Journal of the American Medical Informatics Association, 27(12), 2020, 1834–1843
doi: 10.1093/jamia/ocaa194
Advance Access Publication Date: 26 October 2020
Research and Applications

Research and Applications

An approach to predicting patient experience through machine learning and social network analysis

Vitej Bari [1], Jamie S. Hirsch,[1,2,3] Joseph Narvaez,[4] Robert Sardinia,[4] Kevin R. Bock,[1] Michael I. Oppenheim,[1,2] and Marsha Meytlis[1]

[1]Department of Information Services, Northwell Health, New Hyde Park, New York, USA, [2]Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hofstra University, Hempstead, New York, USA, [3]Institute of Health Innovations and Outcomes Research, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, New York, USA, and [4]Office of Patient and Customer Experience, Northwell Health, Lake Success, New York, USA

Corresponding Author: Vitej Bari, MS, Department of Information Services, Northwell Health, 1981 Marcus Avenue, Suite E100, New Hyde Park, NY 11042, USA; vbari@northwell.edu

Received 21 February 2020; Revised 26 June 2020; Editorial Decision 3 August 2020; Accepted 8 September 2020

ABSTRACT

Objective: Improving the patient experience has become an essential component of any healthcare system's performance metrics portfolio. In this study, we developed a machine learning model to predict a patient's response to the Hospital Consumer Assessment of Healthcare Providers and Systems survey's "Doctor Communications" domain questions while simultaneously identifying most impactful providers in a network.
Materials and Methods: This is an observational study of patients admitted to a single tertiary care hospital between 2016 and 2020. Using machine learning algorithms, electronic health record data were used to predict patient responses to Hospital Consumer Assessment of Healthcare Providers and Systems survey questions in the doctor domain, and patients who are at risk for responding negatively were identified. Model performance was assessed by area under receiver-operating characteristic curve. Social network analysis metrics were also used to identify providers most impactful to patient experience.
Results: Using a random forest algorithm, patients' responses to the following 3 questions were predicted: "During this hospital stay how often did doctors. 1) treat you with courtesy and respect? 2) explain things in a way that you could understand? 3) listen carefully to you?" with areas under the receiver-operating characteristic curve of 0.876, 0.819, and 0.819, respectively. Social network analysis found that doctors with higher centrality appear to have an outsized influence on patient experience, as measured by rank in the random forest model in the doctor domain.
Conclusions: A machine learning algorithm identified patients at risk of a negative experience. Furthermore, a doctor social network framework provides metrics for identifying those providers that are most influential on the patient experience.

Key words: patient experience, electronic medical record, physician-patient relations, machine learning, social network analysis

INTRODUCTION

In addition to the underlying quality of care delivered, a major goal of healthcare organizations is to provide a positive experience to every patient during their medical journey. Positive patient experience is a healthcare quality aim proposed by the Institute of Medicine,[1] and patient experience is positively associated with clinical effectiveness and patient safety.[2] Patient experience information can there-

fore be used to drive strategy for transforming practices as well as to drive overall system transformation.[1] The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey is the first national, standardized, publicly reported survey of patients' perspectives of hospital care, and represents an objective measure of patient experience. The Centers for Medicare and Medicaid Services has developed HCAHPS[3] star ratings to assess excellence in healthcare quality. The results of these surveys are used as marketing tools for healthcare organizations, as compensation determinants for physicians, as a direct medium for comparison for healthcare organizations, and to improve the quality of patient experience.[4,5]

Previous work has shown that various factors are associated with patient experience, including complaints about a care provider; the courtesy of support staff; the time providers spend with a patient; the duration of wait time; the clarity of discharge information; the cleanliness of the treatment area; and the patient gender, language, education, and health status.[6–10] Of the 19 questions on the HCAHPS patient experience survey, 3 deal directly with provider communication. Effective doctor-patient communication is a central clinical function in building a therapeutic relationship, and is essential for delivery of high-quality health care. It is therefore unsurprising that the group of doctors caring for a patient has a significant effect on a patient's overall satisfaction.[7] Patients whose physicians provide information in a comprehensible manner are more likely to acknowledge health problems, understand their treatment options, modify their behavior accordingly, and follow their medication schedules.[11–14]

There is also an increasing appreciation of the importance of provider-based training to improve provider interactions and communication with patients. Identifying the most impactful providers can facilitate prioritization of training and education. Identifying provider leaders who can optimize uptake of improved communication practices across clinical teams[15,16] can enable quality improvement. Focusing communication training on the highest-priority providers—as identified by connectedness in the social network—can facilitate diffusion of best practices throughout,[17] thereby improving provider-patient communication and ultimately patient experience.

This study aimed to identify 2 novel interconnected approaches to improve patient experience. The first approach identifies patients who are at risk of a negative experience. More specifically we predicted patient experience survey scores to provider-related questions. In the prediction component, patients' risk of a negative experience was determined based on 2 types of features: an individual patient's personal demographic and the specific providers caring for the patient. The second approach tests the hypothesis that within each network, doctors who are very connected are likely most influential on patient experience. Social network analysis is used to determine the "connectedness" of each provider.

The predictions from both approaches will be implemented in production in the hospital to improve patient experience. In the first approach, if the model predicts that a negative patient experience is likely to take place (in the provider domain) the hospital will intervene by having a customer service specialist come to the patient's room. The customer service intervention will mitigate the risks of a negative patient experience. Thus, real-time service recovery will be facilitated. In the second approach, when influential providers are identified, they will receive special bedside manner training. The goal of both of these approaches—a patient-centric, real-time experience prediction and a focus on provider education and instruction—is to facilitate improvements in the experience for hospitalized patients.

MATERIALS AND METHODS

Study population

The patient cohort consisted of all patients 18 years of age or older, who were discharged from an 800-bed tertiary hospital within the Northwell Health system between January 1, 2016, and February 29, 2020, and who responded to the postdischarge HCAHPS survey. Patients who expired in the hospital were excluded from the analysis. The HCAHPS survey is offered to all patients across medical conditions, 7-14 days after hospital discharge. The remaining patient population was further processed as described in the Outcomes subsection.

All clinical data were collected from the enterprise inpatient electronic health record database (Sunrise Clinical Manager; Allscripts, Chicago, IL). At Northwell Health, all HCAHPS surveys are administered by Press Ganey Associates (South Bend, IN), and results are returned to the health system. Survey responses and unique patient identifiers that facilitate linkage back to clinical data are stored in Microsoft SQL Server analytics databases.

Study design

Our study comprised 2 parts. In the first part, we identified patients who are at risk for responding negatively to provider communication questions. We developed predictive models using machine learning algorithms, with performance assessed by the area under the receiver-operating characteristic curve (AUC). In the second part, we used social network analysis to identify providers who are most influential on patient experience.

This study was reviewed and approved by the Northwell Health Institutional Review Board (#19-0534). All machine learning models were implemented using the scikit-learn version 0.21.2 machine learning library.[18] The NetworkX version 2.3 python library (https://networkx.github.io/documentation/networkx-2.3/) was used for doctor social network visualizations. All analyses were conducted in Python version 3.7 (Python Software Foundation, Wilmington, DE). There were no missing values in the dataset.

Outcomes: Response variable

Separate models were built to predict the responses to each of 3 HCAHPS questions:

- "During this hospital stay how often did doctors treat you with courtesy and respect?" (courtesy/respect)
- "During this hospital stay how often did doctors explain things in a way that you could understand?" (understand)
- "During this hospital stay how often did doctors listen carefully to you?" (listen)

The response variable was extracted from the HCAHPS survey responses to doctor communication questions. For each question, patients had 4 response options:

- "Always"
- "Usually"
- "Sometimes"
- "Never"

These 4 responses were converted into a binary outcome variable in which "always" (designated "top box") was considered a desired response, while the other 3 possible response choices were grouped into undesirable. The decision to dichotomize responses, favoring only the most positive ("always") was a business decision as the

Table 1. Explanatory variables

Patient	Doc 1	Doc 2	….	Doc N	CCI	LOS	Acuity	Age (y)	Allergies count
Pat 1	4	1		3	6	1	0	21	0
Pat 2	7	0		0	5	2	0	53	0
Pat 3	0	0		0	3	6	3	78	3
Pat4	3	0		2	8	2	0	28	7
Pat 5	1	0		0	2	5	0	57	3
…									
Pat M	0	9		0	1	3	0	19	0

Different features are represented as columns in the matrix. The features consist of 2342 provider features and 22 patient-centric features. For the provider features, each element in the matrix corresponds to the number of interactions between a provider and a patient.

CCI: Charlson Comorbidity Index; Doc: doctor/provider; LOS: length of stay; Pat: patient.

health system is focused on achieving maximum patient satisfaction, with all other responses indicate room for improvement.

The models were trained in a way so that 0 = top box and 1 = all other responses. For each of these questions, approximately 75% of patients gave the desired response and the remaining 25% gave an undesired ("negative") response. In order to balance the dataset, we used data from all of the patients who had an undesired response, but only approximately one-third of patients who had a desired response. The downsampling for desired responses was done using a random sample. For patients with multiple distinct survey responses (from multiple admissions) during the 45 month study period, only 1 visit per unique patient was kept in the analysis in order to ensure that the AUC was not inflated, as the same patient could have been present in the training and testing set.

Explanatory variables

The distinct features that were included in the model are patient interactions with each of 2364 doctors; patient age, gender, race, marital status, language, and religion; admission through the emergency department; Charlson comorbidity index; hospital length of stay; discharge disposition; maximum pain score, difference between the first and the last pain score, average pain score and standard deviation of pain score (and similar metrics for modified early warning score and patient temperature); number of documented allergies; time spent in various hospital locations, including emergency department, intensive care unit, and general nursing units; month of discharge; and prior survey responses, if available (shown in Supplemental Table 5).

These features were included based on internal discussions with our patient experience and clinical teams. Explanatory variables were shown to be impactful based on exploratory data analysis, done by assessing dissimilarities in histogram distributions of feature values for positive and negative responses; and by removing any features with any null values. Additionally, model performance improvement was performed by assessing AUC values following inclusion of features into the model.

Other features were attempted but ultimately not included: patient bed location (door vs window), radiology study delay (time between x-ray order and performance), analgesia use, and laboratory order patterns.

Doctor-provider interactions were calculated by adding together the total number of documents authored and number of orders placed for the particular patient by that care provider. The structure of the dataset as prepared for input into the model is depicted in Table 1, in which each row is a unique patient visit and each column is a model feature.

Model development

In this study, 3 widely used machine learning classification methods were used: random forest (RF),[19] logistic regression (LR), and decision tree.

The first method was LR for binary classification, with a threshold set at a probability of .5.

The second method was decision tree. This model is a tree-like structure, in which leaves represent outcome labels and branches represent conjunctions on features that resulted in those outcomes.

The third method was a RF model, which grows a forest of classification trees for a binary outcome and can provide a probability estimate of membership in each class.[18–20] Random forest uses bagging,[21] in which sampling of features and observations is done with replacement to the original sample. This model fits multiple classification and regression trees on random subsets of patients and random subsets of variables.[20] We used the mode of individual decision tree predictions to obtain a final prediction. The following model hyperparameters were tuned using scikit-learn library tools: number of trees = 2000, maximum tree depth = 100, the number of features considered at each split = 5, the minimum number of samples required to split an internal node = 2, and minimum number of samples required to be at a leaf node = 2. The full set of parameters that were tested are shown in Supplemental Table 6.

We also tested the performance of an XGBoost model and a gradient boosting decision tree model, but ultimately chose not to include these models, as the results were not significantly different from the RF model.

Model validation

The receiver-operating characteristic (ROC) curve[22] is used to evaluate and compare the RF, LR, and decision tree classification models. This metric was considered appropriate because after downsampling, positive and negative classes were of approximately equal size. This metric would not have been appropriate for the original dataset because it was imbalanced. It has been previously reported in the literature that the AUC metric should not be used when the data are heavily imbalanced. The intuition is that false positive rate for highly imbalanced datasets is pulled down due to a large number of true negatives.[23]

An ROC curve plots false positive rates vs true positive rates, and illustrates the performance of a binary classifier system, as its

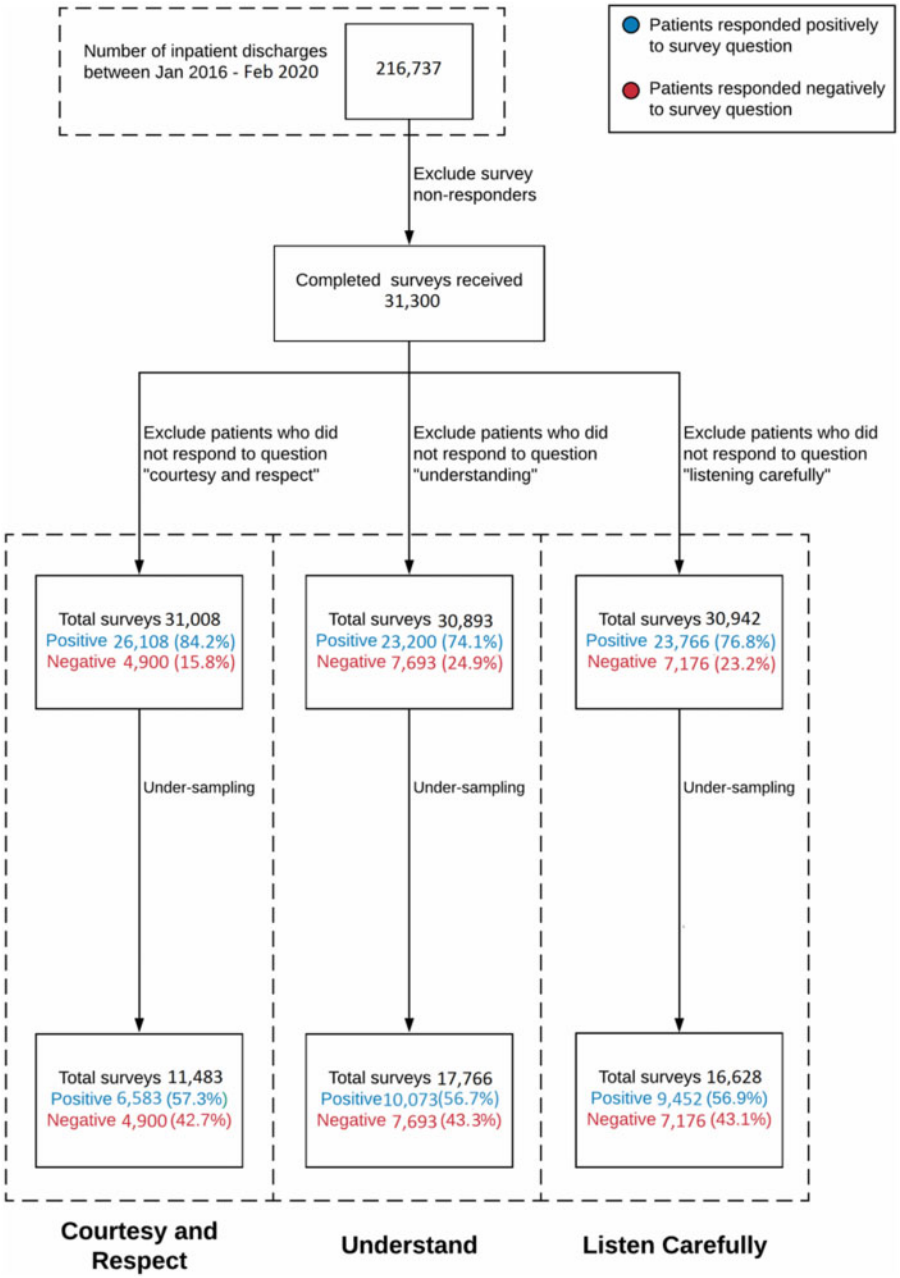

Figure 1. Data flow used in the machine learning model. This diagram shows how the original patient population is reduced for the input to the machine learning model.

discrimination threshold is varied.[22] In this study, desired patient ratings and undesired patient ratings were considered a 2-class prediction problem (binary classification). 30% of the data was set aside as test data. The remaining 70% identified as training data was used for 5-fold cross validation as part of hyperparameter tuning. Finally, once hyperparameters were finalized, model performance was tested on the test data mentioned previously. 95% confidence intervals were calculated by bootstrapping with 1000 iterations.

Assessing variable importance

Mean decrease in impurity and mean decrease in accuracy was used for assessing variable importance:

- Mean decrease in impurity measures the extent of purity for a region containing data points from possibly different classes. Node impurity represents how well the trees split the data.
- Mean decrease in accuracy assesses the effect of scrambling a specific feature on the model AUC.

Prospective validation

A prospective validation was performed using the RF model on the same study population as described previously. The model was trained on data from January 1, 2016, to December 31, 2018, and tested on data from January 1, 2019, to February 29, 2020. The predicted survey responses were compared with actual survey responses, which were received 6 weeks later. We used the same out-

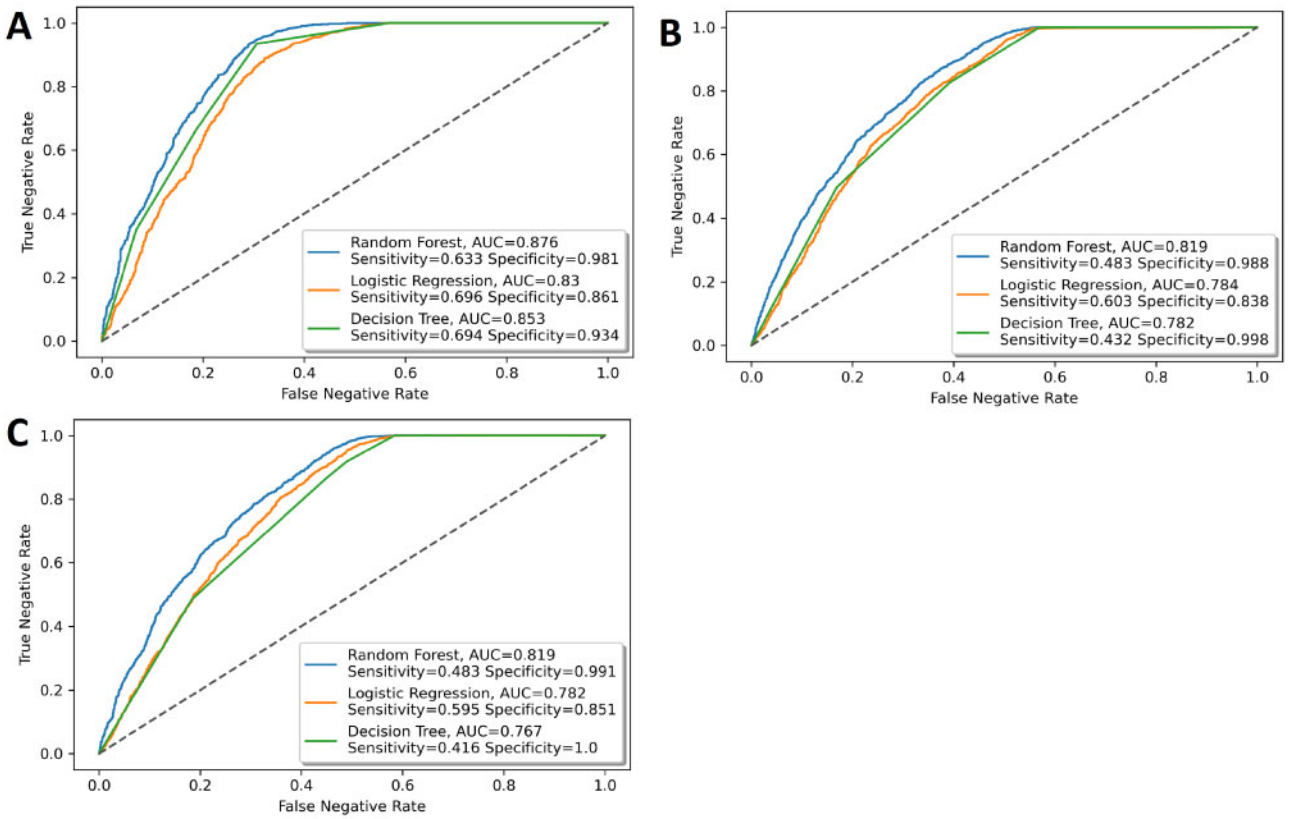

Figure 2. Area under the receiver-operating characteristic curve (AUC) for the prediction of responses of Hospital Consumer Assessment of Healthcare Providers and Systems questions in the doctor domain. The performance of the random forest model is compared with the performance of the logistic regression model and decision tree model. The questions are the following: During your hospital stay, how often did doctors (A) treat you with courtesy and respect, (B) explain things in a way that you could understand, and (C) listen carefully to you?

comes, explanatory variables, and model validation approach as described previously.

Constructing social networks

Social network analysis is a powerful tool that captures hidden channels of collaboration, information flow, and communication between network actors.[24] Using this technique, we can examine patient sharing and collaboration of healthcare providers. This enables assessment of the working relationship of providers involved in a patient's care.

The term *social network* refers to the articulation of a social relationship among individuals such as doctors. We defined a social interaction as occurring if 2 doctors treated the same patient during the same visit. Thus the total number of interactions was the count of patient visits shared by each doctor dyad. This information was extracted from the data matrix illustrated in Table 1. We then used these interactions to build nondirectional weighted social networks. For each provider, the node degree was the number of other providers they interacted with at least once between 2016 and 2019. For each connected dyad, the weight of their connecting edge was calculated as the number of times 2 doctors shared a patient. The social networks included 200 000 interactions.

Social network metrics

Social network analysis metrics were used to assess the global characteristics of the network. Measures of centrality identified nodes with important roles in the network and greater access to other nodes. Centrality was calculated by degree centrality and was used to identify the influencers and providers with greater control over the flow of information in the network. In graph theory and network analysis, indicators of centrality identify the most important vertices within a graph. Degree centrality assigns an importance score based purely on the number of links held by each node.

Pearson's correlation coefficient was calculated between social network node degree and the average RF feature rank.

RESULTS

During the approximately 4-year study period, described in the study population in the Materials and Methods, there were 216 737 patients discharged from the study hospital, of whom approximately 31 300 (14.4%) subsequently completed and returned the HCAHPS survey. After downsampling, 11 483, 17 766 and 16 628 patient-surveys were used for the courtesy/respect, understand, and listen question, respectively (Figure 1).

Supplementary Table 5 shows all the features that were included in the machine learning models, aside from the individual providers, for the 3 unique questions stratified by response type. Patients responding negatively tended to skew older, had similar gender breakdown, had longer length of stay, were admitted more often through the emergency department, had a higher comorbidity burden (Charlson Comorbidity Index), and had similar pain scores.

The out-of-sample ROC curves of the 3 models to predict doctor ratings are shown in Figure 2. The RF model outperformed the LR

Table 2. Retrospective and prospective validation test performed using the random forest model for the 3 models

Parameters	Validation	Question 1 (courtesy and respect)	Question 2 (understand)	Question 3 (listen carefully)
AUC	Retrospective	0.876 (0.865-0.886)	0.819 (0.809-0.829)	0.819 (0.808-0.829)
	Prospective	0.874 (0.861-0.886)	0.767 (0.755-0.780)	0.781 (0.768-0.794)
Sensitivity	Retrospective	0.633	0.483	0.483
	Prospective	0.808	0.72	0.728
Specificity	Retrospective	0.981	0.988	0.991
	Prospective	0.759	0.652	0.659
PPV	Retrospective	0.96	0.968	0.975
	Prospective	0.762	0.649	0.664
NPV	Retrospective	0.782	0.715	0.717
	Prospective	0.805	0.723	0.723

Values are AUC (95% confidence interval).
AUC: area under the receiver-operating characteristic curve; NPV: negative predictive value; PPV: positive predictive value.

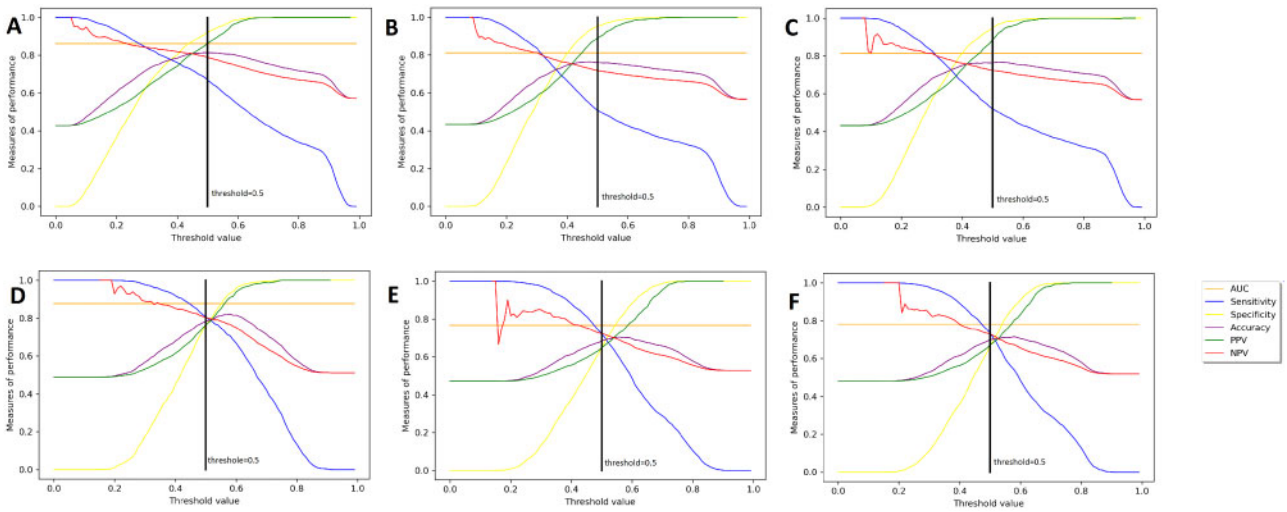

Figure 3. Metric optimizer revealing how the performance characteristics vary with threshold adjustments over a constant area under the receiver-operating characteristic curve (AUC) for each of the 3 models. This type of adjustment can be used in business decisions to find the right balance between false negative and false positive cases. (A) Retrospective of question 1: courtesy and respect; (B) retrospective of question 2: understanding; (C) retrospective of question 3: listen carefully; (D) prospective of question 1: courtesy and respect; (E) prospective of question 2: understanding; (F) prospective of question 3: listen carefully. NPV: negative predictive value; PPV: positive predictive value.

and decision tree models on all 3 questions. The RF model had the following performance. The response to the first question, "How often did doctors treat you with courtesy and respect?" (courtesy and respect), had an AUC of 0.876 (95% confidence interval [CI], 0.865-0.886); "How often did doctors explain things in a way that you could understand?" (understand) had an AUC of 0.819 (95% CI, 0.809-0.829); and "How often did doctors listen carefully to you?" (listen carefully) had an AUC 0.819 (95% CI, 0.808-0.829). Other performance metrics are shown in Table 2.

We further performed a prospective validation of the RF machine learning model. The predicted survey responses were compared with actual survey responses, which were received up to 6 weeks after hospital discharge. The performance is summarized here. The response to the first question, "How often did doctors treat you with courtesy and respect?" (courtesy and respect), had an AUC of 0.874 (95% CI, 0.861-0.886). The response to the second question, "How often did doctors explain things in a way that you could understand?" (understand), had an AUC of 0.767 (95% CI, 0.755-0.780), The response to the third question, "How often did doctors listen carefully to you?" (listen carefully), had an AUC of 0.781 (95% CI, 0.768-0.794). Other performance metrics are also shown in Table 2. The performance of the prospective validation for Questions 2 and 3 is slightly lower than the performance of the retrospective validation. This is probably due to the fact that some newer providers from the prospective validation test set may not have been included in the training data.

We also show a metric optimizer revealing how the performance characteristics vary with threshold adjustments over a constant AUC in Figure 3. In all of our models, the threshold is set to 0.5, as indicated by the black vertical line. It can be seen that there is a tradeoff between sensitivity and specificity. By moving the threshold to the left or to the right, we can increase one of these metrics at the cost of decreasing the other one. Similarly, there is a tradeoff between positive predictive value and negative predictive value. The top 3 subplots show the results of the retrospective validation, while the bottom 3 subplots show the results of the prospective validation. It can be seen on top plots that the threshold falls to the right of the crossover point and as a result the specificity is higher than the sensitivity. However, in the bottom plots, the

Table 3. Feature rank of input features by decrease in impurity and in accuracy based upon the random forest model for the 3 patient experience questions

	Question 1 (courtesy and respect)		Question 2 (understand)		Question 3 (listen carefully)	
	Rank decrease in impurity	Rank decrease in accuracy	Rank decrease in impurity	Rank decrease in accuracy	Rank decrease in impurity	Rank decrease in accuracy
Age	11	27	9	8	9	12
Gender	77	69	57	344	58	45
Race	35	38	33	45	35	16
Primary language	122	133	104	341	151	924
Marital status	34	18	19	154	20	161
Religion	18	26	26	32	24	37
Length of stay	15	10	13	10	14	9
Admit via ED	7	9	7	7	7	8
CCI	13	19	11	12	12	10
Number of allergies	59	34	51	254	62	238
Pain with activity						
Max	29	77	25	208	28	38
Average	31	99	28	24	29	20
Standard deviation	17	14	16	212	18	44
Difference between first and last record	25	416	23	170	21	40
Pain at rest						
Max	30	22	27	217	26	96
Average	32	66	24	37	30	19
Standard deviation	19	13	17	109	17	31
Difference between first and last record	23	198	22	62	23	160
Month	8	7	15	9	13	15
Previous Response	16	23	20	15	16	23
Temperature Recorded						
Max	2	1	2	4	5	1
Average	4	5	1	6	4	6
Standard deviation	5	6	3	2	3	2
Difference between first and last record	9	12	8	13	8	13
MEWS						
Max	1	3	4	5	1	3
Average	3	2	5	1	2	4
Standard deviation	6	4	6	3	6	5
Difference between first and last record	12	11	12	14	11	14
Location hours						
Inpatient	14	8	14	16	15	7
ED	10	16	10	11	10	11
ICU	49	65	32	361	36	411

CCI: Charlson Comorbidity Index; ED: emergency department; ICU: intensive care unit; MEWS: modified early warning score.

threshold falls to the left of the crossover point and this is why the specificity is lower than the sensitivity. If we wanted to increase the specificity of the prospective validation, this could be done by increasing the threshold.

Both Table 3 and Figure 4 show the RF feature rank by 2 methods for assessing variable importance: mean decrease in impurity and mean decrease in accuracy. The ranks of the top 15 features of the 3 models are shown in Figure 3, which reveal a mix of provider and patient-specific factors. Variables with low ranks along these 2 dimensions are most important for tree building and prediction. Because individual doctors correspond to specific features, their name labels were replaced with a label that indicates their specialty. Each doctor is their own point in the plot.

Social network analysis

The social network analysis metrics identify the global structure of the network, the influencers within the network and uncovers specialists that work more closely. The provider social network included 2463 doctors. Using shared patients to define provider interactions, we found that providers had a mean degree of 401, meaning that over the 4-year study period (described in the Materials and Methods in the study population subsection), doctors on average interacted with 401 other doctors (see Table 4).

The results in Table 4 and Figure 5 uncover the characteristics and interactions of subnetworks of specialists. Table 4 shows that obstetrics and gynecology doctors have the lowest degree, and Figure 5 shows that these providers have the fewest interactions with other specialties. Obstetricians and gynecologists mostly interact with surgeons. Table 4 shows that emergency medicine doctors have the highest degree, and Figure 5 shows that these providers tend to interact the most with providers in other specialties. Emergency medicine doctors interact with both internal medicine doctors and surgeons; however, the latter 2 do not interact as much with each other. Furthermore, anesthesiologists interact with all the specialties.

In order to test the hypothesis that within each network, doctors who are very connected are likely most influential on patient experi-

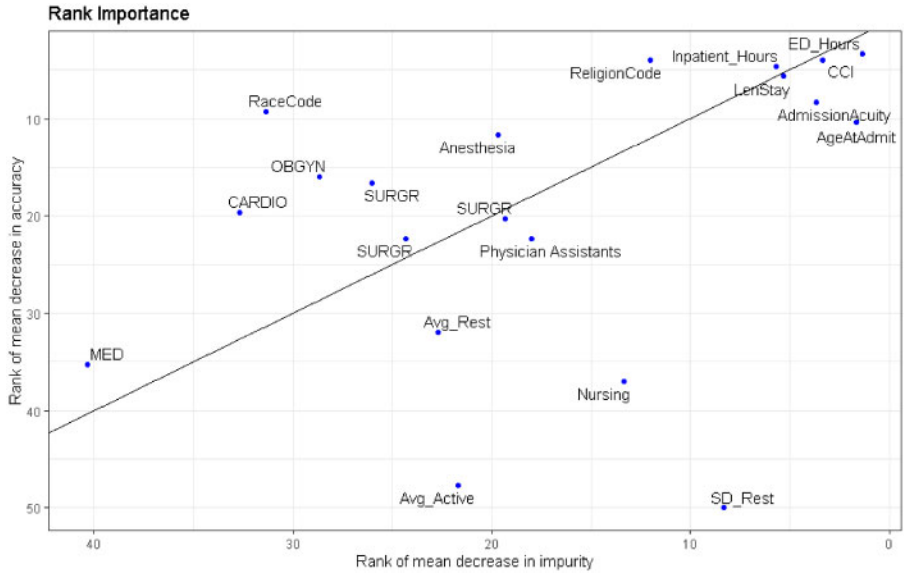

Figure 4. Average feature rank over all 3 models. Feature rank by mean decrease in impurity (Gini impurity) is plotted against feature rank by mean decrease in accuracy. Gini impurity is computed during model training, based on how much each feature decreases the weighted impurity in a tree. Then, the impurity decrease from each feature is averaged over all the trees in the forest. Finally, the features are ranked according to this average measure. Rank by mean decrease in accuracy is computed by scrambling each feature and then measuring how much this decreases the accuracy of the model. When low-ranking variables are scrambled, it has little or no effect on model accuracy, while scrambling high-ranking variables significantly decreases accuracy. AdmissionAcuity: binary flag indicating if patient was admitted through the emergency department vs electively; Avg_Active: average pain score during activity; CARDIO: cardiology; CCI: Charlson comorbidity index; ED_Hours: number of hours spent in the emergency department; Inpatient_Hours: number of hours spent in an inpatient unit; LenStay: overall hospital length of stay; MED: internal medicine; OBGYN: obstetrics/gynecology; SD_Rest: the standard deviation of the pain score at rest; SURGR: surgery.

Table 4. Degree of various populations of doctors

Population	Average degree
All doctors	401.1
Medicine doctors	410.9
Surgery Doctors	376.8
Emergency medicine doctors	491.9
Obstetrics/gynecology doctors	164.0
Anesthesia Doctors	443.7
Other doctors	433.8
Doctors with rank 1-50 as measured by decrease in accuracy of the random forest model	858.9

Providers who have higher degree interact more with other providers and have a larger influence on the patient experience.

ence, we examined the relationship between node degree and provider rank in the RF model. In Table 4 the top 50 ranked doctors by decrease in RF accuracy have a higher average degree than the average degree of all other subpopulations. Providers with higher node degree appear to have an outsized influence on patient experience, as measured by rank in the RF model. Thus, individuals who are very connected (higher node degree) are likely to hold the most information and appear to be most influential on patient experience, as shown by their rank in the RF model. Furthermore, social network node degree is correlated to the average RF feature rank with a correlation coefficient of –0.75 (see Figure 6).

DISCUSSION

Using an RF machine learning algorithm, we developed a patient experience-based stratification tool that can identify patients at risk for a negative in-hospital experience. The patient survey responses to 3 HCAHPS questions in the provider communication domain were predicted and validated on out-of-sample data retrospectively. Predictions made with a RF model outperformed predictions made with a LR model and decision tree model for all 3 survey questions. The RF model AUC ranged from 0.819 to 0.876. The predictions were also prospectively validated, with the AUC ranging from 0.874 to 0.78. The predictive model developed can be used as a basis to build a risk stratification tool to predict patient experience. Such a risk stratification tool could enable earlier evaluation and intervention (service recovery), thus mitigating the chances of negative survey responses.

The performance of the prospective validation for questions 2 and 3 is slightly lower than the performance of the retrospective validation due to the fact that some newer providers were not in included in the training data. This issue will be resolved in production by retraining the model on a daily basis.

Additional patient and clinical features, including additional demographics, comorbidities, and laboratory results, will likely improve the predictive performance, and will be explored for future versions. Future models may be improved further by binning providers into provider stereotypes, thus eliminating the need for individual providers to be features in the model and enhancing the overall model generalizability.

We have shown in our feature rank analysis that some individual providers have more impact on patient experience than others. The interconnectedness of providers (ie, their structural integration into the network) significantly influences their communication and interaction, and therefore holds valuable information for hospitals. We have also shown that these more impactful providers have higher network centrality than less impactful providers. Thus, providers

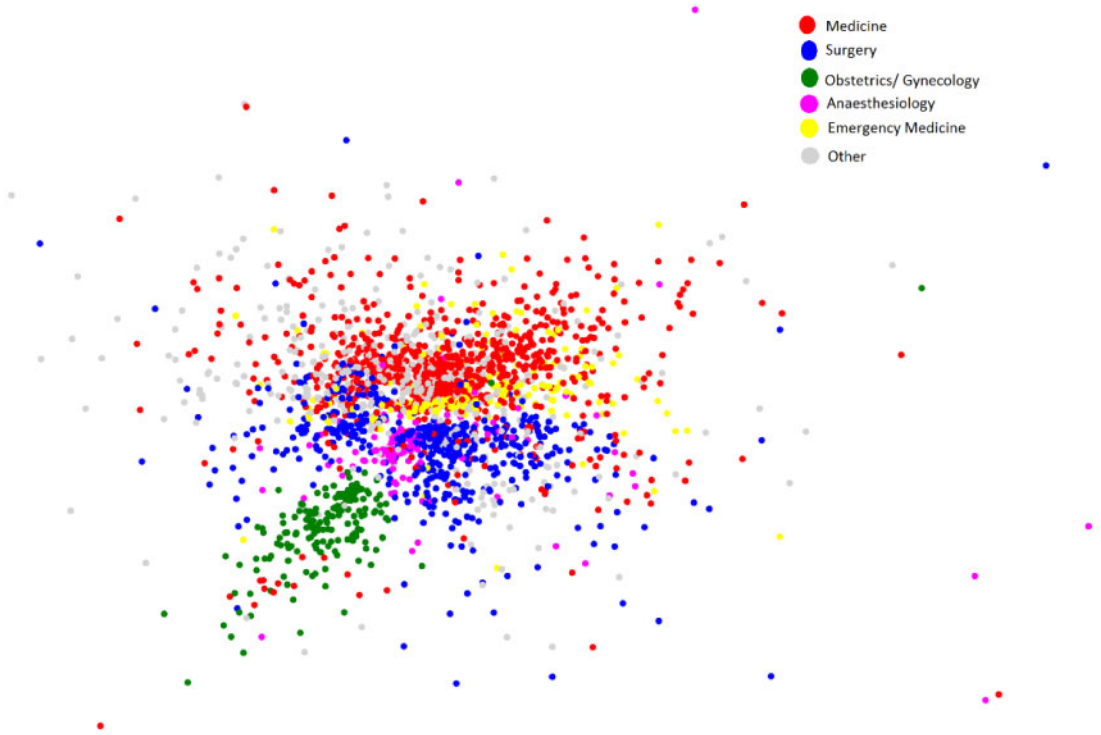

Figure 5. Social network analysis of provider interactions for a single hospital. Each point represents an individual provider and the distance between the points indicates how often the providers interact with one another. Providers who interact frequently are close together while providers who interact infrequently are far apart. Within each network, individuals who have more interactions are also more connected and more influential.

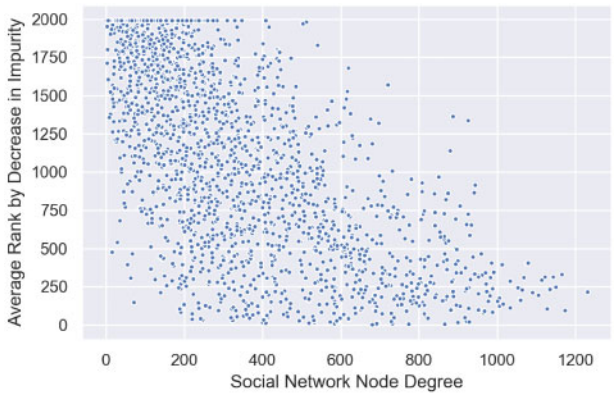

Figure 6. Correlation plot of social network node degree vs average random forest model feature rank. The correlation coefficient is −0.75.

who have more centrality in the social network also have more impact on patient experience, and may be targets for training or other interventions.

Limitations

While we attempted to obtain a large number of survey results over many years, only a small percentage of patients responded to the HCAHPS survey, and prior research has shown evidence of nonresponse bias in patient experience surveys.[25] Our research was also limited to a single institution, and additional research is needed to confirm our conclusions and the generalizability to other settings.

Furthermore, the provider social network analysis was based on the attribution of specific orders, documentation, or actions within the electronic health record. In a clinical operating environment, in which providers are heavily supported by house staff and advanced care providers (eg, residents, fellows, physician assistants, nurse practitioners), it is likely that the relationship, influence, or intensity of certain supervising providers (network nodes) will be underrepresented. The degree of underrepresentation will depend on how much a given provider delegates clinical ordering and documentation responsibilities to their supporting staff.

CONCLUSION

In this article, we propose a dual approach to improving patient experience, by concomitantly identifying patients at risk of having a negative experience and providers who are most influential on patient experience. This will enable hospitals to address patient experience from both the provider side and patient side. Patients who are at risk of a negative experience will receive appropriate interventions, while providers who are influential can undergo additional communication training. In the future, we plan on quantifying the business impact of the models in production via A/B testing. While prior studies using multivariate regression modeling identified factors that contribute to positive and negative patient experience,[26] and others have found an association between social media posts and HCAHPS survey responses,[27] this study is the first to propose an actionable model for predicting patient experience and provider influence on patient experience.

AUTHOR CONTRIBUTIONS

VB, JSH, and MM conceived and designed study. JN, RS, KRB, and MIO contributed materially to study design. VB, JN, RS, and MM were responsible

for data management. VB, MM, and JSH analyzed the data. VB, JSH, and MM wrote the article. VB, JSH, JN, RS, KRB, MIO, and MM reviewed the manuscript and contributed to revisions. All authors reviewed and interpreted the results and read and approved the final version.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Browne K, Roseman D, Shaller D, *et al*. Analysis and commentary measuring patient experience as a strategy for improving primary care. *Health Aff (Millwood)* 2010; 29 (5): 921–5.
2. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open* 2013; 3 (1): e001570.
3. Medicare.gov. Survey of patients' experiences (HCAHPS). https://www.medicare.gov/hospitalcompare/Data/Overview.html Accessed September 15, 2019.
4. Olivero W, Wang H, Vinson D, *et al*. Correlation between Press Ganey scores and quality outcomes from the national neurosurgery quality and outcomes database (lumbar spine) for a hospital employed neurosurgical practice. *Neurosurgery* 2018; 65 (CN_suppl_1): 34–6.
5. Zusman EE. HCAHPS replaces Press Ganey survey as quality measure for patient hospital experience. *Neurosurgery* 2012 ; 71 (2): N21–4.
6. Stelfox HT, Gandhi TK, Orav EJ, *et al*. The relation of patient satisfaction with complaints against physicians and malpractice lawsuits. *Am J Med* 2005; 118 (10): 1126–33.
7. Etier BE Jr, Orr SP, Antonetti J, *et al*. Factors impacting Press Ganey patient satisfaction scores in orthopedic surgery spine clinic. *Spine J* 2016; 16 (11): 1285–9.
8. Younesian S, Shirvani R, Tabatabaey A. Factors predicting patient satisfaction in the emergency department: a single-center study. *J Emerg Pract Trauma* 2017; 4 (1): 3–8.
9. Hall MF, Press I. Keys to patient satisfaction in the emergency department: results of a multiple facility study. *Hospital Health Serv Adm* 1996; 41 (4): 515–33.
10. Elliott MN, Lehrman WG, Goldstein E, Hambarsoomian K, Beckett MK, Giordano LA. Do hospitals rank differently on HCAHPS for different patient subgroups? *Med Care Res Rev* 2010; 67 (1): 56–73.
11. Stewart MA. Effective physician-patient communication and health outcomes: a review. *CMAJ* 1995; 152 (9): 1423–33.
12. Bull SA, Hu XH, Hunkeler EM, *et al*. Discontinuation of use and switching of antidepressants: influence of patient-physician communication. *JAMA* 2002; 288 (11): 1403–9.
13. Ciechanowski PS, Katon WJ, Russo JE, *et al*. The patient-provider relationship: attachment theory and adherence to treatment in diabetes. *Am J Psychiary* 2001; 158 (1): 29–35.
14. Bogardus ST Jr, Holmboe E, Jekel JF. Perils, pitfalls, and possibilities in talking about medical risk. *JAMA* 1991; 281 (11): 1037–41.
15. Giordano LA, Elliott MN, Goldstein E, *et al*. Development, implementation, and public reporting of the HCAHPS survey. *Med Care Res Rev* 2010; 67 (1): 27–37.
16. Baldwin DC Jr, Daugherty SR. How residents say they learn: a national multi- specialty survey of first-and second-year residents. *J Grad Med Educ* 2016; 8 (4): 631–9.
17. Sullivan P, Saatchi G, Younis I, Harris ML. Diffusion of knowledge and behaviors among trainee doctors in an acute medical unit and implications for quality improvement work: a mixed methods social network analysis. *BMJ* 2019; 9 (12): e027039.
18. Fabian P, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30.
19. Breiman L. Random forests. *Mach Learn* 2001; 45 (1): 5–32.
20. Duda RO, Hart PE, Stork DG. *Pattern Classification*. New York: Wiley; 2012.
21. Jiang R W, Tang X, Wu, *et al*. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics* 2009; 10 (S1): S65.
22. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn* 1997; 30 (7): 1145–59.
23. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10 (3): e0118432.
24. Ostovari M C-J, Steele-Morris PM, Griffin, *et al*. Data-driven modeling of diabetes care teams using social network analysis. *J Am Med Inform Assoc* 2019; 26 (10): 911–9.
25. Tyser AR, Abtahi AM, McFadden M, *et al*. Evidence of non-response bias in the Press-Ganey patient satisfaction survey. *BMC Health Serv Res* 2016; 16 (1): 350.
26. McFarland DC, Ornstein KA, Holcombe RF. Demographic factors and hospital size predict patient satisfaction variance—implications for hospital value-based purchasing. *J Hosp Med* 2015; 10 (8): 503–9.
27. Huppertz JW, Otto P. Predicting HCAHPS scores from hospitals' social media pages: a sentiment analysis. *Health Care Manag Rev* 2018; 43 (4): 359–67.