiting room, reception area, or hallway).

We also used the validation set to perform an assessment of the performance of these terms in rapidly filtering reviews pertaining to particular service attribute values. Using the "smoke scores" generated for each review in the validation set, we ranked the reviews in the validation set from most likely (highest score) to least likely (lowest score) to refer to each service attribute value. In Figure 3, we provide lift charts showing the performance of the hospital-specific terms as opposed to generic sentiment analysis. These lift charts show each method's performance at detecting true positive reviews for each service attribute value (i.e., the number of true positives detected in the top *N*-ranked reviews). Unigrams, bigrams, and trigrams are compared to the AFINN negative sentiment dictionary and a random chance baseline. Methods showing quicker vertical ascent on the y-axis are better

Figure 3: Lift charts for each service attribute.



performing. We also show AUC values in Table 5, which confirm the utility of these methods relative to the AFINN negative sentiment dictionary. In Table 6, we explore the manner in which performance varies as a function of the sample size. Thus, we performed an analysis on smoke term generation for “waiting time unacceptable” in which we varied the training set size. In each trial, we created a balanced training set in which half of the reviews pertained to unacceptable waiting times. We retained the top 200 unigrams (removing stop words), bigrams, and trigrams for evaluation purposes, although we also note the growth in the number of candidate smoke terms at each training set size. Our results indicate that the text analytic techniques perform well even at small sample sizes. However, tagging a greater number of reviews has potential to improve performance further as the size and coverage of the lexicon increases, particularly for longer *n*-grams. In our study, our coders contributed a total of 69.87 person-hours and generated 6,791 tags. They worked at a rate of 97.20 reviews per hour on average with a standard deviation of 32.79 reviews per hour.

In each chart, it is clear from the degree of “lift” observed, or the bump in each curve, that the specialized text analytic methods typically outperform both sentiment analysis (AFINN negative) and the random chance baseline, and thus H3 is supported. In each case, the unigram method provides the greatest AUC (see Table 5), although the bigram and trigram methods also frequently outperformed the baseline methods as well. While most of these charts illustrate similar trends, we did note a few unusual results, such as the poor performances of the “incident type emergency” bigrams and “environment dirty/unsanitary” trigrams. A possible reason for the poor performance is that, due to the specificity of bigrams and trigrams relative to unigrams, overfitting occurs more frequently. Fortunately, in each case in which one method performed poorly, another method was observed to perform quite well.

Interestingly, although bigrams and trigrams typically performed worse than unigrams, they often offered similar levels of performance over a top-ranking set of reviews. However, due to the specificity of these terms, matching reviews were exhausted, and the remainder of reviews were classified at the rate of random chance. However, there is still some utility in these methods because the specificity of these terms may be a useful diagnostic tool, responding to the call by Greaves et al. (2013). For example, consider the unigrams generated versus the trigrams generated for “diagnosis misdiagnosed.” While unigrams such as “wrong” or “misdiagnosed” may reveal reviews concerning incorrect diagnoses quickly, they do not detail the nature of the misdiagnosis. Although trigrams may not yield as many matches, terms such as “a stroke and” or “the surgery after” may be more illuminating as to the nature of the diagnosis. Thus, although unigrams may provide superior performance, bigrams and trigrams may offer enhanced interpretability. The decision-maker may determine whether they prefer rapid sorting (unigrams) or more detailed insights (bigrams/trigrams).

CONCLUSIONS AND IMPLICATIONS

Our results demonstrated a particular set of service attributes present in Facebook reviews of hospitals. Although these service attributes overlap somewhat with

Table 5: Comparison of area under the curve (AUC) values across methods and service attribute values.

Service attribute	Service attribute value	Unigrams	Bigrams	Trigrams	AFINN negative
Waiting time	Unacceptable	0.87	0.83	0.72	0.75
Incident type	Emergency	0.84	0.47	0.71	0.62
Incident type	Birth	0.92	0.84	0.79	0.60
Incident type	Surgery	0.86	0.78	0.51	0.62
Incident type	Urgent	0.78	0.68	0.57	0.57
Treatment	Harmful	0.87	0.74	0.63	0.81
Communication	Bad	0.82	0.77	0.66	0.77
Environment	Dirty	0.77	0.64	0.48	0.69
Diagnosis	Misdiagnosed	0.94	0.82	0.71	0.85
Cost	Too expensive	0.90	0.82	0.78	0.76

Table 6: Area under the curve (AUC) scores for "waiting time unacceptable" across training set sizes.

	AUC (count of unique candidate terms in training set)		
Training set size (count of positives)	Unigrams	Bigrams	Trigrams
100 (*50*)	0.84 (*1,285*)	0.79 (*3,850*)	0.63 (*4,508*)
200 (*100*)	0.85 (*1,881*)	0.83 (*6,921*)	0.66 (*8,786*)
300 (*150*)	0.86 (*2,457*)	0.83 (*9,938*)	0.66 (*13,211*)
400 (*200*)	0.86 (*2,873*)	0.83 (*12,526*)	0.71 (*17,259*)
500 (*250*)	0.86 (*3,247*)	0.83 (*15,065*)	0.72 (*21,460*)
600 (*300*)	0.86 (*3,633*)	0.83 (*17,479*)	0.72 (*25,558*)

traditional metrics such as HCAHPS, they also offer a new perspective as to patients' views on service quality. These service attributes span the reviews very well, as 77.92% of the reviews mentioned at least one of the service attributes. The specific mix of service attributes detected is unique to our paper and helps to unify the variety of service attributes studied in past works. Importantly, we showed that not all service attribute values affected overall ratings equally. In particular, communication, waiting time, and the helpfulness of treatments seemed to be highly associated with patients' star ratings. We also generated prevalent key terms for service attribute values, which can assist health care policy makers and health care providers to rapidly monitor specific service concerns and adjust their policies or resources to better serve their patients. This study highlights that these terms derived from service attributes are actionable and informative compared to the nonspecific emotion-focused terms supplied by traditional sentiment analysis.

Our study is subject to several limitations. First, undergraduate students tagged reviews manually in this study. When considering a complex medical case, the students may overlook reported service issues; for instance, some students may misinterpret reviews that actually discuss a medical error such as misdiagnosis by unknowingly claiming these reviews as only symptoms rather than tagging them as service concerns. However, the high degree of agreement with the authority tagger suggests that the tags are of generally high quality. Relatedly, the process of generating smoke terms is expensive in that it requires substantial labor as well as computational efforts. Although we did not pursue this extension, future work might examine the use of artificial intelligence to build a broad lexicon from a seed (Asghar, Ahmad, Qasim, Zahra, & Kundi, 2016). Another possible limitation is the choice of sampling reviews from the top 100-listed hospitals in the United States. Hospitals outside the top 100-listed may share different performance characteristics. We cannot guarantee that our findings would apply to all hospitals. However, the top 100 hospitals, which this study focuses on, likely share many similar service and performance characteristics with a great variety of other hospitals. A final possible limitation is that we employed a tagging protocol at the review-level.

That is, we asked the student taggers to indicate whether an entire review referred to a service attribute as opposed to the sentence-level or phrase-level. We expect that a more granular approach might reduce false positives, but it would also require an enormous tagging effort to address the same number of reviews with such specificity. Therefore, analysis at the review-level is quite common in the literature (Abrahams et al., 2015; Goldberg & Abrahams, 2018; Mummalaneni et al., 2018).

Gathering data from patients' online reviews implies that patients have self-selected their participation in online discussions. As such, it is crucial for health care administrators and policy makers to be aware that the findings are unlikely to be fully representative of the population of the hospital service concerns. Instead, this discovery of specific service attributes should be viewed as supplementary to the conventional data collecting exercises (e.g., HCAHPS, patient surveys, etc.). The rapid discovery and investigation of the specific service attributes in this study is unique, and it would otherwise have been difficult for practitioners to analyze voluminous real-time (unstructured) data.

Future work may also look to categorize service-specific categories into leading, in-delivery, and trailing. In our study, we focused mostly on leading and in-delivery service attributes. A leading service attribute is one that occurs before the service occurs, while an in-delivery service attribute occurs while the service is occurring (e.g., doctor is talking with a patient). A trailing service attribute might be the medical outcome (e.g., successful treatment versus unsuccessful treatment). The following questions could be addressed in future studies: would a patient that provides an initial low online rating feel the same way if one month later the treatment were to be deemed a success? In contrast, if a patient initially provided a high online rating but later experienced a serious side effect, would that be conducive to a poor rating? Gathering a rating before and after the service could help to assess these questions.

Future research may extend this text classification technique to web-based physician reviews (e.g., RateMDs, HealthGrades.com, etc.) for all specialty areas to provide health care providers and researchers with better insight into consumers' thoughts regarding different medical specialties. Additionally, these techniques may also have value in considering online consumer reviews of other products/services (Abrahams et al., 2015; Goldberg & Abrahams, 2018). It would also be interesting to further investigate the relationship between web-based ratings of hospitals and with traditional types of hospital services acquired from the conventional survey results (e.g., HCAHPS).

Equipped with a stronger understanding of specific service attributes that influence online star ratings, managers will be better able to focus on key words and phrases of the service attributes that can improve the patient experience and the overall reputation of the hospital. If the online rating is strongly associated with a specific service attribute, then this identifies where administrators should focus to enhance the patient experience as opposed to other aspects with a lesser influence. In our study, communication, waiting times, and perceptions of treatment helpfulness seemed to be especially important factors. Before making policy changes, administrators should not only target the influential factor, but they should also focus on the cost of targeting each factor to make a final decision as to the most effective policy to implement.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

APPENDICES

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.

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems*, *54*(1), 87–97.

Adams, D. Z., Gruss, R., & Abrahams, A. S. (2017). Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International Journal of Medical Informatics*, *100*, 108–120.

Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R., & Kundi, F. M. (2016). SentiHealth: Creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, *5*(1), 1–23.

Bardach, N. S., Asteria-Peñaloza, R., Boscardin, W. J., & Dudley, R. A. (2012). The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ Quality & Safety*, *22*(3), 194–202.

Bardach, N. S., Lyndon, A., Asteria-Peñaloza, R., Goldman, L. E., Lin, G. A., & Dudley, R. A. (2015). From the closest observers of patient care: A thematic analysis of online narrative reviews of hospitals. *BMJ Quality & Safety*, *25*(11), 889–897.

BrightLocal (2016). Local Consumer Review Survey 2016.

Brody, S., & Elhadad, N. (2010). Detecting salient aspects in online reviews of health providers. *AMIA Annual Symposium Proceedings*, 202–206.

Campbell, L., & Li, Y. (2017). Are Facebook user ratings associated with hospital cost, quality and patient satisfaction? A cross-sectional analysis of hospitals in New York State. *BMJ Quality & Safety*, *27*(2), 119–129.

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

Davis, Z. E. (2018). *Toward a healthcare services ecosystem*, Virginia Tech.

Devarakonda, M. V., Mehta, N., Tsou, C.-H., Liang, J. J., Nowacki, A. S., & Jelovsek, J. E. (2017). Automated problem list generation and physicians perspective from a pilot study. *International Journal of Medical Informatics*, *105*, 121–129.

Dobrzykowski, D. D., Callaway, S. K., & Vonderembse, M. A. (2015). Examining pathways from innovation orientation to patient satisfaction: A relational view of healthcare delivery. *Decision Sciences*, *46*(5), 863–899.

Doing-Harris, K., Mowery, D. L., Daniels, C., Chapman, W. W., & Conway, M. (2016). Understanding patient satisfaction with received healthcare services: A natural language processing approach. *AMIA Annual Symposium Proceedings*, 524–533.

Fan, W., Gordon, M. D., & Pathak, P. (2005). Effective profiling of consumer information retrieval needs: A unified framework and empirical comparison. *Decision Support Systems*, *40*(2), 213–233.

Farhadloo, M., Patterson, R. A., & Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems*, *90*, 1–11.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Fürnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*, *2*(Mar), 721–747.

Galar, M., Fernández, A., Barrenechea, E., & Herrera, F. (2014). Empowering difficult classes with a similarity-based aggregation in multi-class classification problems. *Information Sciences*, *264*, 135–157.

Glover, M., Khalilzadeh, O., Choy, G., Prabhakar, A. M., Pandharipande, P. V., & Gazelle, G. S. (2015). Hospital evaluations by social media: A comparative analysis of Facebook ratings among performance outliers. *Journal of General Internal Medicine*, *30*(10), 1440–1446.

Goldberg, D. M., & Abrahams, A. S. (2018). A Tabu search heuristic for smoke term curation in safety defect discovery. *Decision Support Systems*, *105*, 52–65.

Greaves, F., Laverty, A. A., Cano, D. R., Moilanen, K., Pulman, S., Darzi, A., & Millett, C. (2014). Tweets about hospital quality: A mixed methods study. *BMJ Quality & Safety*, *23*(10), 838–846.

Greaves, F., Pape, U. J., King, D., Darzi, A., Majeed, A., Wachter, R. M., & Millett, C. (2012). Associations between Web-based patient ratings and objective measures of hospital quality. *Archives of Internal Medicine*, *172*(5), 435–436.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, *15*(11), e239.

Griffiths, A., & Leaver, M. P. (2018). Wisdom of patients: Predicting the quality of care using aggregated patient feedback. *BMJ Quality & Safety*, *27*(2), 110–118.

Hao, H., & Zhang, K. (2016). The voice of Chinese health consumers: A text mining approach to web-based physician reviews. *Journal of Medical Internet Research*, *18*(5), e108.

Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., … Bourgeois, F. T. (2015). Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Quality & Safety*, *25*(6), 404–413.

Hu, N., Koh, N. S., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision Support Systems*, *57*, 42–53.

Hu, N., Pavlou, P. A., & Zhang, J. J. (2017). On self-selection biases in online product reviews. *MIS Quarterly*, *41*(2), 449–471.

James, T. L., Calderon, E. D. V., & Cook, D. F. (2017). Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Systems with Applications*, *71*, 479–492.

Jonsson, P., Johansson, M., Ancarani, A., Di Mauro, C., & Giammanco, M. D. (2011). Patient satisfaction, managers' climate orientation and organizational climate. *International Journal of Operations and Production Management*, *31*(3), 224–250.

Jung, Y., Hur, C., Jung, D., & Kim, M. (2015). Identifying key hospital service quality factors in online health communities. *Journal of Medical Internet Research*, *17*(4), e90.

King, M. A. (2015). *Ensemble learning techniques for structured and unstructured data*, Virginia Tech.

Lagu, T., Goff, S. L., Craft, B., Calcasola, S., Benjamin, E. M., Priya, A., & Lindenauer, P. K. (2016). Can social media be used as a hospital quality improvement tool? *Journal of Hospital Medicine*, *11*(1), 52–55.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

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.

Mummalaneni, V., Gruss, R., Goldberg, D. M., Ehsani, J. P., & Abrahams, A. S. (2018). Social media analytics for quality surveillance and safety hazard detection in baby cribs. *Safety Science*, *104*, 260–268.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*,

Pai, R. R., & Alathur, S. (2018). Assessing mobile health applications with twitter analytics. *International Journal of Medical Informatics*, *113*, 72–84.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, *49*(4), 41–50.

Paul, M. J., Wallace, B. C., & Dredze, M. (2013). What affects patient (dis) satisfaction? Analyzing online doctor ratings with a joint topic-sentiment model. *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI*. 53–58.

Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., ..., Merchant, R. M. (2016). What can Yelp teach us about measuring hospital quality? *Health Affairs (Project Hope)*, *35*(4), 697.

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

Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., & Rosa, T. C. (2016). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, *85*(1), 80–95.

Rosen, P. (2016). The patient as consumer and the measurement of bedside manner. *NEJM Catalyst*, *10*, 332.

Rozenblum, R., Bates, D. W. (2013). Patient-centred healthcare, social media and the internet: The perfect storm? *BMJ Quality & Safety*, *22*(3), 183–186.

Rozenblum, R., Greaves, F., & Bates, D. W. (2017). The role of social media around patient experience and engagement. *BMJ Quality & Safety*, *26*(10), 845–848.

Thirumalai, S., & Sinha, K. K. (2009). Customization strategies in electronic retailing: Implications of customer purchase behavior. *Decision Sciences*, *40*(1), 5–36.

Wagland, R., Richardson, A., Armes, J., Hankins, M., Lennan, E., & Griffiths, P. (2015). Treatment-related problems experienced by cancer patients undergoing chemotherapy: A scoping review. *European Journal of Cancer Care*, *24*(5), 605–617.

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.

Winkler, M., Abrahams, A. S., Gruss, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision Support Systems*, *90*, 23–32.

Wolf, J. R., & Muhanna, W. A. (2011). Feedback mechanisms, judgment bias, and trust formation in online auctions. *Decision Sciences*, *42*(1), 43–68.

Yang, M., Kiang, M., & Shang, W. (2015). Filtering big data from social media–Building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics*, *54*, 230–240.

Yang, P.-C., Lee, W.-C., Liu, H.-Y., Shih, M.-J., Chen, T.-J., Chou, L.-F., & Hwang, S.-J. (2018). Use of Facebook by hospitals in Taiwan: A nationwide survey. *International Journal of Environmental Research and Public Health*, *15*(6), 1188.

Zhang, W., Deng, Z., Hong, Z., Evans, R., Ma, J., & Zhang, H. (2018). Unhappy patients are not alike: Content analysis of the negative comments from China's good doctor website. *Journal of Medical Internet Research*, *20*(1), e35.

Nohel Zaman is Assistant Professor of Information Systems and Business Analytics in the College of Business Administration at Loyola Marymount University, Los Angeles. His research interests delve in the field of text analytics for enhancing consumer product quality and hospital services. He received his PhD in Business Information Technology from Virginia Tech. He earned a BS in Business Administration and MS in Economics from University of Texas at Dallas and an MS in Computational Science and Engineering from North Carolina A&T State University.

David M. Goldberg is Assistant Professor of Management Information Systems in the Fowler College of Business at San Diego State University. He received his doctoral and bachelor's degrees from Virginia Tech. His current research interests are in the areas of text mining, machine learning, decision support systems, and expert systems. He has published in *Decision Support Systems*, *Information Technology & Management*, *Safety Science*, *International Journal of Information Technology & Decision Making*, and others.

Alan S. Abrahams is Associate Professor of Business Information Technology in the Pamplin College of Business at Virginia Tech. Dr. Abrahams's primary research interest is in developing text analytic techniques for quality improvement. Dr Abrahams is a Senior Editor at *Decision Support Systems*. He has published over 35 peer-reviewed articles in diverse journals, including *Production & Operations Management*, *Decision Support Systems*, *Safety Science*, and others. He holds a PhD in Computer Science from the University of Cambridge and a Bachelor of Business Science degree with honors in Information Systems from the University of Cape Town.

Richard Essig is a fourth year Marketing PhD candidate in the Pamplin College of Business at Virginia Tech in the consumer behavior tract. Richard's research focuses on product ratings, persuasion, and how firm size impacts consumers' manufacturing perceptions and purchasing behavior. Richard holds an MBA from the Fuqua School of Business at Duke University, an MS in Finance from the University of Delaware, and a BS in Biological Sciences from West Chester University of Pennsylvania.