



2017

Perfect ratings with negative comments: Learning from contradictory patient survey responses

Andrew S. Gallan

DePaul University, agallan@depaul.edu


Marina Girju

DePaul University, MGIRJU@depaul.edu

Roxana Girju

University of Illinois at Urbana-Champaign, girju@illinois.edu

Follow this and additional works at: <http://pxjournal.org/journal>

 Part of the [Applied Statistics Commons](#), [Business Intelligence Commons](#), [Marketing Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

Recommended Citation

Gallan, Andrew S.; Girju, Marina; and Girju, Roxana (2017) "Perfect ratings with negative comments: Learning from contradictory patient survey responses," *Patient Experience Journal*: Vol. 4 : Iss. 3 , Article 6.

Available at: <http://pxjournal.org/journal/vol4/iss3/6>

This Research is brought to you for free and open access by Patient Experience Journal. It has been accepted for inclusion in Patient Experience Journal by an authorized editor of Patient Experience Journal.

Perfect ratings with negative comments: Learning from contradictory patient survey responses

Cover Page Footnote

The authors have not received financial support of any kind in preparing this research. They have received support from the host organization in the form of data sets of patient survey responses.

Perfect ratings with negative comments: Learning from contradictory patient survey responses

Andrew S. Gallan, *DePaul University*, agallan@depaul.edu

Marina Girju, *DePaul University*, MGIRJU@depaul.edu

Roxana Girju, *University of Illinois Urbana-Champaign*, girju@illinois.edu

Abstract

This research explores why patients give perfect domain scores yet provide negative comments on surveys. In order to explore this phenomenon, vendor-supplied in-patient survey data from eleven different hospitals of a major U.S. health care system were utilized. The dataset included survey scores and comments from 56,900 patients, collected from January 2015 through October 2016. Of the total number of responses, 30,485 (54%) contained at least one comment. For our analysis, we use a two-step approach: a quantitative analysis on the domain scores augmented by a qualitative text analysis of patients' comments. To focus the research, we start by building a hospital recommendation model using logistic regression that predicts a patient's likelihood to recommend the hospital; we use this to further evaluate the top four most predictive domains. In these domains (personal issues, nurses, hospital room, and physicians), a significant percentage of patients who rated their experience with a perfect domain score left a comment categorized as not positive, thus giving rise to stark contrasts between survey scores and comments provided by patients. Within each domain, natural language analysis of patient comments shows that, despite providing perfect survey scores, patients have much to say to health care organizations about their experiences in the hospital. A summary of comments also shows that respondents provide negative comments on issues that are outside the survey domains. Results confirm that harvesting and analyzing comments from these patients is important, because much can be learned from their narratives. Implications for health care professionals and organizations are discussed.

Keywords

Patient experience, likelihood to recommend, patient comments, patient surveys, text analytics, natural language analysis

Introduction

Patient experience (PX) professionals are tasked with improving patient and family experiences while in medical care, and often turn to patient surveys and feedback to uncover issues or to get a sense of whether improvement is occurring.¹ PX professionals are also constantly looking to increase their organization's ability to understand patients' voices and generate actionable items that improve their patients' experiences.² While patient surveys are but one method of systematically collecting information about patient perceptions of care,³ they are nonetheless a cornerstone of measuring patient experience.⁴

Based on previous literature⁵ and insights from discussions with PX professionals, this research explores why patients give perfect top-box domain scores (only the highest rating on every item in a particular survey category), yet provide negative comments on important issues during their stay in the hospital. Thus, this paper investigates what patients are attempting to tell health care organizations even when they provide the highest scores, and suggests what can be done to address the issues that patients raise.

What was found, utilizing a large dataset from almost two years of in-patient survey data from a large hospital system in the US, is that a significant percentage of patients provide perfect domain scores only to follow up with negative comments.

Given the apparent contradiction between perfect domain scores and negative comments, and the potential magnitude of the problem, the goals of this analysis are to:

1. Understand what patients who provide positive experience scores and negative comments are trying to tell the health care organization.
2. Identify the key negatives prevalent in overall positive hospital experiences.
3. Based on findings, suggest ways to systematically harvest and understand patient comments.

This research contributes to the patient experience literature by: (a) Expanding an understanding of a quality patient experience; (b) Highlighting issues with providing patient care in a hospital even when receiving excellent experience ratings; and, (c) Understanding that PX professionals can learn more about patient experiences and

potentially identify hidden issues by paying careful attention to their comments, even when the ratings are perfect.

Patient Data

In order to explore the research questions, we utilized vendor-supplied in-patient survey data collected from eleven (11) different hospitals in a single health care system in the U.S. These data capture patients' perceptions of various aspects of their experience during a hospital stay. Through a combination of closed and open-ended questions, the study participants not only rate their experiences but also provide a variety of comments, thus giving more in-depth details on ten different domains of interest: admission, room, meals, nurses, physicians, tests/treatments, visitors/family, personal issues, discharge and overall assessment. Unlike HCAHPS data, the vendor data pairs domain-specific survey responses with domain-specific comments, and thus allows matching of each patient's domain ratings with domain comments. In each domain, patients answered anywhere between two and six questions (five-point continuous survey items – from "Very Poor" to "Very Good"), and wrote one or more comments. The dataset includes all survey responses (N = 56,900) collected from January 2015 through October 2016, containing a total of 91,281 comments on all ten domains of the hospital experience: meals, test/treatment, admission, discharge, visitor/family, personal issues, nurses, hospital room, physicians, and overall hospital experience.

Research Method

Our research involved a two-step approach. First, quantitative analyses were conducted on the structured data collected in the survey. This was to generate an understanding of which of the ten domains had the highest influence on patients' likelihood to recommend the hospital; the resulting domain ranking was used to focus and direct the second step in the research analysis. Descriptive statistics were used to summarize all ratings across all patient experience domains and patients who provided top box ratings on their experience were identified. We then marked patients who rated their likelihood to recommend as top box ("Likelihood of your recommending this hospital to others"). Finally, using binary logistic regression, we built a hospital recommendation model that analyzes the influence of all domain ratings on patients' likelihood to recommend. The dependent variable, likelihood to recommend, was set to one (1) for patients with top box rating (patients rating their likelihood to recommend as "Very Good") and zero (0) otherwise. We used patients' demographics (education, ethnicity, overall and mental health level) and seasonality (year, seasons) as control variables. The model identified drivers of patients' likelihood to recommend the hospital

at the highest level (top box) and helped rank these drivers on their relative magnitude of influence. We used these findings to guide us in the next step of our research plan. All details on this phase of the analysis are presented in Appendix 1.

In the second step in our research, we conducted text analyses on the responses to open-ended questions in the survey, namely comments describing patients' domain experiences as well as overall assessment with the hospital stay. We identified several challenges of analyzing free-form patient comments in hospital reviews and relied on established Natural Language and Linguistics research⁶ to address them:

1. **Language Understanding:** Free-form patient comments abound in shortenings, abbreviations, stylistics, incomplete sentences, and figurative language (sarcasm, metaphors, etc.). Further, partly due to the same medical settings, patients use the same vocabulary across many domains (e.g. nurses, physicians, rooms) to characterize the quality of the services they receive. Decoding and separating the true meaning of these words, phrases, or sentences is thus challenging.
2. **Language Ambiguity:** A single patient comment may refer to multiple aspects of the hospital experience, including care providers, facilities, and even family and friends. Comments may also exhibit high 'language variability' (they say the same thing in many different ways) - which can easily go beyond describing the hospital care service. In negative comments, people may complain about things that are not related to the hospital and health care.

The approach we chose is appropriate because it mitigates the risks posed by the issues raised above. Specifically, we started by looking at the distribution of comments by sentiment, as coded by the survey vendor – negative, positive, neutral and mixed comments. We then identified the negative comments of patients who gave top box domain ratings, the objective of our analysis. These comments were cleaned; stop words (e.g. prepositions, determinants, special characters) and proper names (e.g. doctors' or nurses' names) were removed because they were inconsequential to our analysis and would have decreased the accuracy in the text analysis stage. Further, in order to discover common themes across all patients, we aggregated the comments by domain (e.g. all patients' comments on nurses, all comments on room experience, etc.) and then analyzed them using Term Frequency and Inverse Document Frequency (TF-IDF), a popular information retrieval and text mining technique effective on large unstructured data.⁵ For each domain, TF-IDF generated a list of terms (alpha-numeric strings) and their relative usage frequencies. It is important to point out that, unlike other text analyses techniques that simply calculate word frequencies, TF-IDF penalizes common words that

may appear often yet have little importance (e.g. is, that, of, etc.) while it assigns higher ratings to words that are meaningful for the medical domain (e.g. nurse, blood, care, IV, etc.). We retained the top ranked 100 terms in the TF-IDF list, parsed them and uncovered several sub-clusters indicative of clear semantic categories which represent the main topics provided by patients. For more details on the techniques utilized in this phase of the research, please see Appendix 2. Finally, we used these topics to generate a deeper understanding of patients' comments and to provide suggestions for improving patients' in-hospital experiences. For each domain of care, we created word cloud maps to help us visualize patients' comments (see endnote for justification).

Results

High level descriptive analysis on the survey data shows that of the 50,900 patients participating in this study, 30,488 left at least one comment (54%). Respondents provided anywhere between one and thirteen comments across all ten domains of the health care experience, with an average of three comments per patient (std. dev. = 2.28). See Figure 1 for distribution statistics.

Of all written comments, 30.5% (27,830/91,281) are negative, 47.75% are positive, and the remaining 21.75% are mixed, neutral or not classified (Table 1). It is interesting to note that the negative comments are significantly longer than those that are positive – on average they have 105.91 characters (std. dev. = 101.05) versus 59.31 characters (std. dev. = 63.68). This is

Figure 1. Distribution of Patients by Number of Written Comments (n = 30,488 patients)

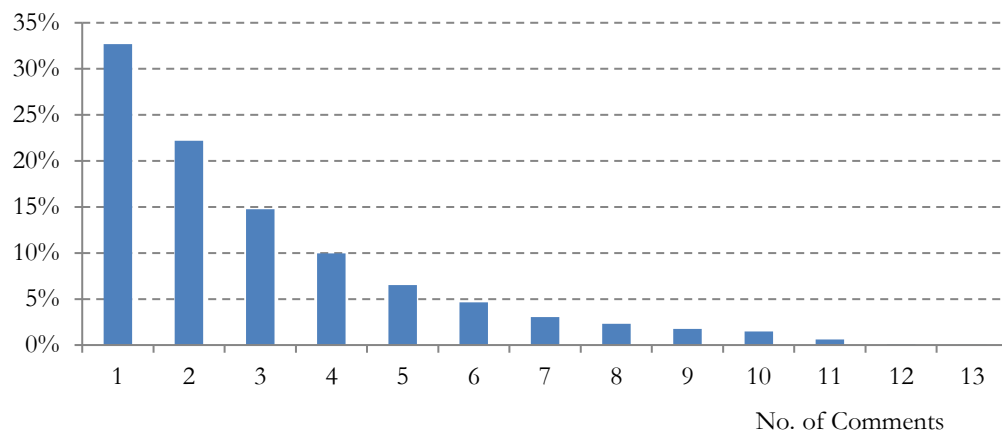


Table 1. Distribution of Comments by Length (n=91,281 comments)

Sentiment	No. of Comments	Min	Max	Mean	Std. Dev.
Positive	43586	1	2095	59.31	63.68
Negative	27830	3	2884	105.90	101.05
Other Neutral	9798	4	858	43.87	37.24
Mixed	7975	8	3839	162.35	182.94
NA	2092	2	2250	87.41	119.98

Table 2. Distribution of Ratings by Domain (n=27,458)

PX Domain	Mean	Std. Dev.
Admission	4.52	0.66
Room	4.36	0.60
Meals	4.29	0.66
Nurse	4.64	0.56
Tests/treatment	4.52	0.57
Visitor/Family	4.57	0.60
Physician	4.48	0.70
Discharge	4.44	0.66
Personal Issues	4.52	0.62

indicative of the fact that patients are more likely to describe their negative experiences in more detail than they do with positive experiences.

Patients rated their experiences across all ten domain of health care. After controlling for missing information, we find that a total of 27,458 patients had complete data that can be used in the quantitative stage of the analysis. As shown in Table 2, patients gave very high ratings on all the health care domains, where seven out of nine had averages well above 4.5 (on a 5-point scale). Patients provided the highest ratings for nurse care (average = 4.64, std. dev. = 0.56) and for having their visitors and family well treated (average = 4.57, std. dev. = 0.6). Patients were least satisfied with the hospital room and meals, as seen in their average ratings of 4.36 (std. dev. = 0.6) and 4.29 (std. dev. = 0.66), respectively.

While the magnitude of these domain ratings definitely looks very encouraging to health care professionals and administration, the high rate of negative comments (30.5%) highlights a contradiction and warrants further investigation to expose hidden issues. In order to provide targeted suggestions for improvement, we (1) identified which health domains influence patients' likelihood to recommend a hospital, and, (2) ranked the domains by the magnitude of their influence.

Quantitative Results

We built a hospital recommendation model using all the domain ratings and the likelihood to recommend the hospital. The dependent variable, Likelihood to Recommend, was the answer to the survey statement "Likelihood to Recommend this Hospital to Others," and was coded one (1) when the patient gave the highest rating of "Very Good" (roughly 30% of patients) and zero (0) otherwise. The dependent variable in this analysis is an

important outcome to all health care organizations and particularly important to the research host organization as it recently rolled out a net promoter-like score across the entire enterprise. Table 3 shows the results of the logistic regression and, through the standardized coefficients, calculates the magnitude of impact for each domain in the health experience (odds ratios).

$$Likelihood\ to\ Recommend_i = \beta_0 + \beta_1 Admission_i + \beta_2 Room_i + \beta_3 Meals_i + \beta_4 Nurse_i + \beta_5 Test_i + \beta_6 Visitor_i + \beta_7 Discharge_i + \beta_8 PersIssues_i + \beta_9 Seasonality_i + \epsilon_i,$$

where i is the patient

As shown in Table 3, all health domains have a statistically significant influence on the likelihood to recommend the hospital (all p-values are smaller than 0.05). Ideally, the hospital should focus on improving the ratings on all domains in order to get the maximum increase in likelihood to recommend. However, the results point out that some domains are more important than others, and thus have a bigger impact on increasing likelihood to recommend. The standardized coefficients in Table 3 identify personal issues (coefficient=0.42), nurse care (coefficient=0.38), room environment (coefficient=0.27), and physician care (coefficient=0.19) as the top four domains with the highest impact on likelihood to recommend. Even more revealing, the odds ratios show that small increases in patients' experiences on these top four domains would in fact double or triple their likelihood of being a promoter. For example, patients who increase their rating by one unit on the Personal Issues domain are 3.67 times more likely to give a top box recommendation to the hospital (nurses care = 3.55 more likely; room environment = 2.30 times more likely; and, physician care = 1.69 more likely). That is, if the hospital focuses on improving patients' experience vis-à-vis personal issues just by one unit on the rating scale, in return the hospital will have patients who are 3.67 times more likely to recommend it. Improvements in the other

Table 3. Hospital Recommendation Model – Likelihood to Recommend (n=27,458)

Parameter	Estimate	Pr > ChiSq	Standardized Coefficient	Odds Ratio Estimate	Odds Ratio Interval [95% Conf. Limits]	
Intercept	-21.2355	<.0001				
Admission	0.2644	<.0001	0.0940	1.303	1.220	1.391
Room	0.8344	<.0001	0.2724	2.303	2.106	2.519
Meals	0.1911	<.0001	0.0685	1.211	1.128	1.299
Nurse	1.2677	<.0001	0.3789	3.553	3.202	3.942
Test/Treatment	0.1164	0.0257	0.0353	1.123	1.014	1.245
Visitor/Family	0.2571	<.0001	0.0829	1.293	1.192	1.403
Physician	0.5221	<.0001	0.1949	1.686	1.574	1.806
Discharge	0.371	<.0001	0.1310	1.449	1.341	1.566
Personal Issues	1.3003	<.0001	0.4211	3.670	3.292	4.092

c-stat = 0.92, Percent Concordant = 0.927

Table 4. Distribution of Respondents by Top Box Ratings and Comments

	Respondents with Perfect Ratings		Patient Comments		
	No.	%	Negative %	Positive %	Other %
Personal Issues	15262	52.24%	30.11%	51.13%	18.76%
Nurse	17842	61.08%	18.25%	66.07%	15.69%
Room	9306	31.85%	52.21%	27.68%	20.12%
Physicians	16612	56.87%	20.20%	61.18%	18.62%

domains also increase recommendations but to a smaller degree. To focus our research, we continue to evaluate the top four domains as they provide health care organizations the biggest return on investment of time and effort: personal issues, nurses, hospital room, and physicians.

Qualitative Results – Patient Comments

To further understand how to improve patients' experiences in the top four domains of inpatient care, we identified respondents providing top box ratings and looked at comments they provided to better understand their ratings. As shown in Table 4, a significant proportion of patients provided perfect (all top-box) domain scores (ranged from 32% for Room to 61% for Nurses). Overall, 23% of patients (6614 respondents) gave a perfect rating on all four service domains. Upon assessing the comments they provided, it is important to note that the negative comments also make up sizeable proportions, ranging from 30% (Nurses) to 52% (Room).

To understand these results and expose topics discussed by the patients in their negative comments, we follow up with in-depth text analyses in each of the four domains of interest.

Personal issues

The TF-IDF analysis on patients' negative comments in this domain identified the most relevant words and their relative usage frequencies. We represent them visually in Figure 2.

Further parsing shows that words cluster in four semantic categories that highlight the major topics in patients' negative comments. As shown in Table 5, patients first complain about issues with supporting staff (e.g. social workers, chaplains, lactation consultants). Second, they indicate that patients do not understand their medications. Third, and to a lesser degree, they are concerned about the hospital room and meals.

Comparing this to the survey data, we find that these topics were not covered in the five questions that patients had to rate for this health domain, e.g. controlling pain, meeting emotional needs. Thus, the textual analysis helps us not only discover patients' major complaints but also provides insights into patient perceptions of personal issues beyond the survey items provided.

Figure 2: Word Cloud of Common Phrases Describing Problems for Personal Issues



Table 5: Examples of Complaint Topics about Personal Issues

Topic	Descriptive words/phrases
Supporting Staff Issues	Chaplain: anxious, not supportive, in a hurry, rude, not present, pushy Lactation consultant: late, never available, rude Social workers: absent, not helpful Photographer: pushy Pharmacy: rude, poor service
Medications	Explained too fast, hard to understand (strong accent), administered late, wrong
Room	Bad beds, no privacy, no safety (people without IDs, things stolen), crowded, desolate, unsafe, noisy
Meal	Bland, cold, late, never delivered

Figure 3: Word Cloud of Common Phrases Describing Problems for Experience with Nurses



Table 7: Examples of Complaint Topics about Experience with Hospital Room

Topic	Descriptive words/phrases
Beds	Uncomfortable, terrible, lumpy, awful, worn-out, bloody, dirty linen, not changed
Bathroom	Poor drainage, bad water pressure, cold water, leaky toilets, no riser seat, faucet not working
Room	Noisy, unsafe, too cold, not well lit, windows cracked, moldy wall, call button defective, phone/TV not working, loud roommates, too many people come in, visitors not IDed

Figure 5: Word Cloud of Common Phrases Describing Problems in Patients’ Experience with Physicians



Table 8: Complaint Topics about experience with Physicians

Topic	Descriptive words/phrases
No/Poor Communication	No communication among doctors in hospital; conflicting communication; no explaining with patient, family members, patient’s regular MD
Resident doctors	Reprimanded nurses, pharmacists, other doctors in front of patients Cocky, arrogant, pompous prick, condescending, unskilled, not knowledgeable, poor/terrible bedside manner, poor greeting, not listening
Physician Behavior	Only interested in themselves, Come in when they want, bother Real/rudest jerk, atrocious, zero personality, very unprofessional, unskilled, uninformed, piece of work, confrontational, S.O.B., condescending, very callous, abrupt, uncaring, not friendly, awful, defensive, unpleasant, neglected, unconcerned, disrespectful, very cruel Was confused, did not understand, took phone call during visit Did not visit for days, came and left right away, always in a hurry Wrong meds, no meds, incorrect info, wrong test, no tests, wrong diagnostic Too many doctors and hospitalists see the patient Seriously overbilled, stay with patient just to bill

Physicians

Patients’ negative comments on their experience with physicians are summarized in the Figure 5 Word cloud. Compared to comments in all the other domains, here patients used the strongest words and descriptions. Patient comments can often be raw and tough to digest. However,

the raw nature of patient comments is also a reason they are so valuable to analyze.

The comments in this domain clustered among three major topics: (a) communication; (b) residents; and, (c) physician behaviors (Table 8). Communication was by far

the most important topic in all the comments analyzed across all domains and the words that described it had the highest relative usage frequencies. It is also essential to note that patients identified issues with residents in the hospital, which are not assessed individually in the survey, and therefore cannot be assessed quantitatively.

Overall experience

As a test for consistency in our analyses, in addition to the negative comments in the four domains analyzed above, we also analyzed those written in the survey section for Overall Experience. We find that Overall Experience comments identify the same topics as in the other domains and they highlight the same issues – supporting staff, nurses, room, and physicians/hospitalists. Further, patients use similar relevant words in describing their overall experience. This result suggests that by focusing on the top four domains, a high quality and representative picture of in-patient experiences can be created.

Results of reviewing the comments in this section also brought forward some exemplary negative comments from patients who provided perfect Overall Rating of Care scores (Table 9). Together with the results already shown in the top four domains, this clearly shows that patients may be inflating their ratings, such that health care organizations may not completely understand patient experience issues without including analysis of comments by domain.

Discussion

This research set out to gain insights from the patients who provide perfect ratings yet write negative comments when describing their experiences with hospital care. Data from a large in-patient study showed that while many rate their health experience highly (average 4.5 out of 5), a large segment of patients (from 18% to as high as 52%) also leave negative comments attached to these ratings. Understanding the negative comments of high rating patients is of great significance as much can be learned from their narratives.

There are two major results that emanate from our analysis. First, we identify several issues that are prevalent in health care processes. Previous research established that nurses and physicians greatly impact health experiences,⁷ but our analysis illustrates that patients also stress that the quality of care they receive from allied health professionals (nurse aides, pre and post-op nurses, lactation consultants, technicians, pharmacists, social workers and chaplains). Similarly, in numerous comments, a large number of patients complain about hospital beds, a problem of such magnitude that some patients announce they won't voluntarily return to the hospital. Our analysis also discovers that communication among health professionals is a major problem that spans all domains of health care (e.g., nurses do not communicate with other nurses, physicians do not talk to nurses). Patients even suggest that this lack of communication is a significant source of errors in hospital patient care, e.g. wrong meds, tests, etc. None of these topics are covered in the inpatient survey instrument, so they would have not been discovered without analyzing patients' negative comments.

Second, in our analysis, we find evidence that patients may be inflating their ratings of their health care experience. Although they use very strong words to describe negative experiences in the hospital, participants in the survey do not allow these negative issues to reduce their ratings. This raises an important question: Are hospitals managing domain scores while having blinders on regarding other issues? That is, of course, only one potential factor that could produce a disparity between scores and comments, but it should be considered as recent research has shown that survey domains do not capture all of what patients believe are part of their experiences.⁸

Several implications emerge directly from this research. First, PX professionals should analyze their patients' comments as they can benefit significantly from matching them to survey ratings. As shown, this is particularly true when analysis is performed at the domain level. Second, inpatient surveys should be supplemented to capture more

Table 9: Selected Patient Comments with Perfect Ratings on Overall Rating of Care

*"I didn't want one person to bring down all my ratings."
 "You should spend more money on staff not building."
 "Would have been nice if you had a Physical Therapy Unit. I had to transfer to [another hospital] for PT."
 "The only complaint I have is the beds. They are terrible. So uncomfortable."
 "The only complaint I have is the patient rooms are very GROSS."
 "Only negative was staff filling cabinets in room at 3:45 am."
 "Poor communication between nurses and physicians."
 "Staff did not work as a team bickering or arguing."*

information on patients' perceptions of their experiences. It may be beneficial for many health care organizations to enlist the help of a professional firm to help them make sense of patient comments. Many vendors provide services based on Artificial Intelligence (AI) and Natural Language Processing (NLP) to dig more deeply into patient comments to extract insights into care experiences. Furthermore, client organizations can benefit by having instant access to dashboards that present these insights in digestible formats, reducing the data-to-implementation time and increasing ROI.

This research reveals a potentially fertile area for learning more about patient perspectives of care: contradictory scores and comments. While this research documents this phenomenon and starts the journey to utilize these types of data for deeper understanding and learning, much more can be done. For instance, while patient comment categories were analyzed, and Word clouds were produced to visualize what patients are telling their health organization, an important question is yet to be addressed: Why do some patients give stellar ratings and then write negative comments? We next discuss two plausible explanations supported by analysis of patient comments.

First, patients who are highly loyal to an organization may not want to decrease their ratings, based on an understanding that ratings are important to the organization. This may manifest in the behavior we see here: top box ratings with comments that provide feedback that a patient feels will help an organization improve. Supportive of this line of thinking is that loyal patients plan on visiting a particular hospital again; thus, they may not want to "punish" an organization's ratings, but are highly motivated to share insights from their care experience that may help them avoid similar conditions in the future. If this is the case, the comments analyzed here are from the most loyal patients an organization has, and thus should be given priority in determining areas to improve.

Second, a patient may see a health domain as being predominantly great, but spoiled by "one bad apple." For instance, within the Nurses domain, patients may see a wide variety of health professionals as "nurses," and thus may feel overwhelmed from receiving care from so many "nurses." If only one out of 20 or more "nurses" displays unsatisfactory behaviors, the patients may not discount their ratings, but would provide this type of feedback in the comments section. This is supported by the patient quote provided previously: "I didn't want one person to bring down all my ratings." If this is the case, organizations can identify individuals on staff that may need evaluation, training, and possible intervention.

An important question that emerges from our research is: Why do negative comments matter if I'm getting top box

scores anyway? There are several important issues to consider when answering this question. First, top-box responders who provide negative comments may represent a great number of patients who don't provide any comments, regardless of their scoring. As such, this "tip of the iceberg" theory necessitates that PX professionals listen to feedback to determine areas for improvement. Second, organizations may identify problems with specific employees, and may wish to take further action to ameliorate them. Third, many PX professionals may feel that they have reached a ceiling or plateau with their scores and percentile rankings. They may no longer know where to turn to drive incremental improvement. Patient comments, particularly from those who otherwise had a good experience, provide a source from which to select a new area on which to focus efforts. Finally, as physician transparency gains significant support,⁹ managing an online presence and reputation become more critical. As redacted yet unedited patient comments become easily accessed by the public, incentives to better understand patient perceptions increase. That is, understanding and then improving issues over time will undoubtedly benefit physician profiles through improved patient comments. PX professionals ultimately need to wrestle with the question of whether their job is manage top box responses or patient experiences.

In conclusion, we hope that patients' voices will be heard, and their feedback, wherever it appears and in whatever form, will be collected and analyzed in order for health care organizations to learn more about what patients and families expect when they are admitted to the hospital. Patients deserve the best care, and they often work hard to communicate that to health care organizations. PX professionals should do whatever they can to listen as closely as possible. With this paper we describe one approach on how to analyze and make the most of patients' structured and unstructured feedback.

Endnote

Word clouds have been used regularly in academic literature, and are a common visual description of free-text comments, and a valuable tool to summarize findings in studies that perform text analytics. Examples of articles with word clouds include: Horwitz, Leora I., et al. (2013), "Quality of Discharge Practices and Patient Understanding at an Academic Medical Center." *JAMA Internal Medicine* 173.18: 1715-1722, and Maramba, Inocencio Daniel, et al. (2015), "Web-Based Textual Analysis of Free-Text Patient Experience Comments from a Survey in Primary Care." *JMIR Medical Informatics* 3.2.

References

1. Cornwell, Jocelyn (2015), "Reframing the Work on Patient Experience Improvement," *Patient Experience Journal*, 2 (1), Article 3.
2. Sandager, Mette, Morten Freil, and Janne Lehmann Knudsen (2016), "Please Tick the Appropriate Box: Perspectives on Patient Reported Experience," *Patient Experience Journal*, 3 (1), Article 10.
3. LaVela, Sherri and Andrew S. Gallan (2014), "Evaluation and Measurement of Patient Experience," *Patient Experience Journal*, 1 (1), 28-36.
4. Edwards, Kelly J., Kim Walker, and Jed Duff (2015), "Instruments to Measure the Inpatient Hospital Experience: A Literature Review," *Patient Experience Journal*, 2 (2), Article 11.
5. Sandager, Mette; Freil, Morten; and Knudsen, Janne Lehmann (2016) "Please tick the appropriate box: Perspectives on patient reported experience," *Patient Experience Journal*: Vol. 3 : Iss. 1 , Article 10.
6. Martin, J. H., and Jurafsky, D. (2000), Speech and Language Processing, *International Edition*, 710.
7. Feo, Rebecca, Philippa Rasmussen, Rick Wiechula, Tiffany Conroy, and Alison Kitson (2017), "Developing Effective and Caring Nurse-Patient Relationships," *Nursing Standard*, 31 (28), 54-63.
8. Ranard, Benjamin L., Rachel M. Werner, Tadas Antanavicius, H. Andrew Schwartz, Robert J. Smith, Zachary F. Meisel, David A. Asch, Lyle H. Ungar, and Raina M. Merchant (2016), "Yelp Reviews of Hospital Care Can Supplement and Inform Traditional Surveys of the Patient Experience of Care," *Health Affairs*, 35 (4), 697-705.
9. Lee, Vivian (2017), "Transparency and Trust - Online Patient Reviews of Physicians," *The New England Journal of Medicine*, 376 (3), 197-199.

Appendix 1. Detailed Description of Quantitative Modeling

The likelihood to recommend analysis performed on these data was conducted according to the following steps. Guided by the survey structure, for each experience domain in the survey we created an aggregated domain measure that shows the average experience rating of each patient in that domain, e.g. we created a nurse domain by averaging each patient's experience rating on all statements pertaining to the health service provided by nurses, etc. Then, we visualized the continuous distributions of all aggregated measures. This step of the analysis gave us a summary of the patient data and a clear picture of the composition, average, minimum, and maximum ratings of experience on all health domains monitored. We also identified our variable of interest, likelihood to recommend, analyzed its frequency distribution and checked that any missing data occurred at random.

The next step of the analysis focused on creating patients' model of likelihood to recommend. Based on each patient's response to likelihood to recommend, a patient was determined to be a promoter or other. In line with Net Promoter Score (NPS) research and with the goal to identify the factors that would encourage a patient to become a promoter (as compared to becoming a detractor or passive), we filtered out all the neutral PLS recommendations (passives) and focused our analyses on promoters and detractors.

We used binary logistic regression to model the likelihood to recommend, where a patient's category (promoter/detractor) was the dependent variable and demographics and composite ratings on each of the survey domains were the independent/explanatory variables. We checked all the assumptions of binary logistic regression – through plots, we confirmed the linear relationship between the dependent variable and all explanatory variables; we looked for potential multicollinearity problems and developed correlation analyses for all continuous aggregated measures, through chi-square analyses we checked for potential association among discrete variables. We then built a series of logistic regression models to discover any confounding variables and potential interaction effects. We finished by building the final logistic regression models with standardized and regular parameters. We confirmed the models were statistically significant and their parameters were significantly different from zero. For each model, we created Odds Ratios to quantify the effect of each explanatory factor. We used the magnitude of the standardized coefficients to rank the factors affecting the patient's likelihood to recommend category.

Appendix 2. Detailed Description of Models for Word Informativeness

In the second step of our procedure, we conducted text analyses on the open-ended questions in the survey, namely the comments describing patients' domain experiences (e.g. comments for experience with nurses, physicians, room, meal, etc.) as well as their overall assessment with the hospital stay. We started by looking at the distribution of comments by their sentiment, as coded by the data provider - negative, positive, neutral and mixed comments. We then filtered the negative comments of patients who gave top box domain ratings, the objective of our analysis. These comments were cleaned – meaning, we preprocessed the comments to remove common stop words (e.g. prepositions, determinants, special characters - \$, /, ..., #). Proper names (e.g. doctors' or nurses' names) were also removed since they were inconsequential to our analysis and would have decreased the accuracy in the text analysis stage. Further, in order to discover common themes across all patients, we aggregated the comments by domain (e.g. all patients' comments on nurses, all comments on room experience, etc.).

Given the large number of patient comments and the various types of domains (i.e., comment codes), the best approach is to identify clusters capturing the relevant topics of discussion. In this approach, documents are commonly represented as a sparse vector over the entire feature set of all distinct terms in all input documents (i.e., a term here is defined as a word). However, such an approach comes with two shortcomings: (1) high dimensionality (i.e., a large number of features) and (2) feature sparsity (i.e., features appearing in only few comments or comment codes).¹ To address these issues, we have decided to use the popular TF-IDF analysis metric to extract features by generating a sparse representation of the comments. Moreover, we also reduced the feature space by removing sparse terms. More specifically, we started by converting each patient comment into a set of representative features (i.e., important terms). Researchers have previously shown the importance of medically relevant features to the understanding of the underlying meaning of the text, specifically pinpointing the importance of attribute or feature extraction.^{2,3} We, too, combine medical and feature relevance to the document.

TF-IDF is a widely used and effective metric in information classification and retrieval that seeks to emphasize the importance of a word to a document in a large unstructured data collection.^{4,5} The idea is simply to multiply the term frequency (TF) with the inverse document frequency (IDF) calculated from the entire corpus as shown in Equation 1:

$$\text{TF-IDF}(t) = \text{tf}(t,d) \times \log(N/n_t), \quad (1)$$

where $\text{tf}(t,d)$ is the frequency of term t in document d , N is the total number of documents in the collection, and n_t is the number of documents in which the term t appears.

For each domain, TF-IDF generated a list of terms (alpha-numeric strings) and their relative usage frequencies, inverse document weight per term. Then, we used the TF-IDF weights as generated by Equation 1. The result was a sparse vector representation of the document.

It is important to point out that, unlike other text analysis techniques that simply calculate word frequencies, TF-IDF penalizes common words that may appear often, yet have little importance (e.g. 'is', 'that', 'of', etc.) and assigns higher ratings to words that are meaningful for the medical domain (e.g. 'nurse', 'blood', 'care', 'IV', etc.).

Moreover, we also performed feature reduction to narrow the feature space to a subset of representative features, to filter out noise while preserving meaning without negatively affecting prediction performance. For example, some sparse features like 'hospital' and 'experience' are too general and less relevant to the comment code 'nurse'. Here we followed the approach of Saif et. al.⁶ who define a sparse feature as the number of documents in which the feature appears, divided by the total number of documents in the corpus, as in Equation 2:

$$\text{Sparsity} = n_t/N, \quad (2)$$

where n_t is the number of documents in which the term t appears and N is the total number of documents in the collection. Thus, a term with 0.90 sparsity appears in at least 90% of the documents. Through empirical tests performed on a separate development subset, we chose a sparsity of at least 0.90 to filter out less relevant terms.

With these results we have shown that a quantitative analysis of the free-form textual patient comments can be effective in identifying novel and more detailed topics that other questionnaire-based approaches cannot. Our results are in sync with prior research^{2,7,8} in that we have shown improved results with a reduced (and hence a more representative) feature space. However, unlike Elmessiry et al,² we show that, by applying feature reduction to the collection of comments resulted after

filtering the negative comments of patients who gave top box domain ratings split per comment codes, we reduce noise without getting to a point where the terms are too few to perform any meaningful analysis (e.g., identify a topic). The next step was to take the top ranked 100 terms in the TF-IDF list, parse them and uncover several sub-clusters indicative of clear semantic categories. They represent the main topics discussed by patients in their comments. Finally, we used these topics to get a deep understanding of patients' comments and to provide suggestions towards improving patients' in-hospital experiences. For each domain of care, we created Word cloud maps to help us visualize patients' comments.

We followed here the approach of Doyle et al.⁹ and Lopez, et al.¹⁰ who developed a complex taxonomy of patient comments (see Lopez et al., Table 1). It includes general themes of *overall excellence*, *negative sentiment*, and *professionalism* – characterized by specific factors, such as *interpersonal manner* (e.g., *friendly*, *helpful*, *trustworthy*, *time spent with doctor during appointment*), *technical competence* (e.g., *knowledgeable*, *detailed*, *efficient*), and *system issues* (e.g., *appointment access*, *wait time*, *practice environment*). Doyle, et al. have followed a similar approach of topic categories in a meta-analysis of patient experience research based on search terms. Their terms are classified based on their aspectual classes: (1) relational (similar to *interpersonal manner: emotional and psychological support*, *patient-centered decisions*, *clear information*, and *transparency*) and (2) functional (similar to *professionalism*, *technical competence*, and *systems issues: effective treatment*, *expertise*, *clean environment*, and *coordination of care*). Other researchers have identified similar semantic themes: Greaves et al.¹¹ has used topics like *overall recommendation*, *cleanliness*, and *treatment with dignity to label* 6,412 free-text online comments about hospitals from the English National Health Service. Using a corpus of 33,654 online reviews of 12,898 New York-based medical practitioners, Brody et al.¹² identified words associated with both specialty-independent themes (e.g., *recommendation*, *manner*, *anecdotal*, *attention*, *scheduling*) and specialty-specific themes (e.g., general practitioner: *prescription and tests*, dentist: *costs*, obstetrician/gynecologist: *pregnancy*). Our semi-automatic approach to identifying novel topics in patient comments would help hospital decision makers reach faster conclusions than any manual approach. And this, in turn, will facilitate further improvements.

Appendix 2 References

1. Aggarwal CC, Yu PS. (2000), "Finding Generalized Projected Clusters in High Dimensional Spaces," Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data; The 19th ACM International Conference on Management of Data (SIGMOD); May 15-18, 2000; Dallas, TX. pp. 70–81.
2. Elmessiry A, Cooper WO, Catron TF, Karrass J, Zhang Z, Singh MP (2017), "Triaging Patient Complaints: Monte Carlo Cross-Validation of Six Machine Learning Classifiers," *JMIR Medical Informatics*, 5(3):e19. doi:10.2196/medinform.7140.
3. Wilcox AB, Hripcsak G. (2003), "The Role of Domain Knowledge in Automating Medical Text Report Classification," *J Am Med Inform Assoc.*, 10(4): 330–338.
4. Rajaraman A, Ullman JD. (2012), *Mining of Massive Datasets*. Cambridge: Cambridge University Press.
5. Dumais S, Platt J, Heckerman D, Sahami M. (1998), "Inductive Learning Algorithms and Representations for Text Categorization," *Proceedings of the 7th International Conference on Information and Knowledge Management; The 7th International Conference on Information and Knowledge Management (CIKM)*; Nov 2-7, 1998; Bethesda, MD. pp. 148–155.
6. Saif H, Fernández M, He Y, Alani H. (2014), "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter," *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*; The 9th International Conference on Language Resources and Evaluation (LREC); May 26-31, 2014; Reykjavik, Iceland. pp. 810–817.
7. Liu T, Liu S, Chen Z, Ma WY. (2003), "An Evaluation on Feature Selection for Text Clustering," *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*; The 20th International Conference on Machine Learning (ICML); Aug 21-24, 2003; Washington, DC. pp. 488–495.
8. Cho H, Lee JS. (2016), "Data-Driven Feature Word Selection for Clustering Online News Comments," *Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp)*; The 3rd International Conference on Big Data and Smart Computing (BigComp); Jan 18-20, 2016; Hong Kong. pp. 494–497.
9. Doyle C, Lennox L, Bell D. (2013) "A Systematic Review of Evidence on the Links between Patient Experience and Clinical Safety and Effectiveness," *BMJ Open*. 2013;3:e001570–0.
10. López A, Detz A, Ratanawongsa N, Sarkar U. (2012), "What Patients Say About Their Doctors Online: A Qualitative Content Analysis," *J Gen Intern Med*. 2012;27: 685–92.
11. Greaves F, Millett C, Nuki P. (2014), "England's Experience Incorporating 'Anecdotal' Reports From Consumers into Their National Reporting System: Lessons for the United States of What to Do or Not to Do," *Medical Care Research and Review*. 2014; 71: 65S–80S.
12. Brody S, Elhadad N. (2010), "Detecting Salient Aspects in Online Reviews of Health Providers," *AMIA Annu Symp Proc*; 2010. pp. 202–6.