1 Introduction

In an increasingly volatile healthcare economy, politicians and healthcare service providers alike are struggling to balance a financially sustainable system of care with one that provides adequate service to its patients. This is especially challenging as "healthcare quality" comes with a complex set of methodological difficulties (Weheba et al., 2020). One imperative aspect of effective healthcare outcomes is surveyed patient satisfaction, as measured by a survey called the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS). The HCAHPS, which is updated four times a year, is highly influential in determining Medicare reimbursement rates and shaping individual hospital plans for staff training (Shah & Hoque, 2024). As our healthcare system moves further away from the damage done by the COVID-19 pandemic over the last four years, it's worth investigating the research question: What factors are most effective in predicting a state's HCAHPS patient satisfaction scores?

2 Background

2.1 HCAHPS

The HCAHPS assessment is a 29-question survey given to patients at over 3,000 hospitals in the United States on a quarterly basis to determine public satisfaction with hospital standards, ranging from service provider interactions to environment and cleanliness. All hospitals that treat Medicare patients are required to publicly disclose their HCAHPS reports or risk lowering the amount of pay they receive from Medicare (Giordano et al., 2009). States receive an aggregated score based on the reports from every hospital within that state. In theory, this score ranges from 0 to 5, with 0 being the poorest patient satisfaction and 5 being the best. In practice, the current lowest aggregate score is 2.17 in Washington, D.C., and the highest score is 4.22 in South Dakota. The influential nature of HCAHPS makes it imperative to investigate what factors can most effectively predict the statewide score. Table A1 in the appendix provides a list of each state's HCAHPS score.

2.2 Variable Selection

We chose 15 variables to include in this analysis based on existing scholarly literature that justifies a potential connection between the predictor variables and patient satisfaction. These variables fall into four broad categories: access to healthcare, healthcare spending, the provider environment, and miscellaneous factors.

2.2.1 Access to healthcare

Access to adequate healthcare is inextricably linked to patient satisfaction outcomes (Faezipour & Ferreira, 2013). Thus, it's no surprise why availability

of hospital beds, physicians, and nurses in relation to population size are all strongly linked to patient satisfaction rates (Xesfingi & Vozikis, 2016). Additionally, regional shortages of hospitals, especially in rural areas as opposed to urban centers, also have a major connection to healthcare consumers' perception of hospital systems (source). Studies have also found potential links between aging, healthcare access, and patient outcomes (source).

2.2.2 Healthcare spending

When people spend money on healthcare, they want to make sure that they are receiving a high quality service- or perhaps they spend money on healthcare because they are satisfied with the service they have received in the past. In either case, healthcare spending is tied to patient perceptions of hospitals, regardless of who is doing the spending. Patient satisfaction is linked with government healthcare spending, possession of healthcare insurance, and even salaries paid to physicians and nurse practitioners (source).

2.2.3 The provider environment

An increasingly contentious aspect of healthcare policy and patient satisfaction lies with the volatility in regulation for nurse practitioners. Different states have different regulations for how much diagnostic and prescriptive freedom nurse practitioners can exercise without the supervision of a physician. This has resulted in instability in the relationship between doctors and nurses, and various studies have found different links between nurse practitioners' scope of practice and patient outcomes, in both positive and negative directions (source).

2.3 Exploratory Analysis

In order to get a fundamental understanding of the potential correlation between the predictor variables and HCAHPS scores, we conducted preliminary research and analysis. Figure 1 below provides a visual representation of each state's HCAHPS score from the most recent survey, which was conducted on April 8th, 2024.

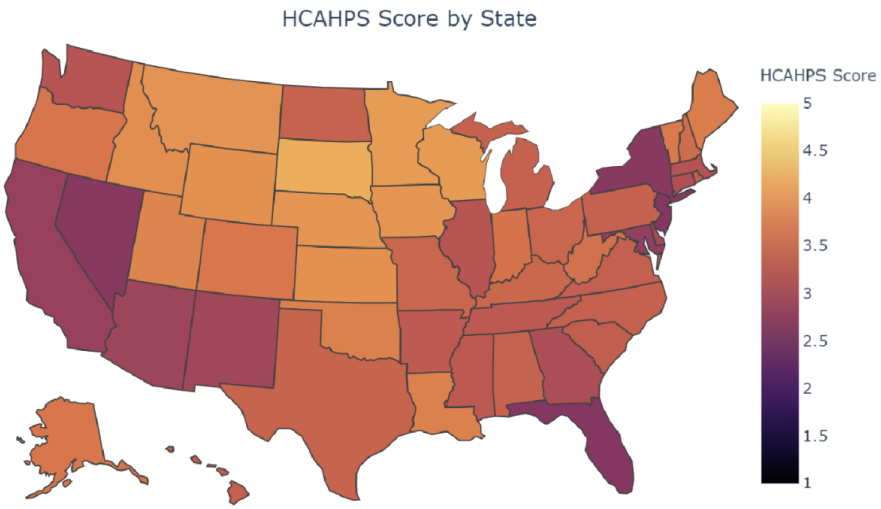

Figure 1: Heatmap of HCAHPS score by state

We investigated the individual correlations between each predictor variable and HCAHPS scores. The correlation coefficients between the predictor variables and HCAHPS scores are listed in Table 1 below.

Table 1: Correlation coefficients between predictor variables and HCAHPS scores

Predictor Variable	**Correlation Coefficient**	**p-value**
Urban Percent	-0.58536708	6.428×10^{-06} *
Hospitals Per Capita	0.48043376	0.0003596*
Average Nurse Practitioner Salary	-0.34990318	0.01184*
Medical Schools	-0.42531940	0.001863*
Healthcare Spending Per Capita	-0.15811835	0.2678
Staffed Beds Per Capita	-0.10888035	0.4469
Gross Patient Revenue	-0.38384860	0.005425*
Doctors Per Capita	-0.16142406	0.2578
Median Age	-0.05683453	0.692
Nurse Practitioners Per Capita	0.06042984	0.6736
Average Doctor Salary	-0.08890918	0.535
Healthcare Insured Population	-0.01912582	0.894
Nurse Practitioner Programs	-0.28973829	0.03918*
Restricted Practice Environment	-0.16634517	0.2434
Reduced Practice Environment	0.02921381	0.8387

Correlation coefficients that are statistically significant at $p < 0.05$ are marked with an asterisk.

3 Methodology

3.1 Variables

We investigated 15 predictor variables and one outcome variable in this analysis.

Hospitals per capita: The number of hospitals in every state, divided by the state's total population. This data was collected from the American Hospital Directory.

Median age: The median age in years of a state's total population. This data was found in the World Population Review.

Staffed beds per capita: The number of beds available for patient use in every state, divided by the state's total population. This data was collected from the American Hospital Directory.

Gross patient revenue (GPR): The total patient revenue reported by all hospitals in a state, measured in dollars. This data was collected from the American Hospital Directory.

Medical schools: The number of medical schools in a state. This data was found on a report from the website Medical School Headquarters.

Nurse practitioner programs: The number of nurse practitioner programs in a state. This data was collected from a report on a website called Nursing Process.

Physicians per capita: The number of practicing, licensed physicians in every state, divided by the state's total population. This information was collected from the Bureau of Labor Statistics.

Nurse practitioners per capita: The number of practicing, licensed nurse practitioners in every state, divided by the state's total population. This information was collected from the Bureau of Labor Statistics.

Spending per capita: The amount of money the state spends on healthcare annually, divided by the state's total population, measured in dollars. This data was collected from the Kaiser Family Foundation.

Average physician salary: The average salary of physicians in a state, measured in dollars. This information was found in a Becker's Hospital Review report.

Average nurse practitioner salary: The average salary of nurse practitioners in a state, measured in dollars. This data was collected from an IntelyCare report.

Healthcare insured population: The percent of a state's population that has healthcare insurance. This data was collected from the Kaiser Family Foundation.

Urban percent: The percentage of a state's population that lives in an urban rather than rural environment. This information was found on a report from Visual Capitalist.

Practice environment: The dependent variable in this analysis is the practice environment for nurse practitioners in each state. The American Academy of Nurse Practitioners defines every state as having a "full", "reduced", or "restricted" regulatory environment. This was encoded into two dummy variables. For "dummyrest", the restricted practice environments were coded as 1, and the reduced and full environments were coded as 0. For "dummyred", the reduced practice environments were coded as 1, and the restricted and full environments were coded as 0. The full practice environment was used as the reference group. The data was collected from the American Academy of Nurse Practitioners.

HCAHPS: The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) is an annual state by state tool used to measure patient satisfaction with hospitals. Every state receives a score between 0 and 5, with 0 being the poorest patient satisfaction and 5 being the highest patient satisfaction. This was the outcome variable in this analysis.

For "per capita" measurements, population data by state was collected from a report by Brittanica.

3.2 Model Fitting

We utilized four types of models to predict the HCAHPS score of each state. First, we used all 15 predictor variables to construct a multiple linear regression model. This model serves as a "baseline" to which we can compare the other models, and is also the most easily interpretable. Next, we created a Lasso regression model in order to combat overfitting and to create a sparse model in which the coefficients of variables with less predictive power are shrunk to zero. Thirdly, we used all 15 predictor variables to train a random forest model, in order to account for potential nonlinear relationships between predictors and HCAHPS scores and to control for any interactions between coefficients. Finally,

we used Recursive Feature Elimination (RFE) to recursively eliminate features with low predictive power and built a regression model using the coefficients for the 11 predictor variables that were selected by the RFE. To compare and evaluate the performance of the four models, we used a 10-fold cross validation. The models will be compared using the Root Mean Squared Error (RMSE) and R2 value. The RMSE will allow us to identify the model that optimizes model complexity, and the R2 value will determine the model that best fits the data.

The regression models will be constructed using the following "baseline" equation.

$$hcahps = a_0 + (a_1)hospitals - (a_2)age - (a_3)beds - (a_4)gpr - (a_5)medschools + (a_6)npprograms - (a_7)doctors - (a_8)nurses - (a_9)spending - (a_{10})doctorsalary - (a_{11})nursesalary + (a_{12})insured - (a_{13})urban - (a_{14})dummyrestricted - (a_{15})dummyreduced + \varepsilon \quad (1)$$

4 Results

4.1 Multiple Linear Regression

We used all 15 predictor variables to construct a basic linear regression model based on the baseline equation. Table 2 below shows the coefficients and p-values of the predictors.

Table 2: Coefficients and p-values of linear regression model

Variable	Coefficient	p-value
(Intercept)	-2.517e-02	0.991303
hospitalspercap	6.674e+04	0.000501*
medianage	-1.953e-02	0.384817
staffedbedspercap	-3.376e+02	0.053480*
gpr	-1.649e-10	0.861443
medschools	-7.257e-02	0.028401*
npprograms	4.388e-02	0.004564*
docspercap	-9.343e+01	0.262790
npspercap	-1.445e+02	0.499987
spendpercap	-4.483e-05	0.263822
avgdocsalary	-1.391e-06	0.326249
avgnpsalary	-3.250e-06	0.614131
hcinsuredpop	6.893e-02	0.006578*
urbanpercent	-1.142e-02	0.022645*
dummyrest	-1.501e-01	0.359086
dummyred	-4.623e-01	0.005097*

Seven of the 15 predictor variables were statistically significant at $p < 0.05$. These variables are marked with an asterisk. The full list of statistics for the variables can be found in Table A3 in the appendix.

4.2 Lasso Regression

We utilized a Lasso regression model to remove any variables that did not play a significant role in predicting the HCAHPS score of a given state. The Lasso model eliminated gross patient revenue and retained the other 14 predictor variables, resulting in the following coefficients in Table 3.

Table 3: Coefficients of Lasso regression model

Variable	Coefficient
(Intercept)	1.098687e+00
hospitalspercap	5.575790e+04
medianage	-1.411521e-02
staffedbedspercap	-3.485709e+02
gpr	0
medschools	-3.412103e-02
npprograms	1.771710e-02
docspercap	-8.287415e+01
npspercap	-2.324536e+01
spendpercap	-2.892382e-05
avgdocsalary	-1.863002e-07
avgnpsalary	-1.988614e-06
hcinsuredpop	4.833523e-02
urbanpercent	-1.077416e-02
dummyrest	-4.946403e-02
dummyred	-2.067782e-01

The model was constructed by identifying the lambda value with the lowest root mean squared error using 10-fold cross validation. This optimal lambda value was 0.01109091. This indicates a weak regularization effect. This aligns with the fact that the model only eliminated one of the 15 predictor variables. The p-values are not reported because Lasso regression already incorporates feature selection for the whole model and is not meant for hypothesis testing at the level of individual variables.

4.3 Random Forest

In order to account for any potential nonlinear correlations between the predictor variables and the HCAHPS score, interactions between the predictors, and to identify which variables played the most significant roles in predicting the score, we trained a random forest model. Table 4 below shows the importance scores assigned to each variable in the model.

Table 4: Random forest variable importance scores

Variable Name	Importance
Urban Percent	4.72390
Hospitals Per Capita	1.00297
Average Nurse Practitioner Salary	0.72792
Medical Schools	0.45063
Healthcare Spending Per Capita	0.44928
Staffed Beds Per Capita	0.42673
Gross Patient Revenue	0.39192
Doctors Per Capita	0.28572
Median Age	0.27572
Nurse Practitioners Per Capita	0.26203
Average Doctor Salary	0.24881
Healthcare Insured Population	0.15123
Nurse Practitioner Programs	0.08186
Restricted Practice Environment	0.01203
Reduced Practice Environment	0.01137

The importance scores quantify the mean decrease in accuracy of the model when a given variable is not utilized to predict the outcome variable. A higher importance score means a higher mean decrease in accuracy when the variable is not used, so higher scores mean that the variable has higher predictive power. This indicates that, when using the random forest model, the percentage of a state's population that lives in urban areas and the number of hospitals per capita are the factors that contribute most to model accuracy.

4.4 Recursive Feature Elimination (RFE)

We constructed an RFE model to produce a regression equation that had fewer predictors in order to avoid overfitting and identify which predictors were most effective in predicting the HCAHPS score of a given state. It does this by using 10-fold cross validation to train a series of linear models and eliminate the variables with the lowest coefficients. The RFE model selected 11 out of the 15 predictor variables. Gross patient revenue, healthcare spending per capita, average doctor salary, and average nurse practitioner salary were eliminated by the model. The analysis resulted in the coefficients in Table 5 below.

Table 5: Coefficients and p-values of model derived from RFE analysis

Variables	Coefficient	p-value
(Intercept)	6.980e-01	0.739252
hospitalspercap	6.379e+04	0.000407*
staffedbedspercap	-3.789e+02	0.011950*
npspercap	-1.289e+02	0.526920
docspercap	-1.392e+02	0.070111*
dummyred	-3.460e-01	0.020567*
dummyrest	-5.612e-02	0.696573
hcinsuredpop	5.316e-02	0.012806*
medschools	-7.072e-02	0.016361*
npprograms	3.696e-02	0.005525*
medianage	-2.240e-02	0.300033
urbanpercent	-1.277e-02	0.003931*

Eight of the 11 selected predictor variables were statistically significant at p < 0.05. These variables are marked with an asterisk. The full list of statistics for the variables can be found in Table A4 in the appendix.

4.5 Cross Validation

We used a 10-fold cross validation to compare and evaluate the models. The root mean squared error (RMSE) and R2 value of all four models are stated in Table 6 below. The detailed descriptive statistics of the model evaluation tools are included in Table A5 in the appendix.

Table 6: Evaluation of models to predict HCAHPS score

Model	RMSE	R^2
Linear regression	0.3706752	0.4580556
Lasso regression	0.3193341	0.5798507*
Random forest	0.3280300	0.4741711
RFE	0.3106873*	0.5288606

Lower RMSE indicates a model that best balances simplicity with complexity in order to ensure accurate prediction without overfitting. Higher R2 indicates a model that best fits the existing data. Thus, the RFE model was the best-performing model using RMSE and the Lasso regression was the best-performing model using R2.

5 Conclusion

The results of this analysis indicate that the Lasso model best fits the existing data, and the RFE model has the best predictive power for unknown HCAHPS scores. It is beyond the scope of this paper to claim whether there is a causal relationship between the predictor variables and patient satisfaction, but we find that the number of hospitals per capita, staffed beds per capita, physicians per capita, urban population, number of medical schools, healthcare insured population, and reduced practice environment are the most powerful predictors of patient satisfaction, given the most recent data. A particularly interesting result is that gross patient revenue was eliminated from both models, meaning that it likely has very weak power when attempting to predict patient satisfaction. Observing these trends highlights some interesting implications and avenues for potential further investigation.

5.1 COVID-19

Immediately following the beginning of the COVID-19 pandemic, HCAHPS scores (unsurprisingly) declined dramatically (source). However, after 2021, the scores began to rise again as hospitals began to adapt to the crisis and adopted strategies to optimize patient experiences in the far from ideal circumstance (source). Considering that HCAHPS scores took a sharp detour for two years before settling back on track, potential future analysis could examine the effect of these predictor variables prior to the COVID-19 pandemic in order to discover whether they predicted patient satisfaction outcomes in the same way that they do in the present. Understanding how the roles of these predictors have changed over time could be key to understanding whether our healthcare system is returning to the same methods it utilized prior to the pandemic, or adopting a new paradigm of care entirely.

5.2 Local Measurement

A given state is not a monolith of healthcare conditions; different counties within a state have vastly different experiences with access to high quality healthcare (source). Surveying patient satisfaction on a county level will likely yield a broad range of results that cannot be captured by a statewide statistic; including the urban vs. rural population in the analysis in this paper is only the first step towards a more comprehensive understanding of the variations in the patient experience at more local levels. A potential future analysis could utilize county level data in order to have a more thorough understanding of variations and predictions for patient satisfaction. Additionally, HCAHPS is surveyed at each hospital before being aggregated by county and state; another potential modification for future analysis could utilize measurements at the hospital level,

rather than the state or county level.

As our healthcare system continues to adjust to a post-COVID sense of normalcy, we hope that this analysis will be beneficial in determining what factors can predict a more effective and high quality patient experience.

6 References

7 Appendix

Table A1: HCAHPS scores by state

State	HCAHPS
Alabama	3.40
Alaska	3.67
Arizona	2.88
Arkansas	3.26
California	2.82
Colorado	3.67
Connecticut	3.08
DC	2.17
Delaware	3.00
Florida	2.61
Georgia	3.05
Hawaii	3.36
Illinois	3.19
Indiana	3.61
Iowa	3.98
Kansas	3.93
Kentucky	3.46
Louisiana	3.76
Maine	3.72
Maryland	2.78
Michigan	3.37
Minnesota	4.04
Mississippi	3.25
Missouri	3.44
Montana	3.95
Nebraska	3.97
Nevada	2.64
New Hampshire	3.52
New Jersey	2.61
New Mexico	2.93
New York	2.64
North Carolina	3.37
North Dakota	3.40
Oklahoma	3.76
Ohio	3.44
Oregon	3.65
Pennsylvania	3.39
Rhode Island	3.36
South Carolina	3.33
South Dakota	4.22
Tennessee	3.25
Texas	3.42
Utah	3.81
Vermont	3.69
Virginia	3.35
Washington	3.18
West Virginia	3.58
Wisconsin	4.03
Wyoming	3.93

Table A2: p-values of linear regression model

Variable	Coefficient	Std. Error	t-value	p-value
(Intercept)	-2.517e-02	2.292e+00	-0.011	0.991303
hospitalspercap	6.674e+04	1.740e+04	3.836	0.000501
medianage	-1.953e-02	2.220e-02	-0.880	0.384817
staffedbedspercap	-3.376e+02	1.689e+02	-1.998	0.053480
gpr	-1.649e-10	9.378e-10	-0.176	0.861443
medschools	-7.257e-02	3.174e-02	-2.286	0.028401
npprograms	4.388e-02	1.448e-02	3.031	0.004564
docspercap	-9.343e+01	8.209e+01	-1.138	0.262790
npspercap	-1.445e+02	2.120e+02	-0.682	0.499987
spendpercap	-4.483e-05	3.948e-05	-1.136	0.263822
avgdocsalary	-1.391e-06	1.397e-06	-0.996	0.326249
avgnpsalary	-3.250e-06	6.388e-06	-0.509	0.614131
hcinsuredpop	6.893e-02	2.385e-02	2.890	0.006578
urbanpercent	-1.142e-02	4.788e-03	-2.385	0.022645
dummyrest	-1.501e-01	1.615e-01	-0.929	0.359086
dummyred	-4.623e-01	1.547e-01	-2.989	0.005097

Multiple R-squared: 0.731, Adjusted R-squared: 0.6157
F-statistic: 6.34 on 15 and 35 DF, p-value: 3.467e-06

Table A3: p-values of RFE model

Variable	Coefficient	Std. Error	t-value	p-value
(Intercept)	6.980e-01	2.082e+00	0.335	0.739252
hospitalspercap	6.379e+04	1.650e+04	3.867	0.000407
staffedbedspercap	-3.789e+02	1.437e+02	-2.637	0.011950
npspercap	-1.289e+02	2.020e+02	-0.638	0.526920
docspercap	-1.392e+02	7.475e+01	-1.862	0.070111
dummyred	-3.460e-01	1.433e-01	-2.414	0.020567
dummyrest	-5.612e-02	1.428e-01	-0.393	0.696573
hcinsuredpop	5.316e-02	2.037e-02	2.609	0.012806
medschools	-7.072e-02	2.818e-02	-2.509	0.016361
npprograms	3.696e-02	1.258e-02	2.938	0.005525
medianage	-2.240e-02	2.133e-02	-1.050	0.300033
urbanpercent	-1.277e-02	4.164e-03	-3.066	0.003931

Multiple R-squared: 0.6998, Adjusted R-squared: 0.6151
F-statistic: 8.264 on 11 and 39 DF, p-value: 3.126e-07

Table A4: Descriptive statistics of evaluation tools

RMSE

Model	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Linear	0.1958999	0.2927375	0.3537815	0.3706752	0.4556817	0.5310463
Lasso	0.1776066	0.2615233	0.2957125	0.3193341	0.3696481	0.4830503
Random Forest	0.1011962	0.3084216	0.3367423	0.3280300	0.3870528	0.4493962
RFE	0.2374771	0.2643863	0.2948705	0.3106873	0.3504836	0.4272403

R^2

Model	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Linear	0.0001480487	0.3401612	0.4832217	0.4580556	0.6516408	0.6968737
Lasso	0.0445451439	0.5171739	0.6032024	0.5798507	0.7421082	0.9242257
Random Forest	0.0003427442	0.1256972	0.4137896	0.4741711	0.8459433	0.9415322
RFE	0.0537718481	0.5024046	0.5689611	0.5288606	0.6735165	0.8560477