Healthcare Quality under QHP Individual Market Analysis

By Brendon Hahm

Background

Health Plans

Insurance plans are divided into three different markets: Individual, small group, and large group markets. Individual markets pertain to policyholders that are directly purchasing insurance from the insurer as a single entity. Small group markets pertain to businesses for 50 or fewer policyholders. Large group markets pertain to businesses with more than 50 policyholders. Group health insurance is divided as such because the risk of a single policyholder is spread more thinly across the rest of the insurance pool and every new policyholder that does not actualize costs, implies cheaper overall costs for the rest of the pool.

Qualified Health Plans

Qualified health plans (QHP) were defined under the ACA separately for each of the three groups and are certified by the Health Insurance Marketplace. Upon certification, insurance plans can participate in the Federally-Facilitated Marketplace (FFM or The Marketplace) where health plans meeting requirements for coverage, cost sharing, and other areas are available to consumers. The FFM and QHPs essentially allows consumers to centrally locate health plans to meet baseline quality and benefit coverage. One of the most central requirements of QHP is a minimum set of covered benefits such as ambulatory patient services, emergency services, mental health, preventive health services, and many other benefits. These minimum required benefits are called essential health benefits (EHB) and all plans in the FFM must meet or exceed these requirements.

Medical Loss Ratio

Efficiency in healthcare is a large concern for both consumers, insurers, and the government. However, an efficiency metric that is of particular interest for consumers and the government is the medical loss ratio (MLR). Conventionally, MLR is understood as the ratio of health care claims over premiums. This ratio shows the degree to which a health plan's revenue from premiums are being utilized towards their insurance pool's health care and services. From the perspective of the insurer, MLR measures the profitability of a health plan, as all other revenues from the plan go towards overhead costs.

Given the interest of the government in health plan efficiencies, the ACA established requirements of MLR for health plans. However, the formulation of MLR under the ACA is slightly different from the conventional formulation where the numerator includes both health care claims and quality improvement expenses while the denominator consists of premiums subtracted by taxes, licensing, and regulation fees. From here on, we will use the ACA formulation of MLR. The differences in the formulations are shown below:

$$MLR = \frac{health\ care\ claims}{revenue\ from\ premiums}$$

$$ACA\ MLR = \frac{health\ care\ claims + quality\ improvement\ expenses}{revenue\ from\ premiums - taxes,\ licensing,\ and\ regulation\ fees}$$

Basic interpretations of this ratio are that lower MLR implies higher profitability for the insurer and a lower proportion of premiums going towards medical care and services. A higher MLR implies less profitability for the insurer and a higher proportion of premiums going towards medical care and services. Notably, a MLR of greater than one implies a plan that is not profitable and runs at a profit net loss. This scenario is entirely possible due market strategies by insurers for temporary market share gains. However, since these plans are not self sufficient, these typically are temporary and not sustainable.

Under the ACA, MLR requirements are typically enforced at .8 for individual and small group plans and .85 for large group plans. However, there are rate review provisions that may impose even greater MLR minimums given premium rate increases of a plan. Failure to meet MLR requirements lead to required rebates proportional to the amount in premiums paid for each policyholder. States may require a minimum MLR higher than what is federally mandated but cannot lower any MLR.

Insurers must also report the MLR and the variables that constitute the ratio annually at an aggregated state level. While the reporting of this information is a key insight to the quality and sustainability of health plans offered, the enforcement of these reporting requirements vary by state and are often unenforced for most plans in the individual and small market groups.

Quality Metrics

Along with coverage of benefits, quality of care is one of the most important questions facing how we provide and distribute care. However, the quality of care is both an extremely important and difficult measurement to define and measure. Especially with the shifting of focus towards preventive care and many other evolutions in healthcare, our metrics of quality must be holistic, well defined, and patient centered.

The ACA takes three quality ratings from QHP insurers and calculates the overall rating based on these quality ratings. 34 quality measures, with 24 being clinical and 10 being surveyed from members of the QHP, are used for the three quality ratings: Medical Care, Member Experience, and Plan Administration. While the measurements taken are not publicly available, the quality ratings are available for all QHP's on the FFM.

Project Overview

Introduction

The market for medical care is unique in that uncertainty runs rampant in nearly every respect of its provision. This fundamental uncertainty mixed with the high cost nature of medical care brings about the most clear solution, insurance. Insurance allows pools of people to diversify their risk and be affordable for an individual while covering the costs across the insurance pool. However, the fundamental uncertainty of medical care is often argued to be the very reason medical care can not be provisioned purely by private means as argued by Kenneth Arrow in *Uncertainty and Welfare Economics of Medical Care*. This along with the adverse incentives of many necessary actors in the healthcare system leads to what we have now, a highly regulated and complicated insurance model for providing health care. No other part of healthcare does this uncertainty matter more than in the provision of health insurance.

In this project, I explore data on qualified health plans for the individual market offered on the FFM that have been made publicly available by Healthcare.gov. The totality of the data used is both large, but incomplete so analysis is limited in multiple regards that are laid out later in this report. This project can be broken down into three parts. The first part explores the relationships between essential health benefit coverage, MLR, and quality metrics. These are defined and regulated to ensure minimum quality and efficiency in health plan offerings on the FFM. Without significant relationships, there is no reason for the FFM and no reason for QHP's as defined by the ACA.

The second part of the project uses the data on QHP's to create a predictive model for quality. Quality is an essential metric for insurers on the FFM due to its important role in marketability. Quality scores are displayed on the FFM to aid consumers in finding the best, affordable, and quality health plan for them. The third and final part of the project examines data available from 2019 to 2022 and identifies trends and performance of insurers and states across the time period.

Data Overview

Ever since the creation of the FFM, data and information on qualified health plans have been slowly made publically available on Healthcare.gov. The data used in this project range from 2019 to 2022 and are divided into three groups. General information on QHP's offered in a given year are provided as Landscape datasets in xlsx format. These datasets offer information on basic plan information, EHB percent of total premium, and premium scenarios. The other kind of datasets offered for QHP's are public use files (PUF) that are publicly available datasets on various metrics and reporting for QHP's on the FFM. Notably, participation in the reporting of these PUF's range greatly by state and are often not reported especially for the individual market. The PUF's used in this project are the quality metrics for 2019-2022 and the MLR dataset for 2019. The usage of various datasets are summarized by the following table:

Year	QHP Individual Landscape	Quality PUF	MLR
2019	x	x	x
2020	x	x	
2021	x	x	
2022	x	x	

Missing Values

As mentioned before, the reporting and requirements of participation in PUF's vary greatly. So it must be mentioned that since much of this project required combining the QHP datasets to the PUF datasets, the many missing values for certain columns that arose from finding the intersection of these data was inevitable. In trying to preserve as much data as possible, a significant amount of columns or variables were not considered in the project due to a high frequency of missing values that resulted from the necessary merging of data required for the analysis and work done in the project.

The key notes about missing values for each of the three parts of project are as follows:

- EHB, MLR, and Quality Analysis
    - Only 2019 health plans that contain data on all three variables, where each come from different datasets, are considered
    - Missing values are dropped vertically for columns with a high degree of missing values then horizontally rows with any missing values
- Predictive Modeling
    - Only 2023 health plans that have data in the QHP Individual Landscape and Quality PUF datasets are considered
    - Missing values are dropped vertically for columns with a high degree of missing values then horizontally rows with any missing values
- General Trends
    - Only health plans that have data in the QHP Individual Landscape and Quality PUF datasets with data from each year 2019-2023 are considered
    - Missing values are dropped vertically for columns with a high degree of missing values then horizontally rows with any missing values

QHP Individual Landscape

The QHP Individual Landscape data records all the qualified health plans on the FFM in the individual market. These datasets record mostly information that is helpful for consumers to either gain a general idea of the plan or to find resources for further information. Notably the dataset also offers other variables for premium scenarios and various levels of actuarial values for silver plans. These variables were problematic since they had many missing values for most QHP's that weren't silver metal tier which affected any predictive modeling or analysis. They were also such specific and granular aspects for plans that including them would lose generality and increase the dimensionality of data to too much of a degree.

Public Use Files

The quality PUF datasets are simple datasets and provided quality metrics in four categories:

1) Overall Rating
2) Medical Care Rating
3) Member Experience Rating
4) Plan Administration Rating

The overall rating is determined formulaically by the other three ratings. The quality metrics correspond to unique health plan identification numbers which were used to merge the datasets.

The MLR data offered on healthcare.gov is only offered up until 2019. This meant that the only year that intersected for all three dataset types was 2019 so only 2019 was used for MLR. The MLR dataset is different from the other two datasets since the data is reported by the insurer aggregated at a state level as opposed to the health plan level like the other two datasets. This consideration was important for the analysis later on.

States included

The states included in all analysis were 39 states consisting of:

Alaska, Alabama, Arizona, Arkansas, Delaware, Florida, Georgia, Hawaii, Iowa, Illinois, Indiana, Kansas, Kentucky, Louisiana, Maine, Missouri, Michigan, Mississippi, Montana, North Carolina, North Dakota, Nebraska, New Hampshire, New Jersey, New Mexico, Nevada, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, South Dakota, Tennessee, Texas, Utah, Virginia, Wisconsin, West Virginia, and Wyoming.

Notably, QHP's from larger population states such as California, New York, and some other smaller population states are not included.

EHB, MLR, and Quality Analysis

Data at Health Plan Level Results

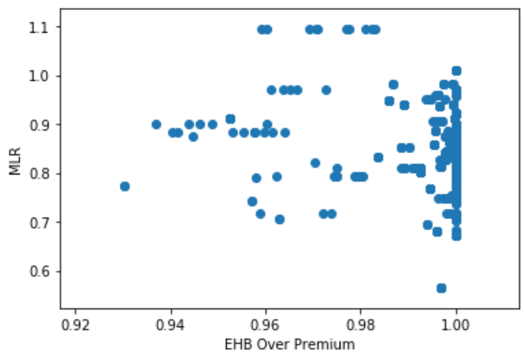
To first make initial analysis possible, we found the intersection of the QHP and quality datasets. Then to combine with the MLR dataset, we assigned the MLR associated with each insurer id for each state to each health plan. The data can be summarized as such:

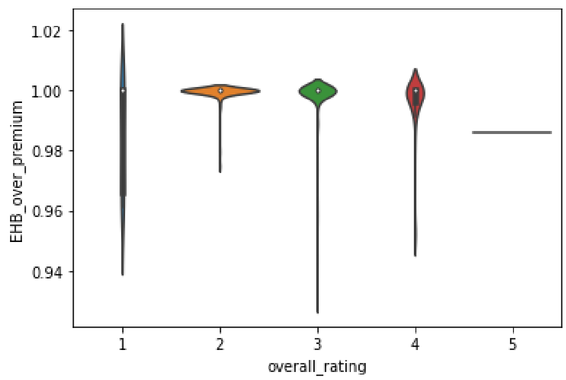
- EHB over Premium
    - Health plan level
- Quality
    - Health plan level
- MLR
    - Insurer per state level
    - Aggregated by total ratio
    - Assigned to each health plan based on insurer and state

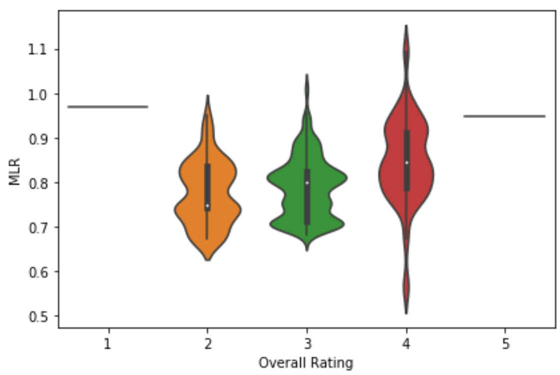
The first test to measure the relationship between EHB, MLR, and Quality Metrics was a simple Pearson correlation coefficient estimate and significance testing. The results are found here:

	r coefficient	Two tailed p-value
EHB_over_premium & mlr	-0.285	1.2836586230781855e-33
EHB_over_premium & overall_rating	-0.149	4.1064244210145736e-10
Overall_rating & mlr	0.252	1.9243885086677498e-26

Interestingly across all three variables we see small coefficient values but highly significance given by the p-values. At 99% significance level, all three correlations show significance. Visualizations of the relationships are shown here:







A few key takeaways from the visualizations reveal that individual market QHP's at either extremes of overall quality seem to have very tight distributions with respect to MLR. We can also see tight distribution for EHB over premiums with overall rating 5 but a very wide distribution of plans with overall rating 1. So we can see that plans with high overall ratings tend to have high MLR and lower EHB over premiums. This makes sense since high MLR means more of the premium goes towards actual medical care and lower EHB over premiums means the plan goes beyond the minimum required EHB's. However, it is important to note that MLR tends to be high with a tight distribution for plans with a 1 overall rating.

Regression analysis on the relationship of EHB and MLR with quality was also performed with the following results:

OLS Regression Results

Dep. Variable:	overall_rating	**R-squared:**	0.080
Model:	OLS	**Adj. R-squared:**	0.077
Method:	Least Squares	**F-statistic:**	29.90
Date:	Mon, 17 Apr 2023	**Prob (F-statistic):**	3.23e-29
Time:	22:51:24	**Log-Likelihood:**	-1622.6
No. Observations:	1732	**AIC:**	3257.
Df Residuals:	1726	**BIC:**	3290.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6784	2.692	0.623	0.533	-3.602	6.959
adult_dental	0.2994	0.132	2.269	0.023	0.041	0.558
child_dental	-0.0049	0.035	-0.141	0.888	-0.073	0.063
EHB_over_premium	-0.8708	2.640	-0.330	0.742	-6.048	4.306
mlr	2.7169	0.328	8.293	0.000	2.074	3.359
average_rebate	0.0003	8.91e-05	3.343	0.001	0.000	0.000

Omnibus:	115.253	**Durbin-Watson:**	0.143
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	190.484
Skew:	-0.507	**Prob(JB):**	4.33e-42
Kurtosis:	4.269	**Cond. No.**	9.37e+04

We can see that the model found significant coefficients for MLRbut not EHB over premium. This shows that at a health plan level, the aggregated MLR still has a significant effect which is an interesting result. It is also interesting to see that coverage beyond essential health benefits shows no significant effect on quality.

It is also notable to mention that we would expect that higher MLR and lower EHB lead to higher quality, however such relationships are only indicated by the parity of the correlation coefficients. The absolute values of the magnitude for both correlations are less than 0.3. It is possible that the relationships are nonlinear but it is also possible that the level at which the data should be examined is at the insurer by state aggregate level as we have for MLR. This approach is explored next.

Data at Aggregated Level Results

For these results, we also merge the QHP and quality datasets initially but aggregate the data at the insurer level by state. Then we merge the dataset with the MLR dataset which is already at the same level. This approach is much more intuitive and in theory should be a more sound approach as the data types are more similar. The data can be summarized as follows:

- EHB over Premium
  - Insurer per state level
  - Aggregated by mean
- Quality
  - Insurer per state level
  - Aggregated by mean
- MLR
  - Insurer per state level
  - Aggregated by total ratio

By merging the data like this, it keeps all data aggregated at the same level. The first test to measure the relationship between EHB, MLR, and Quality Metrics was a simple pearson correlation coefficient estimate and significance testing. The results are found here:

	r coefficient	Two tailed p-value
EHB_over_premium & mlr	-0.253	0.01336
EHB_over_premium & overall_rating	-0.1357	0.1898
Overall_rating & mlr	0.03	0.772

Here we can see that when we aggregate the data instead, the only correlation that remains significant is between EHB over premium and MLR at a 95% significance level. But neither are significantly correlated with quality.

Regression analysis on the insurer by state aggregated data also showed interesting results:

OLS Regression Results

Dep. Variable:	overall_rating	**R-squared:**	0.074
Model:	OLS	**Adj. R-squared:**	0.022
Method:	Least Squares	**F-statistic:**	1.421
Date:	Mon, 17 Apr 2023	**Prob (F-statistic):**	0.225
Time:	22:50:15	**Log-Likelihood:**	-102.59
No. Observations:	95	**AIC:**	217.2
Df Residuals:	89	**BIC:**	232.5
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.7975	14.625	0.260	0.796	-25.262	32.857
adult_dental	0.6426	0.556	1.155	0.251	-0.463	1.748
child_dental	0.1048	0.174	0.603	0.548	-0.241	0.450
EHB_over_premium	-2.3008	14.347	-0.160	0.873	-30.807	26.205
mlr	1.6695	1.427	1.170	0.245	-1.167	4.506
average_rebate	0.0007	0.000	1.781	0.078	-7.84e-05	0.001

Omnibus:	4.163	**Durbin-Watson:**	1.771
Prob(Omnibus):	0.125	**Jarque-Bera (JB):**	3.467
Skew:	-0.427	**Prob(JB):**	0.177
Kurtosis:	3.382	**Cond. No.**	9.51e+04

We see almost the same R^2 but observe no statistically significant coefficient at the 95% significance level. This is a surprising result given the data is all considered at the same level in a more intuitive way. However, it is completely possible that these variables show no relationship at an aggregated mean level but quality of individual plans are affected by the general MLR of each insurer for each state.

Predictive Modeling

In choosing the appropriate method for modeling our data, it became evident that consideration of quality's ordinality would be important. The ordinality of quality could be treated as a continuous or categorical response variable. There are also other methods that account for the ordered quality of ordinal response variables but such methods were not explored. We tested machine learning methods for both continuous and categorical response variables to compare the performances of various models.

We considered multiple methods for modeling and used each quality metric for a response. To validate and score each method, we used repeated stratified k fold cross validation with 5 folds and 3 repetitions and averaged the testing scores across the folds and repetitions. Hyperparameter optimization was then performed for each response variable for each method. The methods with significant results are shown below:

Linear Regression

	Overall Rating	Medical Care	Member Experience	Plan Administration
Testing R^2	0.611	0.681	0.67	0.556
Rounded Testing R^2	0.467	0.581	0.577	0.364

Logistic Regression

	Overall Rating	Medical Care	Member Experience	Plan Administration
Best Average Accuracy	0.779	0.815	0.838	0.806
OneVsRest or multinomial	multinomial	multinomial	multinomial	multinomial

Random Forest Classification

	Overall Rating	Medical Care	Member Experience	Plan Administration
Best Average Accuracy	0.869	0.883	0.917	0.875
Number of estimators	35	37	45	54
Maximum tree depth	22	20	21	20

Decision Tree

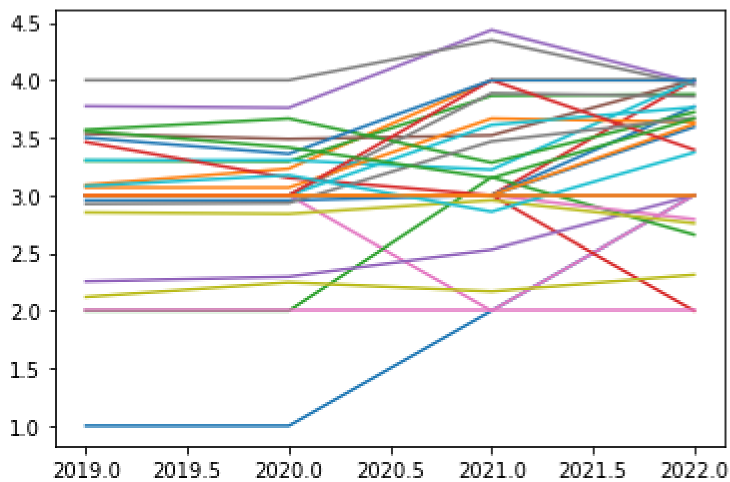
	Overall Rating	Medical Care	Member Experience	Plan Administration
Best Average Accuracy	0.864	0.878	0.917	0.873
Max Depth	24	24	24	18
Most Significant Features	EHB_over_premium state_AZ state_TX	state_AZ EHB_over_premium state_TX	plan_type_HMO state_AZ EHB_over_premium	state_WI state_TX EHB_over_premium

From our results we see that a classification treatment of the ordinal quality metrics performed much better than the continuous treatment. Among the classifiers, the tree methods performed the best and both were nearly identical with respect to accuracy. However, due to the ensemble nature of random forest, the runtime was significantly longer with seemingly no accuracy gains. So the best performing method was the decision tree for each quality metric when also accounting for runtime.

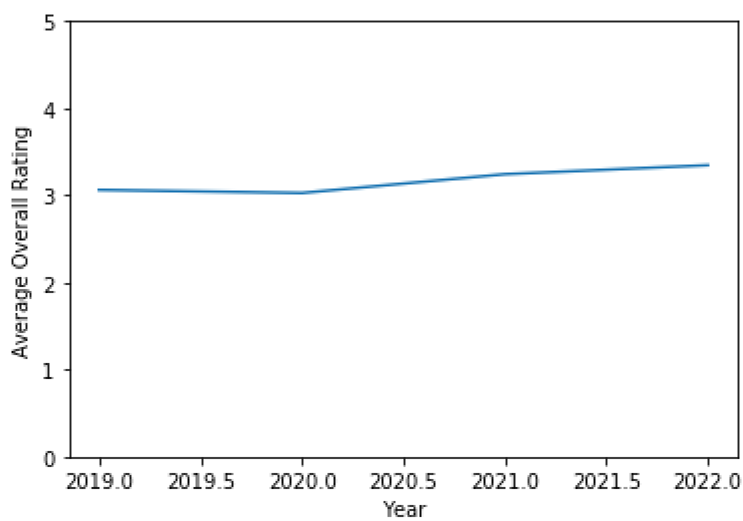
Examining our decision tree model more closely, we can also observe the most significant, or discriminating features of the model. The most discriminating features across the metrics seemed to be state dummy variables (particularly in Arizona, Texas, and Wisconsin), EHB percent of premium, and plan type as HMO. These features being the most significant seem to support the notion of how much variance there is in healthcare systems across different states. Also EHB percent of premium being a top 3 significant feature for each quality metric suggest that this feature is largely important for general quality modeling.

General Trends

For this data we use QHP and quality data from 2019 to 2022 to get time series data over 4 years. Given we have 4 years of data for each health plan (assuming no missing values), our analysis is limited to mostly trends and general insights. First we wanted to understand generally the trend of overall ratings for health plans. Have health plans improved in general quality over time? The following visualization shows the average QHP quality scores per state over time.



What we see seem to be slight increases over time to quality scores for the average health plan for each state. When looking at average overall scores across the whole data set we see:



This shows a similar trend to the state visualizations where overall QHP's tend to increase slightly over the four years.

The state with the highest average rating considered among the 39 was South Dakota (4.08) with West Virginia showing the greatest positive change (2). We also see West Virginia with lowest average rating (1.75) and Arizona with greatest negative change (-1.46).

Finally we perform a similar analysis but for each insurer for each state. The insurer with the highest average overall rating is Kaiser Permanente in Virginia (5) with the lowest average overall rating being CareSource West Virginia Co. in West Virginia (1). Notably of the top 5 overall ratings, Kaiser were the top 3, and for the bottom 5 Blue Cross Blue Shield were the bottom second, third, and fourth. For each year the highest average insurer are reported below:

Year	Insurer	State
2019	Kaiser Permanente	Virgina
2020	Kaiser Permanente	Virginia
2021	Kaiser Permanente	Virginia
2022	Kaiser Permanente	Georgia

Quality Metrics Analysis

Four quality ratings are evaluated for health plans based on the quality of medical care, member experience, and plan administration. Ideally, these three ratings should be independent of each other and overall rating should be related to each of the other ratings as it is calculated as a weighted combination. The CMS states that medical care is given the greatest weight for the overall rating. To further investigate the independence of scoring we observe the correlation matrix of the scores to be:

	overall_rating	medcare_rating	member_exp_rating	plan_admin_rating
overall_rating	1.000000	0.870514	0.223804	0.250568
medcare_rating	0.870514	1.000000	0.070576	0.291489
member_exp_rating	0.223804	0.070576	1.000000	0.222006
plan_admin_rating	0.250568	0.291489	0.222006	1.000000

Unsurprisingly, we see that overall rating is highly correlated with medical care rating and much less correlated with member experience and plan administration. Excluding the overall rating, we can also see that there is very little correlation between the three ratings. To further explore this some simple regression analysis was performed on the data and the results are shown below:

OLS Regression Results

Dep. Variable:	overall_rating	**R-squared:**	0.791
Model:	OLS	**Adj. R-squared:**	0.791
Method:	Least Squares	**F-statistic:**	5177.
Date:	Wed, 19 Apr 2023	**Prob (F-statistic):**	0.00
Time:	21:27:14	**Log-Likelihood:**	-1243.5
No. Observations:	4103	**AIC:**	2495.
Df Residuals:	4099	**BIC:**	2520.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0356	0.037	28.131	0.000	0.963	1.108
medcare_rating	0.6734	0.006	118.432	0.000	0.662	0.685
member_exp_rating	0.1443	0.008	18.822	0.000	0.129	0.159
plan_admin_rating	-0.0557	0.009	-5.965	0.000	-0.074	-0.037

Omnibus:	446.352	**Durbin-Watson:**	0.064
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	601.245
Skew:	-0.905	**Prob(JB):**	2.76e-131
Kurtosis:	3.494	**Cond. No.**	42.6

We can see that since the R^2 is 0.791 (0.83 when predictions are rounded), the overall rating is not simply a weight sum of the three other ratings and only about 79% (83% when predictions are rounded) of the variance of overall ratings are explained by the three other ratings. However in this linear model, all coefficients of the ratings are statistically significant with medical care having a large coefficient value of 0.67. Interestingly, plan administration rating has a negative coefficient. Even though the coefficient is very low, a negative coefficient with a statistically significant p-value is indicative of possible drawbacks of the analysis as done here. Regardless, we can see that when modeled as a linear model, medical care rating is in fact the highest weighted rating component of the overall rating and all the ratings are in fact statistically significant.

Discussion and Conclusion

Summary and Analysis

The main goal of this project was to gain insight into the following three questions:

1) Do characteristic requirements of qualified health plans have effects on their quality?
2) Can the QHP data publicly available be modeled to predict quality metrics?
3) What are the general trends of qualified health plans?

At an individual health plan level, MLR and EHB are significantly correlated to quality. However, MLR and EHB are also significantly correlated with each other and in our regression model we see that only MLR has statistical significance at a 95% significance level. We can also see that our model only has a R^2 of 0.08 implying that while MLR is significant in explaining the variance of quality, it is not nearly strong enough alone to explain more than 10%.

At the insurer state level, we see that only EHB and MLR are significantly correlated with each other with a coefficient of -0.25 with a p-value of 0.134. Our regression model also found that neither EHB or MLR have statistically significant coefficients. Yet the model explains a similar level of variance in quality with a R^2 of 0.074.

It is important to note, that for EHB, we measure EHB over premium. That is, the proportion of money going to essential health benefits over the money from premiums going to medical care. That means that all QHP's already cover all the established EHB's and that our measurement of EHB over premium goes below 100% if your plan covers more than EHB's. So it is not alarming to see that EHB over premium has no statistically significant effect at either level since it may be indicative that quality does not change after EHB's are met. In other words, we can not measure the effect of EHB on quality from this data, only the effect of exceeding EHB on quality.

We will also note that MLR alone is not measured in this project. Even though we have data on MLR, there is the implicit dampening effect imposed on it due to the mandatory rebate punishment if an insurer is found to be non compliant. This rebate therefore biases our estimate of the effect of MLR on quality since MLR is not just the ratio when in noncompliance to the required minimums. We can only understand MLR as how the ACA defines and enforces it.

For our predictive modeling we found that classification models performed better for scoring than our regression methods. Our final chosen model is the Random Forest Classification for all quality metrics due to its strong performance and theoretical benefits for this problem. Random Forest has implicit feature selection properties that are useful due to the high dimensionality of the data. This aspect is especially helpful given that many features are likely collinear. Random Forest also tends to have better generalizability to testing sets due to it being an ensemble method that makes it more robust to small changes in the training sample. For all these reasons, Random Forest ended up being the preferred choice.

Conclusion

The ACA aimed to improve the quality of care while also simultaneously offering higher coverage rates nationally. From the results of this project, we can see that MLR is an important consideration to quality while coverage exceeding essential health benefits shows no signs of significant effects on quality. We also produced a generalized model that insurers and consumers can use to help predict quality of current health plans with .87 testing accuracy. Finally we found general trends that give us insight into the quality trends and performance of states and insurers across 39 states.