Q Manage Health Care
Vol. 21, No. 1, pp. 1–11
© 2012 Wolters Kluwer Health | Lippincott Williams & Wilkins

Feasibility of Real-Time Satisfaction Surveys Through Automated Analysis of Patients′ Unstructured Comments and Sentiments

Farrokh Alemi, PhD; Manabu Torii, PhD; Laura Clementz, MA; David C. Aron, MD, MS

This article shows how sentiment analysis (an artificial intelligence procedure that classifies opinions expressed within the text) can be used to design real-time satisfaction surveys. To improve participation, real-time surveys must be radically short. The shortest possible survey is a comment card. Patients' comments can be found online at sites organized for rating clinical care, within e-mails, in hospital complaint registries, or through simplified satisfaction surveys such as "Minute Survey." Sentiment analysis uses patterns among words to classify a comment into a complaint, or praise. It further classifies complaints into specific reasons for dissatisfaction, similar to broad categories found in longer surveys such as Consumer Assessment of Healthcare Providers and Systems. In this manner, sentiment analysis allows one to re-create responses to longer satisfaction surveys from a list of comments. To demonstrate, this article provides an analysis of sentiments expressed in 995 online comments made at the RateMDs.com Web site. We focused on pediatrician and obstetrician/gynecologist physicians in District of Columbia, Maryland, and Virginia. We were able to classify patients' reasons for dissatisfaction and the analysis provided information on how practices can improve their care. This article reports the accuracy of classifications of comments. Accuracy will improve as the number of comments received increases. In addition, we ranked physicians using the concept of time-to-next complaint. A time-between control chart was used to assess whether time-to-next complaint exceeded historical patterns and therefore suggested a departure from norms. These findings suggest that (1) patients' comments are easily available, (2) sentiment analysis can classify these comments into complaints/praise, and (3) time-to-next complaint can turn these classifications into numerical benchmarks that can trace impact of improvements over time. The procedures described in the article shows that real-time satisfaction surveys are possible.

Key words: *XXX, XXX, XXX* **[AQ2]**

Author Affiliations: *Department of Health Systems Administration (Dr Alemi) and Imaging Science and Information Systems Center (Dr Torii), Georgetown University, Washington, District of Columbia; and VA HSR&D Quality Enhancement Research Initiative Center for Implementation Practice & Research Support (Ms Clementz and Dr Aron), Louis Stokes Cleveland Department of Veterans Affairs Medical Center, Cleveland, Ohio.*

Correspondence: *Farrokh Alemi, PhD, Department of Health Systems Administration, Georgetown University, 3700 Reservoir Rd, Washington, DC 20007 (fa@georgetown.edu)*

F.A. started Rapid Improvement Inc, a company focused on providing real-time satisfaction surveys using time to next complaint and sentiment analysis. L.C. and D.A. received grant support from VA Health Services Research and Development Service Quality Improvement Research Initiative. The authors extend special thanks to Laura Johnson, who assisted in preparation of the data. **[AQ1]**

The authors declare no conflict of interest.

DOI: 10.1097/QMH.0b013e3182417fc4

Satisfaction with care is an important patient outcome. Yet, health care managers and local improvement teams often do not have access to timely and reliable data on patient satisfaction. Typically, surveys are mailed over relatively long intervals. Patients require frequent and costly reminders. Satisfaction reports are available months after the patients' experience, and improvement action is delayed unnecessarily. Because it is expensive to collect patients' responses in each case, most surveys rely on a sample of patients and therefore conclusions cannot be derived at the provider level. In addition, these surveys are not frequent enough to allow monitoring of the efforts of improvement teams. As a consequence, it is difficult to relate survey results to specific improvement efforts or to unusual circumstances (eg, a snow storm). We set out to rethink how satisfaction surveys should be done so that the results can be available in real time and at a fraction of current costs.

By real time, we mean surveys that inform managers and improvement teams within a day of care. When reports are continuous and about events that are unfolding, managers and improvement teams can more easily understand the causes of dissatisfaction. If information is collected at every visit, then the surveys need to be radically short so as not to interfere with clinical practices. Short surveys also accomplish a secondary goal of reducing the cost of data collection by increasing response rate and removing the need for costly reminders. The shortest possible survey is to have a single item. Alemi and colleagues[1] suggested a 2-question survey called the "Minute Survey" that had 50% to 75% response rate without the use of reminders. The first question asks patients to rate their overall experience. The second question asks: "Tell us what worked well and what needs improvement? (c)" This open-ended question provides an opportunity for patients to explain the reasoning behind their rating. However, open-ended questions, while a potentially rich source of information, require more efforts in the way of analysis. These questions are easy to collect but difficult to analyze.

Several problems arise from using open-ended questions. First, how could we report numerical benchmarks, using responses to an open-ended question? Clinicians need numerical benchmarks to examine whether they are improving overtime and to compare their operation to other units within and outside the organization. Text data limit the ability of managers and improvement teams to benchmark their performance. Second, improvement teams need to know what to do so as to improve patient satisfaction. To review thousands of patient comments may not be feasible. A method is needed that summarizes the comments. This article describes the use of "Sentiment Analysis" as a method for classifying and grouping comments. We further report how time-to-dissatisfied customer can be used to produce numerical benchmarks for these classified comments. By showing that text data can be used to create numerical benchmarks, we hope to show the feasibility of a low-cost method of getting to patients' satisfaction with care.

METHODS

Sentiment analysis

The framework for analysis of comments is illustrated in Figure 1. First, dated comments are obtained [F1] from many different sources. A human reviewer classifies a sample of these comments into various categories. Sentiment analysis is used to learn which pattern of words that predict a complaint.[2-5] Then, these patterns are used to classify future comments into complaint or praise. There are different methods available for conducting sentiment analysis. Here we focus on a collection of artificial intelligence and text-processing tools, including (*a*) Decision Trees, (*b*) Bagging, (*c*) Support Vector Machine, and (*d*) Naïve Bayes Multinomial. Because there is a limited vocabulary that is useful for commenting about health services, it is possible to anticipate the meaning of various phrases and automatically classify the comments. When a comment is not understood or does not match the limited vocabulary, a human being classifies the comment and the classification system

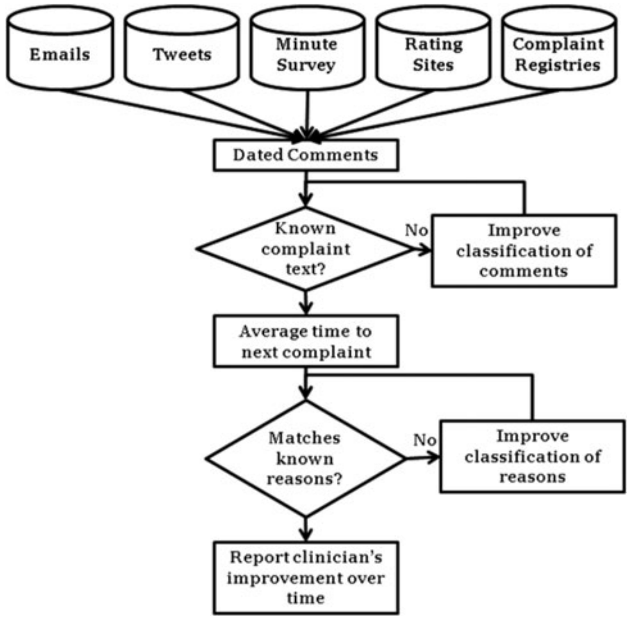
Emails
Tweets
Minute Survey
Rating Sites
Complaint Registries
Dated Comments
Known complaint text?
No
Improve classification of comments
Average time to next complaint
Matches known reasons?
No
Improve classification of reasons
Report clinician's improvement over time

Figure 1. The process of data collection and analysis of sentiments.

is improved by adding the comment to the training data set.

Comments are first classified into praise or complaint and then further classified into various types, such as the following:

1. Physician-related concerns
2. Staff-related issues
3. Getting in to be seen
4. Wait-related issues
5. Cost-related issues
6. Nurse-related concerns
7. Facility-related concerns
8. Privacy-related concerns
9. Other

These categories, or drivers of dissatisfaction, were taken from Consumer Assessment of Healthcare Providers and Systems' Clinician and Group Survey.[6] Other categories can also be selected on the basis of the specific purpose of the analysis. Comments falling into the other category are further analyzed using human raters.

Satisfaction benchmarks

We used the average number of visits-to-next complaints as our overall measure of satisfaction.[7] Complaints are made at different times. It is not easy to summarize the implications of multiple complaints into one single statistic. Traditional surveys calculate rate of dissatisfaction by examining percentage of patients who are not satisfied, combining data from patients in different time intervals. In real-time analysis, this is not possible. There is only 1 observation per visit, and it is not clear how dissatisfaction rates could be calculated from a single observation. One possibility is to measure the average number of visits-to-next complaint. This statistic can be calculated at each visit. It is inversely related to the rate of dissatisfaction per visit:

$$\textit{Rate of dissatisfaction} = \frac{1}{1 + \textit{Average number of visits of to next dissatified patient}}$$

The longer it takes to receive a complaint about a provider, the lower the rate of dissatisfaction per visit.

Source of data

We utilized a publicly available convenience sample consisting of 995 comments from the Rate My MD site (http://www.ratemds.com/) for obstetrician/gynecologists in Maryland, Virginia, and District of Columbia. These comments cover the 5-year period from July 12, 2004, to April 13, 2010. On the site there is no requirement to have a login to review ratings and as such there is no expectation of privacy.[8] In most cases, no login is necessary to post reviews; however, one is required if a user would like to modify it once posting has been completed. Doctors are added as needed to create a review by the users; that is, the list of doctors is user generated. The site defines those allowed to post reviews as users who have first-hand knowledge of the doctor but makes it clear that they have no way to know who actually posts the information. To assist with the quality of the information, the safeguards RateMSs.com has in place include a review of every posting and editing/eliminating comments that are not relevant or defamatory, requiring users create a login when making a posting for a doctor with 25 or more postings listed and automatic deletion of multiple postings from the same computer. Both users and doctors are able to send a rating back to the site for review with explanation why it should be removed and if they create a login, respond to any rating.

The survey on the site asks patients to rate punctuality, helpfulness, and knowledge. It also asks for an overall rating of clinic. In addition to these ratings, patients were asked to enter comments in free-text format. These comments usually contain praise or complaints about doctors, staff, facilities, and processes at clinic, but they are not always reflected in the overall rating. For example, 6% of the patients who gave the highest overall rating still included complaint in their comments, and 33% of the patients who gave the lowest overall rating included praise, besides their complaints.

In this study, comments were manually reviewed by a data curator, and each comment was classified as "Praise," "Complaint," "Both" (Praise and Complaint), or "Neither," exclusively. Among the reviewed comments, 688 were praises, 210 were complaints, and 97 were both. They were relabeled to yield 785 instances for praise (688 + 97) and 307 instances for complaints (210 + 97), where 97 were labeled as both praise and complaints. In addition, comments were coded with 1 or more of the 24 reasons for dissatisfaction. Note that a comment could be coded with several reasons. A list of the reasons [T1] is shown in Table 1. Reasons that did not appear at least 50 times were not included in the Sentiment Analysis.

Table 1

CLASSIFICATION OF COMMENTS

1. MD-related concerns
   a. Doctor's advice and treatment: 425 comments[a]
   b. Time doctor takes: 181 comments[a]
   c. Extent doctor listens to the patient: 61 comments[a]
   d. Explanations provided by the doctor: 168 comments[a]
   e. Other MD-/NP-related concerns: 957 comments
2. Staff-related issues[a]
   a. Staff answers to questions: 5 comments
   b. Staff friendliness and helpfulness: 88 comments
   c. Other staff-related issues: 127 comments
3. Access-related issues[a]
   a. Getting in to be seen: 21 comments
   b. Other access-related concerns: 70 comments
4. Wait-related issues[a]
   a. Time in examination room: 3 comments
   b. Wait for test results: 3 comments
   c. Time in waiting room: 48 comments
   d. Other wait-related issues: 55 comments
5. Pay-related issues
   a. What patient pays: 1 comment
   b. Other pay-related issues: 27 comments
6. Nurse-related concerns
   a. RN friendliness and helpfulness: 13 comments
   b. Other nurse-related concerns: 19 comments
7. Facility-related concerns
   a. Building cleanliness: 7 comments
   b. Comfortable and safe building: 1 comment
   c. Other facility-related concerns: 15 comments
8. Privacy-related concerns
   a. Other privacy-related concerns: 2 comments
9. Unclear complaint: 18 comments
10. Otherwise unclassified: 805

[a]Reasons that occurred frequently. [AQ4]

Comments were first tokenized into words, where non-alpha-numeric symbols (white space, comma, period, hyphen, etc) were used as token delimiters. Letters in these tokens were lowercased, and different numbers ("0," "1," . . . , "9") were converted into one common symbol ("9"). From normalized token sequences of words, phrases were extracted. In the classification literature, these phrases are referred to as *n* grams, where *n* indicates the number of words in the phrase. We examined continuous sets of 1, 2, or 3 g. All *n* grams, except for those that were observed in less than 3 comments, were considered for classification of the comment.

In order for improved performance and efficient computation, a feature selection method was employed to reduce the number of *n* grams considered. We initially tested information gain and chi-square test that have been commonly used for feature selection in document classification. Provided with class-labeled instances, these methods can be used to calculate a score for each *n* gram independently, and candidate features can be sorted on their scores. Top-ranked *n* grams were assumed to be more informative in building machine learning classifiers. We found that information gain and chi-square test ranked *n* grams similarly, and performance of resulting classifiers was similar. For our experiments, therefore, we just employed information gain. With using all or selected features, binary feature vectors were derived, each representing presence (1) and absence (0) of selected *n* gram features in a comment.

Among the multiplicity of classification algorithms,[9] we used decision trees (REPTree), bagging (bootstrap aggregation) with decision trees (Bagging with REPTree), naive Bayes (NaïveBayes-Multinomial), and Support vector machine (SVM) implemented in WEKA machine learning package.[10] The decision tree algorithm has high readability of resulting classification models. Bagging can improve classification performance. Support vector machine and multinomial naive Bayes have been widely used for document classification, which yielded good classification performance in past studies, including ours.[11] In training classifiers using these algorithms, we used default parameters in WEKA, except for SVM. For SVM, we switched the kernel function to RBF, instead of the linear function (the polynomial function with the degree 1) and opted to use logistic models to yield probability-like output scores. This SVM setting yielded improved classification performance in our preliminary experiments.

RESULTS

We obtained comments on 200 physicians and the numbers of comments per physician ranged from 1 to 57 with a mean of 9.59 and standard deviation of 8.10 comments per physician. Each comment was classified into multiple classes. In almost every comment, some component of the comment could not be classified into standardized reasons for dissatisfaction.

We first used the entire data set to calculate information gain of *n* grams ($n = 1$, 2, and 3) and examined them for each category. Part of these results is shown in Table 1. For each reason for dissatisfaction listed in Table 1, we also list the top-10 most informative *n* grams. For example, the most informative 3-gram phrase for classifying "Doctor gives good advice and treatment" was the 3-word phrase: "very_knowledgeable_and."

Next, we examined the classification performance, using different machine learning algorithms. For each category, we conducted 5 repetitions of 2-fold cross-validation tests. Namely, the data set was first partitioned into 2 subsets. Then, classifiers were trained on one subset and tested on the other subset. This process was repeated again by switching the training and the test set. This 2-fold cross-validation test was repeated 5 times using different splits of the training data set. Average performance of 10 runs (5×2) is reported. As performance measures, we employed widely used performance measures, precision, recall, F-measure (F1-measure), and Area Under Receiver Operating Curve (AUC).

We used information gain to limit the number of 3-gram phrases to 1500. Noticing that the performance of naive Bayes was particularly sensitive to the number of *n* grams selected, we also tested data sets with feature sizes of 250, 500, and 1,000 for naive Bayes in (i) and (ii) both. The performance measures (precision, recall, F-measure, and AUC) are found in Table 2. [T2]

For the purposes of demonstrating the procedures for calculation of days till next complaint, we present data for physicians in District of Columbia. We examined days to next complaint, using the reviewer's rating of the comments on the http://RateMDs.com Web site. For 28% of physicians there were no complaints. We assumed that these physicians would have had 1 complaint after the 5 years of the study period. For physicians who had at least 1 complaint, days to complaint ranged from 38 to 1415 days. There was a great deal of variability in days till next complaint for physicians in this market. Some had no complaints and others had nearly monthly complaints. Figure 2 [F2] provides the distribution of days till next complaints.

DISCUSSION

Our study shows that Sentiment Analysis can be applied to qualitative patient satisfaction data. Combined with a measure of time to next complaint, this approach is at least conceptually familiar to those involved in quality improvement efforts.

Table 2

TOP-RANKED *n* GRAMS BY INFORMATION GAIN[a]

Praise	
1	not (0.031), rude (0.021), best (0.019), dr (0.019), wonderful (0.019), love (0.018), great (0.018), information (0.017), cold (0.016), unprofessional (0.015)
2	did_not (0.028), will_not (0.016), dr_<name> (0.014), the_best (0.014), not_go (0.014), i_went (0.014), attitude_problem (0.013), she_didn (0.011), t_want (0.011), i_love (0.011)
3	t_want_to (0.014), know_what_she (0.0.13), not_go_to (0.013), she_did_not (0.012), she_didn_t (0.011), to_come_back (0.011), is_the_best (0.011)
Complaint	
1	but (.066), rude (.060), not (.055), get (.045), wait (.040), then (.033), if (.032), it (.031), back (.031), however (.028)
2	if_you (0.026), did_not (0.025), the_wait (0.024), the_best (0.024), rude_and (0.023), to_get (0.021), the_doctor (0.019), don_t (0.019), get_a (0.018), difficult_to (0.017)
3	i_don_t (0.013), to_come_back (0.012), the_weiting_room (0.011), i_did_not (0.011), she_is_not (0.010), the_wait_is (0.010), i_will_not (0.010), not_go_to (0.010), seemed_to_be (0.010),
Doctor gives good advice & treatment	
1	knowledgeable (0.16), knowledgable (0.035), my (0.021), very (0.018), and (0.018), delivered (0.017), extremely (0.0.17), rude (0.017), excellent (0.016), best (0.015)
2	knowledgeable_and (0.062), very_knowledgeble (0.052), and_knowledgeable (0.039), knowledgeable_i (0.020), knowledgable_and (0.018), extremely_knowledgeable (0.016), knowledge_and (0.015), very_knowledgable (0.014), saved_my (0.012), he_is (0.012)
3	very_knowledgeable_and (0.022), is_very_knowledgeable (0.017), is_a_very (0.013)
Doctor takes enough time	
1	time (0.176), takes (0.078), questions (0.049), answer (0.030), spends (0.027), rushed (0.024), took (0.023), you (0.020), your (0.019), all (0.019)
2	time_to (0.128), the_time (0.080), time_with (0.068), takes_time (0.045), takes_the (0.032), took_the (0.030), questions_and (0.029), to_answer (0.025), feel_rushed (.020), and_takes (0.020)
3	the_time_to (0.080), time_to_answer (0.043), took_the_time (0.032), time_with_you (0.030), time_to_talk (0.025), time_to_listen (0.020), she_took_the (0.017), to_answer_all (0.015), does_not_rush (0.015), of_time_with (0.015)
Doctor explains well	
1	questions (0.211), answer (0.078), answered (0.068), time (0.046), explains (0.028), answers (0.028), takes (0.024), answering (0.022), all (0.020), explained (0.017)
2	my_questions (0.084), to_answer (0.063), questions_and (0.060), time_to (0.045), answered_all (0.037), your_questions (0.024), any_questions (0.021), questions_i (0.021), the_time (0.019), to_explain (0.019)
3	time_to_answer (0.031), my_questions_and (0.026), the_time_to (0.025), answered_all_of (0.023), of_my_questions (0.022), all_my_questions (0.021), to_answer_any (0.016), to_answer_all (0.016), willing_to_answer (0.016), questions_i_had (0.016)
Staff related	
1	staff (.445), office (.090), are (.036), nurse (.034), they (.032), friendly (.030), rude (.027), his (.026), nurses (.026), receptionist (.022)
2	staff_is (0.141), his_staff (0.083), the_staff (0.081), office_staff (0.077), her_staff (0.074), staff_are (0.039), staff_was (0.037), and_his (0.026), the_office (0.025), very_friendly (0.018)
3	her_staff_is (0.035), the_staff_is (0.033), his_staff_is (0.030), the_office_staff (0.030), and_his_staff (0.030), office_staff_is (0.026), his_staff_are (0.015), staff_is_wonderful (0.015), dr_<name>_and (0.015), staff_is_very (0.015)

(*Continues*)

Table 2

TOP-RANKED *n* GRAMS BY INFORMATION GAIN[a] (*Continued*)

Staff friendly and helpful	
1	staff (0.156), friendly (0.049), office (0.034), helpful (0.022), are (0.021), <name> (0.019), also (0.017), his (0.016), wonderful (0.015), they (0.015)
2	staff_is (0.067), his_staff (0.041), very_frinedly (0.032), staff_are (0.028), the_staff (0.024), friendly_and (0.024), her_staff (0.023), staff_was (0.020), dr_<name> (0.019), office_staff (0.019)
3	staff_is_wonderful (0.025), his_staff_is (0.019), staff_is_also (0.018), are_very_friendly (0.018), staff_is_frinedly (0.018), <name>_and_his (0.018), also_very_frinedly (0.018), dr_<name>_and (0.017), his_staff_are (0.016), very_frinedly_and (0.016)
Doctor listens	
1	listen (0.053), listens (0.051), listened (0.041), listener (0.033), concerns (0.016), to (0.010)
2	listens_to (0.041), to_listen (0.033), listened_to (0.023), listen_and (0.025), listen_to (0.018), good_listener (0.016), great_listener (0.012), to_my (0.011)
3	really_listen_to (0.016), time_to_listen (0.013), to_my_concerns (0.013), listened_to_my (0.013), listens_to_me (0.012), listens_to_your (0.012), to_listen_to (0.012), and_really_listens (0.012)
Wait related	
1	wait (0.127), long (0.053), waiting (0.045), office (0.040), hour (0.037), worth (0.035), minutes (0.031), waited (0.028), staff (0.021), but (0.020)
2	the_wait (0.062), to_wait (0.047), on_time (0.042), an_hour (0.029), 99_minutes (0.028), wait_is (0.023), worth_it (0.023), i_waited (0.020), hour_and (0.019), a_long (0.019)
3	had_to_wait (0.021), the_wait_is (0.02), 99_99_minutes (0.017), to_wait_long (0.017), always_on_time (0.013), an_hour_and (0.013), i_have_never (0.011), the_waiting_room (0.011)

[a]Top 10 *n* grams for n = 1, 2, or 3 are shown, and information gain values are shown in parentheses.

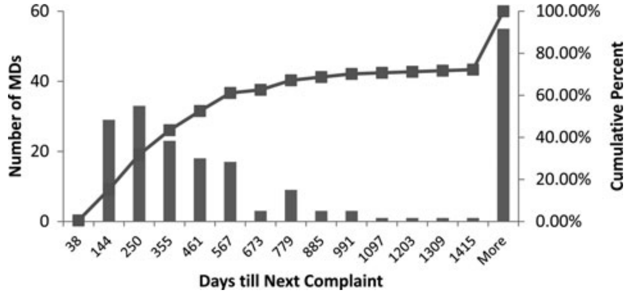

Figure 2. Days to next complaint for obstetrician/gynecologist in District of Columbia based on 5 years of data obtained from RateMDs.com Web site on March 23 2009.

Moreover, data from a single open-ended question can yield information both useful at the individual level and health plan levels to drive improvement. The frequently listed reasons for dissatisfaction uncovered in this study are theoretically sound opportunities for improvements in the physician's practices.

The most surprising finding of the reviewer's classification of comments was that in almost every comment, some component of the comment could not be classified into standardized reasons for dissatisfaction. This suggests that long surveys with preset questions are missing a great deal of information and forcing patients to distort their ideas to fit the questions asked. For example, a patient who wants to complain about parking may not find the question in the long survey and therefore may inadvertently express his frustration as dissatisfaction with another element of clinic services.

The analysis of days-till-next complaint showed that it is relatively easy to benchmark clinicians. This article has laid out procedures for benchmarking and classifying comments. Using various classification procedures, we obtained relatively high rates of correctly classifying comments into complaints or praise. Even though we were dealing with very limited number of comments, the procedures provided

Table 3

ACCURACY OF PREDICTIONS OF REASONS FOR DISSATISFACTION

Predicted Category	Method	Precision	Recall	*F* Measure	AUC
Praise	Decision Trees	0.90	0.99	0.94	0.55
	Bagging	0.89	1.00	0.95	0.74
	Support Vector Machine	0.96	0.84	0.89	0.85
	Naive Bayes Multinomial	0.95	0.93	0.94	0.90
Complaint	Decision Trees	0.59	0.32	0.40	0.65
	Bagging	0.70	0.71	0.53	0.83
	Support Vector Machine	0.62	0.67	0.64	0.84
	Naive Bayes Multinomial	0.73	0.63	0.68	0.84
Doctor gives good advice & treatment	Decision Trees	0.75	0.59	0.66	0.74
	Bagging	0.83	0.56	0.66	0.81
	Support Vector Machine	0.68	0.72	0.70	0.80
	Naive Bayes Multinomial	0.70	0.62	0.66	0.75
Doctor takes enough time	Decision Trees	0.78	0.53	0.63	0.83
	Bagging	0.84	0.56	0.67	0.89
	Support Vector Machine	0.67	0.69	0.68	0.90
	Naive Bayes Multinomial	0.67	0.61	0.64	0.82
Doctor explains well	Decision Trees	0.70	0.70	0.70	0.81
	Bagging	0.69	0.67	0.68	0.87
	Support Vector Machine	0.70	0.66	0.68	0.90
	Naive Bayes Multinomial	0.73	0.54	0.62	0.83
Staff related	Decision Trees	0.91	0.81	0.86	0.90
	Bagging	0.91	0.81	0.87	0.94
	Support Vector Machine	0.87	0.83	0.85	0.96
	Naive Bayes Multinomial	0.86	0.75	0.80	0.91
Staff friendly and helpful	Decision Trees	0.50	0.27	0.34	0.81
	Bagging	0.67	0.35	0.45	0.89
	Support Vector Machine	0.04	0.68	0.48	0.88
	Naive Bayes Multinomial	0.36	0.47	0.41	0.75
Doctor listens	Decision Trees	0.77	0.77	0.74	0.88
	Bagging	0.75	0.68	0.37	0.93
	Support Vector Machine	0.23	0.73	0.34	0.87
	Naive Bayes Multinomial	0.55	0.29	0.37	0.79
Wait related	Decision Trees	0.68	0.56	0.60	0.78
	Bagging	0.69	0.53	0.59	0.87
	Support Vector Machine	0.41	0.73	0.53	0.89
	Naive Bayes Multinomial	0.51	0.52	0.52	0.80

Abbreviation: AUC, area under receiver operating curve.

evidence on viability of identifying complaints among patient comments.

We were able to conduct this study, using publicly available data from the Internet. Because many online ratings are public, benchmarked data summarizing these ratings are available by name. We have not displayed the names of the physicians who performed best or worst but those data could be displayed from the publicly available comments. Finally, we have shown that we could re-create the dimensions used in longer satisfaction surveys from a list of complaints. Using various classification methods, we successfully classified complaints into standardized items found in CAPHS surveys. The classification systems achieved high accuracy, as measured by precision, recall, area under curve, and F-measure. Thus, our study indicates that it is possible to use comments to measure patient satisfaction with care and the output from such analysis would be similar to output from longer surveys.

There are good reasons why satisfaction surveys are not always done in real time. These surveys take significant time to complete. A typical survey takes many days and sometimes months. First, a sample of patients must be selected. Patients must be stratified and samples must be identified by statisticians and not front-line providers. Second, sending and receiving surveys take time. Sometimes, surveys are printed and mailed to the patient; other times patients are surveyed on the phone. Both methods take several days. Third, many patients fail to respond to surveys. Reminders need to be sent. Patients need to be written to or called. For some populations, multiple reminders must be sent, which further adds to the time the survey takes. The survey forms themselves may be long and relatively time-consuming. The Consumer Assessment of Healthcare Providers and Systems questionnaire has 41 questions.[12] Some patients may leave portions of the questionnaire blank or may rate the final set of responses without reading the questions. All of this takes time and delays the availability of reliable information about the patient's satisfaction with care. Compared to longer surveys, the Minute Survey or the Web comments provide a much simpler method of collecting the patient's satisfaction with care.

We believe that the procedures developed apply to different settings (ambulatory, hospital, and health maintenance care); longer satisfaction surveys are often tailored to specific settings. We believe that the open-ended solicitation of comments helps clients express themselves in their own words and prevents forcing clients into artificial response categories. The short survey creates a high response rate and enables real-time data collection without interfering with clinical practices. Many questions about utility of real-time surveys remain, but it is clear to us that real-time satisfaction surveys are practical. The next step is to test these procedures within a clinic.

The availability of data on the Internet raises other issues. Utilization of comments placed on the Internet for quality improvement purposes is relatively new to the medical industry. However, there are many potential sources for data, some clearly public and others perhaps less so. Free-text comments can be found in many different settings: (1) online sites specifically for physician ratings, for example, http://www.RateMDs.com; (2) e-mail content that is accessible to a firm, for example, Google's Gmail has access to comments made by its clients and an employer has access to e-mails written by its employees; (3) social media, for example, Twitter http://twitter.com/docloop; and (4) formal complaints to hospital risk management units or to state agencies that appear on official Web sites, for example, http://www.medbd.ca.gov. We can imagine an environment where search engines, such as Google, Bing, and others, could allow consumers to search for a doctor and receive benchmarked data on their rate of complaints. However, the validity of the comments in terms of their accuracy in representing the population of patients cannot be readily determined. Within the health care community, many have raised concerns about the accuracy of comments expressed on the Web.[13,14]

Although ethically viable, it is understood that providing physician-specific ratings based upon publicly available sources remains a controversial position within the profession (and engendered much discussion among the authors). The growing popularity of physician-rating Web sites and the information they place in public view forces the debate forward. The concept of beneficence, traditionally reserved for patients, can be redirected towards the welfare of the physician. Examples of fears among physician and physician organizations concerning defamation, slander, and general miss information being placed on the Internet can easily be observed in a variety of sources from medical literature[15] to the Internet.[16] In general, the issue is described as the fear that reviews will be excessively negative. Contrary to those views, current findings show that patient reviews and comments are mostly positive and constructive. In one study, 33 different physician-rating Web sites that contained 190 reviews of 81 physicians evaluated both numeric and narrative reviews and found that 88% of the reviews were positive, 6% were negative, and 6% were neutral.[17] Another study also utilized ratings from http://www.RateMDs.com but, from other states, reviewed 16 703 ratings on 6101 providers from 4 major cities over 5 years. The reviews are based upon 5 categories, each scored on a scale of 1 to 5, with 5 as the highest. Providers were found to have a high mean score for each category of 3.7 to 4.0 out of 5.[18] Finally, 68.8% of comments in our study were praises, 21% were complaints, and 9.7% were both.

This study is not without its limitations. First, our analysis was limited to only 1 dimension of quality of care. Patient satisfaction may or may not parallel technical quality of care. Second, the data are representative only of those who provided ratings/comments and this may be a biased sample. Third, we looked at only a sample of comments; our measures of classification accuracy will improve further when several thousands or millions of comments are analyzed. Fourth, we looked at only 1 region of the country and the results may not be generalizable. Fifth, while we have examined how real-time surveys could be carried out through comment cards, we have not actually done so. Furthermore, the linkage between real-time surveys and actual quality improvement at both provider and plan levels has not been demonstrated. Nevertheless, we believe that this article shows that it is theoretically possible to rely on comments and such comments provide rich information.

This study shows that Sentiment Analysis can clarify what has led to dissatisfaction. It is not easy for improvement teams to read and summarize thousands of comments. Long satisfaction surveys typically ask about various areas of dissatisfaction, including access to care, organization of care, clinic environment, and clinicians' performance.[19,20] Thus these surveys provide additional guidance regarding what might be leading to low (or high) dissatisfaction ratings. Before adopting such an approach wholesale, it will be necessary to conduct comparative studies with traditional methods. In addition, as the research progresses, transparency of methods and remaining cognizant of the context of the data are required to avoid alienating members of the medical profession. But the approach we have laid out in this article is promising if it can reduce data collection and increase improvement efforts.

REFERENCES

1. Alemi F, Badr N, Kulesz S, Walsh C, Neuhauser D. Rethinking satisfaction surveys: minute survey. *Qual Manag Health Care*. 2008;17(4):280-291.
2. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2002:79-86. **[AQ5]**
3. Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the Association for Computational Linguistics (ACL)*; 2002:417-424.
4. Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: *Proceedings of the Association for Computational Linguistics (ACL)*. 2005:115-124.
5. Pang B, Lee L. 4.1.2 Subjectivity Detection and Opinion Identification. *Opinion Mining and Sentiment Analysis*. Hanover, MA: Now Publishers Inc; 2008. http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html. **[AQ6]**
6. The CAHPS Database: Preliminary Comparative Data for the CAHPS Clinician & Group Survey. http://www.cahps.ahrq.gov/content/ncbd/pdf/ClinicianGroupPreliminaryReport11_2010.pdf. Accessed March 12, 2011. **[AQ7]**
7. Alemi F, Hurd P. Rethinking satisfaction surveys: time to next complaint. *Jt Comm J Qual Patient Saf*. 2009;35(3):156-161.
8. Kraut R, Olson J, Banaji M, Bruckman A, Cohen J, Couper M. Report of scientific affairs advisory group of the conduct of research on the Internet. 2004;59(2):105-117. **[AQ8]**
9. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. New York, NY: Springer; 2009.
10. Holmes G, Donkin A, Witten IH. *Weka: A Machine Learning Workbench*. Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia; 1994, Retrieved June 25, 2007.
11. Alemi F, Torii M, Atherton M, Pattie DC, Cox K. Reasons-for-appointments can improve case definition of influenza. *J Med Decis Making*. 2011.
12. Cleary PD, Edgman-Levitan S. Health care quality: incorporating consumer perspectives. *JAMA*. 1997;278(19):1608-1612.
13. McCartney M. Will doctor rating sites improve standards of care? No. *BMJ*. 2009;338:b1033.

14. Bacon N. Will doctor rating sites improve standards of care? Yes. *BMJ*. 2009;338:b1030.
15. Black BW, Thompson LA, Saliba H, Dawson K, Paradise Black, NM. An analysis of healthcare providers' online ratings. *Inform Prim Care*. 2009;17:249-253
16. See for example. http//www.medicaljustic.com. Accessed March 11, 2011. [AQ9]
17. Lagu T, Hannon NS, Rothberg MB, Lindenauer PK. Patients' evaluations of health care providers in the era of social networking: an analysis of physician-rating websites. *J Gen Internal Med*. 2010;25(9):942-946.
18. Black BW, Thompson LA, Saliba H, Dawson K, Paradise Black NM. An analysis of healthcare providers' online ratings. *Inform Prim Care*. 2009;17:249-253.
19. Rodriguez HP, Scoggins JF, von Glahn T, Zaslavsky AM, Safran DG. Attributing sources of variation in patients' experiences of ambulatory care. *Med Care*. 2009;47:835-841.
20. Haggerty JL, Pineault R, Beaulieu MD, Brunelle Y, Gauthier J, Goulet F. Practice features associated with patient-reported accessibility, continuity, and coordination of primary health care. *Ann Fam Med*. 2008;6:116-123.

Queries to Author

Title: Feasibility of Real-Time Satisfaction Surveys Through Automated Analysis of Patients' Unstructured Comments and Sentiments
Author: Farrokh Alemi, Manabu Torii, Laura Clementz, David C. Aron

[AQ1]: Please check whether the conflict of interest footnote is OK.
[AQ2]: Please provide 4 to 6 key words.
[AQ3]: Please check whether the running head, affiliation, correspondence, and social titles (Mr/Ms) are OK as set. Also review the conflict of interest statement.
[AQ4]: Please check the footnotes set in Tables 1 and 2.
[AQ5]: For Refs 2, 3, and 4, were these from printed books? If so, please provide publisher name and location. If not, please provide further information on where these refs came from.
[AQ6]: Ref 5: Please check what "4.1.2 Subjectivity Detection and Opinion Identification" indicates. Provide an access date for the URL.
[AQ7]: Ref 6: The Web page doesn't link with the intended page. Please check.
[AQ8]: For ref 8, please provide the name of the journal in which this report was published.
[AQ9]: For ref. 16, please provide the author name(s) and the article title.