Geospatial and Explanatory Models for Heart Failure Admissions, 2016 through 2018

Clemens Scott Kruse
Texas State University
Bradley M. Beauvais
Texas State University
Matthew S. Brooks
Texas State University
Michael Mileski
Texas State University
Lawrence Fulton (✉ lfulton159@gmail.com)
Texas State University https://orcid.org/0000-0001-8603-1913

Research article

Keywords: heart failure, geospatial, obesity, cost analysis

DOI: https://doi.org/10.21203/rs.2.21707/v2

Abstract

Background. About 5.7 million individuals in the United States have heart failure, and the disease was estimated to cost about $42.9 billion in 2020. This research provides geographical incidence models of this disease in the U.S. and explanatory models to account for hospitals' number of heart failure DRGs using technical, workload, financial, geographical, and time-related variables. The research also provides updated financial and demand estimates based on inflationary pressures and disease rate increases. Understanding patterns is important to both policymakers and health administrators for cost control and planning.

Methods. Maps of heart failure diagnosis-related groups (DRGs) from 2016 through 2018 depicted areas of high incidence as well as changes. Spatial regression identified no significant spatial correlations. Simple expenditure forecasts were calculated for 2016 through 2018. Linear, lasso, ridge, and Elastic Net models as well as ensembled tree regressors including were built on an 80% training set and evaluated on a 20% test set.

Results: The incidence of heart failure has increased over time with highest intensities in the East and center of the country; however, several Northern states (e.g., Minnesota) have seen large increases in rates from 2016. The best traditional regression model explained 75% of the variability in the number of DRGs experienced by hospital using a small subset of variables including discharges, DRG type, percent Medicare reimbursement, hospital type, and medical school affiliation. The best ensembled tree models achieved R^2 over .97 on the blinded test set and identified discharges, percent Medicare reimbursement, hospital acute days, affiliated physicians, staffed beds, employees, hospital type, emergency room visits, medical school affiliation, geographical location, and the number of surgeries as highly important predictors.

Conclusions. Overall, the total cost of the three DRGs in the study has increased approximately $61 billion from 2016 through 2018 (average of two estimates). The increase in the more expensive DRG (DRG 291) has outpaced others with an associated increase of $92 billion in expenditures. With the increase in demand (linked to obesity and other factors) as well as the relatively steady-state supply of cardiologists over time, the costs are likely to balloon over the next decade.

1. Introduction

1. *1.1 Demand for Coronary Heart Disease Treatment*

   Coronary heart disease (CHD), cardiovascular disease (CVD), and coronary artery disease (CAD) are leading causes of death in the US, taking the lives of 647,457 in 2017 [1]. Heart disease is the leading cause of death in most developed countries, causing the deaths of one third of those over the age of 35 [2] and one quarter of deaths in the US [3]. Heart disease affects all races and the proportion of deaths attributable to heart disease by race follows: 23.8% non-Hispanic whites, 23.8% non-Hispanic Blacks, 22.2% Asian or Pacific Islander, and 18.4% Native American or Alaskan Native [3]. Incidence of total coronary events in the US increases sharply with age [4,5]. An update of heart disease and stroke in 2016 reported 15.5 million people > 20 years old have CHD [6], which is nearly 6% of that population in the United States [7] . Some of the risk factors for heart disease are high-blood pressure, high cholesterol, and smoking [8]. About 47% of Americans report at least one of these conditions [3]. CHD affects men slightly more than women [9] , and food insecurity (associated with poverty) is an obvious correlational factor [10].

   Heart disease was not a common cause of death at the turn of the 20th century, but the prevalence of coronary atherosclerosis grew until 1960 [11]. In 1900, heart disease was the fourth cause of death, surpassed by infectious conditions [12]. Longevity in our nation increased after 1900 only due to the decrease in infectious diseases [13]. In 1900, less than 5% of Americans smoked, but in 1960 incidence of smoking was 42% [13]. After the 1950s, Americans decreased smoking and reduced cholesterol levels [11]. Deaths from CHD in 1965 decreased from 466 per 100,000 to 345 per 100,000 in 1980: a 26% decrease [14]. Since the 1960s, age-adjusted incidence of heart disease has experienced a steady decline [15], but it is still the number one cause of death in our nation [1]. Mechanisms to track heart disease and predict admissions would be another mechanism to control this killer of Americans: particularly the elderly who are more susceptible to the condition [16].

Heart failure as a subset of heart disease is prevalent in about 6.5 million adults in the United States, and one out of 8 deaths in 2017 were attributed at least in part to heart failure. The annualized cost is estimated to be $30.7 billion in 2012 [17].

1.2. Supply and Payment of Cardiologists

Despite the national average of 383 people per physician in the United States, the number of people per cardiologist is 14,572 [18]. There is certainly an element of artificiality in those numbers because while all people in the U.S. seek some medical care, a much smaller number need specialty care from a cardiologist. However, the message is the same: cardiology is highly specialized and a highly sought area of care.

While the general trend is up for cardiovascular disease (CVD), the growth of those entering cardiology is relatively flat. It is estimated that 40.5% of the U.S. population will have some form of CVD by 2030. This equates to a 3.1% incidence rate and $818 billion in cost of care [19]. A 2018 study of heart failure incidence from 1990 to 2009 revealed that heart failure with reduced ejection fraction (HFrEF) was down, while heart failure with preserved ejection fraction was up (HFpEF) [20]. More recent studies are not readily available.

1.3. Relevant Methods

Explanatory models for healthcare costs have included linear and penalized linear models such as a lasso regression [21] with reasonable success. Other, machine learning techniques such as random forests have also been used to predict and explain CHD events and risk factors successfully [22]. Random forests are an ensemble of tree models used for either regression or classification [23]. This study uses these models for explanatory investigation of CHD in this study as well, as they have proven successful in previous studies of this nature.

Research in public health has leveraged geospatial analysis to look at several aspects of heart disease such as emergency transport and inter-hospital transfer of myocardial infarction [24] as well as individual and contextual correlates of cardiovascular disease [25]. Spatial analysis in the area of public health is conducted at the worldwide, country, and regional levels of analysis [26]. Most often, choropleth maps (maps that depict categorical and numerical data) are used to present one or two data attributes, although dot maps, graduated symbol maps, and isarithmic maps are also commonplace [26]. Spatial regression techniques such as simultaneous autoregressive (SAR) models are often used to document health risks [27], and spatial clustering has been used for leprosy in Brazil [28], measle vaccination in sub-Saharan Africa [29], as well as food and physical activity in the United States [30]. Spatial recognition has been used to identify congenital heart disease in youth aged 4-18 in China as well [31]. Geospatial mapping has been used for describing birthing incidence [32], the opioid epidemic [33], evaluating back surgery growth over time [34], and in many other health-related studies. To date, however, researchers have not conducted a geospatial analysis of heart failure with predictive modeling to provide epidemiological and administrative descriptive and inferential insight as well as economic implications for supply and demand. This research does just that over a three-year window (2016 through 2018).

1.4. Research Question and Significance

This research seeks to understand the geospatial incidence of CHD and to build explanatory models that might account for hospitals' number of heart failure DRGs using technical, workload, financial, and geospatial-temporal variables. Further, the research provides financial and demand estimates based on inflationary pressures and disease rate increases. Understanding patterns is important to both policymakers, epidemiologists, and health administrators alike for cost control and planning efforts. Further, the demand and supply analysis highlight potential shortfalls that may require redress.

2. Methods

2.1. Data

Variables in this study come from the Definitive Healthcare dataset [35]. Definitive Healthcare provided the heart failure data for this study. Diagnostic-related groups (DRGs) associated with heart failure (DRGs 291 ,292, and 293) were selected for inclusion. The Definitive Healthcare datasets contain the Centers for Medicare and Medicaid Services (CMS) Standard Analytical Files (SAF) [35]. State and county-level population data for rate calculations were from the Census Bureau [7,36]. For years 2016 through 2018, there were 13.66, 13.52, and 13.35 thousand hospital observations in the study, respectively. These hospital observations were associated (respectively) with 20.08, 22.74, and 23.46 million DRGs. For the geographical analyses only, the DRG counts were aggregated by county and state for different analyses. These counts were then converted to rates based on the population of the geographic unit, as rates per population base provide a comparison basis across geographical units.

2.2. Variables

The primary variable of interest is admissions for “heart failure” diagnoses as defined by Diagnostic-Related Groups 291, 292, and 293 [37]. The Diagnosis Related Group 291 encompasses “Heart Failure and Shock with Major Complication of Comorbidity (MCC)”; DRG 292 relates to “Heart Failure and Shock with Complication or Comorbidity (CC); DRG 293 pertains to “Heart Failure and Shock without Complication or Comorbidity (CC) / Major Complication or Comorbidity”. The dependent variable is measured at the hospital level and aggregated by county for I mapping. Inpatient claims for heart failure provide a measure of the met demand for services and is suggestive of which areas may need additional funding and resources from health policy decisionmakers.

Variable groups evaluated in the explanatory models included four categories: financial variables, workload variables, technical variables, and geo-spatial temporal variables. All variables are measured at the hospital level by year. Table 1 provides the definitions of the independent variables.

<Insert Table 1 About Here>

2.4. Train and Test Sets

For the explanatory analysis, data were divided randomly using a pseudo-random seed for replication and consistency in model comparison into 80% training and 20% test set of sizes 32,419 and 8,104, respectively. Models were built on the training set and evaluated on the test set. The primary model selection metric of interest was the Root Mean Squared Error (RMSE), a metric which penalizes outlier forecasts heavily.

2.5. Geospatial Analysis

Geospatial maps for the rates of heart failure incident rates for the selected DRGs from 2016 through 2018 were generated at the county and state levels. Rate data adjust for population changes, allowing comparison of incidence rates across counties or states. Population data for each county and the states by year came from Census Bureau estimates [36]. Spatial regression, regression which assumes geographically correlated data, was run using county-level data on the presentation rates for the DRGS similar to Mahara et al. [38]. The significance of changes for 2016 to 2018 (DRG rates) are evaluated by a non-parametric Friedman’s test. The Wilcoxon non-parametric test is preferable and more conservative than repeated samples ANOVA, as normality, homogeneity of variance, and independence assumptions do not hold [39].

2.6. Explanatory Analysis

Linear regression, lasso regression, robust regression, Elastic Net regression, extreme gradient-boosted random forests, and bagging regressors estimate the DRG heart failure admissions. To investigate the bias-variance trade-off [40], we built multiple models on an 80% training and evaluated on a 20% test set. All models are compared based on Root Mean Squared Error (RMSE), which penalizes outliers. The models are exploratory to see which features (workload, financial, technical, and geospatial) might be explanatory.

Lasso regression is a constrained regression that penalizes overfitting using an L1-norm penalty function (absolute value), while ridge regression is similar to lasso regression but penalizes using the L2-norm (squared) [40]. Elastic Net combines both Lasso and Ridge penalty functions [41].

While coefficients are easily interpreted in regression-type models, the data, typically need scaling and transformations with no single best solution available. Unlike tree ensemble models (forests), regression models are unable to find polytomous splits of variables automatically and are not scale invariant. To address the concerns of collinearity, multivariate Box-Cox methods [42] are employed on all quantitative variables simultaneously after location adjustments to make them positive definite.

Random forests are an ensemble of de-correlated tree models. Every tree produces a forecast, and all trees produced are than averaged to produce the estimate. Trees are “pruned,” to prevent overfitting [40]. Figure 1 is an example of a tree with three branches. The tree splits observations by the number of hospital discharges less than or equal to versus greater than or 12,406 initially to obtain the maximum separation (RMSE).

<Insert Figure 1>

Gradient boosted random forests are a special class of ensembled trees. These models use nonlinear optimization to optimize a cost function based on the (pseudo)-residuals of a given function. Unlike random forests, gradient boosted random forests do not produce uncorrelated trees. Instead, the residuals of each tree are re-fitted with the possible independent variables in other tree models. Essentially, the focus is on the residuals. A more complete discussion of gradient boosting is provided in *The Elements of Statistical Learning* [40].

Gradient boosted random forests are scale-invariant, as they find relationships (splits) which the researcher might miss and generate importance metrics for explanatory purposes. These models will, however, overfit the data if the researcher does not restrict the growth of the trees. Cross-validation is necessary.

A Bagging regressor is an ensemble which fits base regressors on random subsets of the original dataset. The estimates from these regressors are then aggregated by voting or averaging to generate a final prediction. The result reduces variance of other block-box estimators by random sampling and ensembling. A good implementation and discussion of bagging regressors is available from the Python scikit-learn module [43].

2.7. Software

All analysis was performed in Anaconda Python Release 3.7 [44], R Statistical Software (inside of Python using the r2py library) [45], and Microsoft Excel 2016 (data wrangling) [29]. Python was used primarily for tree models, while R provided regression analysis and Geographical Information System functions.

3. Results

3.1 Missing Observations

About 2% of quantitative observations were missing, so simple imputation using the mean was employed. This is conservative, as it tends to hide results that might be statistically relevant by reinforcing mean values. For the categorical variables, only ownership was not fully complete. There were only 14 missing observations for this variable, and these were imputed with the mode.

3.2. Descriptive Statistics-Quantitative Data

Descriptive statistics for the quantitative data are provided in Table 2. The average hospital observation during any given year had 1,635.62 observations of DRG 291, 292, and 293 (median of 383). That same average hospital had 146.39 staffed beds (median of 87), 6,996.58 discharges (median of 2,825), 6,348.76 surgeries (median of 4,490), and 34,181.38 acute days (median of 14,051). The average hospital had positive income (in millions) of $17.232 (median of $2.044), significant cash-on-hand ($20.28, median of $1.99), and positive equity. The typical hospital had 1005.53 employees (median of 437) with 231.37 affiliated physicians (median of 104) and was reimbursed 45% by Medicare (median of 42%). Only 9% reimbursement was from Medicaid (median of 6%).

<Insert Table 2 About Here>

Year over year, both DRGs and rates of DRGs per 1000 population increased as illustrated in Figure 2. The significance of the DRG increase is the financial consideration. The significance of the rate of DRG increase is the epidemiological consideration. If the DRG rate is considered a proxy for incidence rate, then there is either a significant increase, a coding issue, or something else. These considerations are found in the discussion section. One might expect the DRG rate graph to remain horizontal (static). Independent variables remained relatively constant year-over-year likely due to repeated measures on the same facilities.

<Insert Figure 2>

3.3. Descriptive Statistics-Categorical Data

California, Texas, and Florida had the largest number of diagnoses for all years and year-over-year, largely due to population size, with averages of 1,669,210; 1,631,021; 1,490,983; respectively. When adjusted per 1000 population, the District of Columbia, West Virginia, and Delaware dominated the with total rates per 1,000 population of 109.47, 102.86, and 94.15, respectively. Utah, Hawaii, and Colorado had the smallest average rates, 26.17, 28.87, and 34.74, respectively. Appendix AIllustrates the rates by state / territory.

Of the hospital observations, 6700 were rural (42%) while 9279 were urban (58%). Most of the hospitals (8342 or 52%) were voluntary non-profits with 29% (4641) proprietary and 18.7% (2996) governmental. The vast majority (11,914 or 75%) had no affiliation with a medical school and were short-term care facilities (9,604 or 60%). Nearly no hospitals were classified as Department of Defense (DoD) or children’s hospitals. Figure 3 depicts the categorical breakout by year.

<Insert Figure 3>

3.4. Descriptive Statistics-Financial Estimates

In FY 2008, the Centers for Medicare and Medicaid (CMS) estimated that heart failure DRGs 291, 292, and 293 national average total costs per case were $10.235, $6.882, and $5.038 thousand, respectively. By FY 2012, CMS increased those estimates to $11.437, $7.841, $5.400 thousand, respectively. In four years, the accumulation rates (1 plus the inflation rate) were 1.139, 1.117, and 1.072 for the DRGs in ascending order. Using these accumulation rates, estimates for 2016, 2017, and 2018 were generated. Table 3 shows these extrapolated estimates.

<Insert Table 3 About Here>

Another method for estimating these costs involved the use of the Federal Reserve Bank of Saint Louis (FRED) producer price index for general medical and surgical hospitals [46]. The annual accumulation rates for 2013 through 2018 were estimated as 1.022, 1.012, 1,007, 1.013, 1.018, and 1.023, respectively. Applying these to the 2012 total costs from CMS results in Table 4 estimates for 2016 through 2018.

<Insert Table 4 About Here>

Both estimates are fairly close. To estimate costs, we used both of these tables separately as upper and lower bounds. Since these total costs represent only CMS costs, the actual financial burden across all payers is likely underestimated as commercial third-party insurers can reimburse up to 90% more than Medicare for the same diagnosis [47]. Figure 4 illustrates the number of DRGs by year, while Figure 5 shows the associated aggregate cost estimates.

<Insert Figure 4>

<Insert Figure 5>

In Figure 4, it is clear that DRG 291, the DRG with the highest average reimbursement rate per case, has increased nonlinearly, while DRG2 292 has seen a small drop, and DRG 293 is flat. In Figure 5, the total cost estimates for 2018 are nearly $66 billion more than 2016 on average. DRG 291, the most expensive DRG, has seen reimbursement increases of $92 billion on average. Reasons for such an increase are explored in the discussion section.

3.4. Descriptive Statistics-Correlational Analysis

Hierarchical clustered correlation analysis of quantitative variables (Figure 6) illustrate tight relationships among many variables. Hierarchical clustered correlation analysis clusters variables based on distance measures (e.g., Euclidean), so that those which are most highly correlated are close in location. These variables are then placed into a correlation plot or correlogram. Figure 6 illustrates that discharges, acute days, and staffed beds are most closely associated with the number of diagnoses, our primary variable of interest.

<Insert Figure 6>

Analysis of the relationship between some categorical variables and the number of diagnoses also proved interesting. Notched boxplots by year and medical school affiliation reveal that a major affiliation experiences a larger number of diagnoses at the .05 level a result that is to be expected. (See Figure 7). Further, voluntary not-for-profits see a larger number of diagnoses (Figure 8).

<Insert Figure 7>

<Insert Figure 8>

3.5. State Level Geospatial Analysis

A descriptive analysis of heart failure over time using geographical informat systems was conducted to evaluate regional differences. Primarily, we were interested in rates per standardize unit in the population of the geographical area. Populations over time were based on Census Bureau estimates for each geographic region [7,36].

While DRG rates per 1000 were not constant over time, the concentrations were fairly consistent. There is a clear bifurcation in the center of the United States separating high and low rates. That bifurcation suggests a clear West-East difference, favoring the West Coast. Washington, D.C. has had (on average) the highest admission rate for diagnoses of heart failure (perhaps, due to the large presence of military and veteran care facilities) followed by West Virginia, Alabama, Mississippi, Michigan, Louisiana, Kentucky, and North Dakota. Of interest is that previous studies indicate these states also see many admissions due to the opioid crisis [33]. Figure 9 shows the admission rates for diagnoses per 100,00 persons by area for all years and for by years on a standardized scale. These diagrams were produced with R [45] using the *usmap* package [48].

<Insert Figure 9>

From 2016 through 2018, the average rate of diagnoses per 1,000 population increased for nearly all states. A Friedman rank sum test (paired, non-parametric ANOVA) of rates by state by year revealed significantly different rates by year by state (c^2_2=70.941, p<.001). Figure 10 illustrates the changes by year and by state.

<Insert Figure 10>

When 2018 data are aggregated at the state / territory level, the DRGs per 1000 paint a slightly different picture, with high-intensities in Washington D.C., West Virginia, Delaware, Mississippi, Kentucky, North Dakota, Michigan, and Missouri (listed in descending order.) Further, evaluating obesity prevalence intensity from the Centers for Disease Control and Prevention (CDC) shows significant correlation between obesity and DRGs per 1000 [49]. A Spearman's test for correlation of obesity prevalence and 2018 DRGs per 1000 was statistically significant with rho=.689, S=6,867.7, *p*<.001.

An exploratory spatial regression model using a first-order Queen contiguity criterion to evaluate the importance of geography wase performed using rolled, Z-scaled, state-level independent variables on the state-level admission rate variable (admissions per population in each state). The final model (after backwards stepwise regression based on Akaike Information Criterion using the *MASS* package in R [50] included mean profit margin for hospitals within the state, total surgical cases, total acute hospital days, total staffed beds, total physicians, mean proportion of Medicare patients, proportion of voluntary not-for-profits (NFP), proportion with a major or graduate affiliation with medical schools, and proportion of urban facilities. The regression was of reasonable effect size, *Adjusted* R^2 = .539. Table 5 summarizes the regression coefficient table.

<Insert Table 5 About Here>

Most important to this preliminary analysis was whether state-level spatial data were important to evaluating admission rates. The spatial map of the standardized residuals [51] is shown in Figure 11. The spatial residual shows little spatial correlation. The visual check was confirmed by a global test for spatial relationships, Moran's I (observed = -0.085, expected = -0.020, p =.510) [52].

<Insert Figure 11>

3.6. County-Level Spatial Analysis

Admissions have significant county variation as one might expect. Figures 12 and 13 depict the sum of the admissions and the admission rate (based on average county populations) for the years 2016 through 2018. The rate of admissions is more useful, as larger populations should likely experience more admissions. The outlier is Winchester County, VA (small population with a major medical center). Charts were generated by the *tmap* package in R [51].

<Insert Figure 12>

<Insert Figure 13>

The heart failure admissions per county population per state for the top five states (West Virginia, Alabama, Mississippi, Michigan, and Louisiana), based on admission rates (year 2018) are shown in Figures 14 through 18, respectively. There is little change in concentration over time. These county maps show that the admissions are generally (as expected) in large metropolitan areas.

For West Virginia, the densest concentration of heart failure admissions is in Logan County. Logan County has a population of 33,674 (2018) and is located south of Charleston. The admission rate density was 0.252 in 2018. In Alabama, Houston county (home to Dothan and a population of 28,838 in 2018) has the highest density, 0.277. Forrest County, Mississippi, home of Hattiesburg and a population of 31,372 in 2018, saw 75,564 admissions for a rate per person of 0.415. The county also has the second largest medical facility based on discharges [35]. Emmet County (home of Petoskey and 32,875 individuals) had the highest density for Michigan (0.329), perhaps because it has a hospital that serves many counties. Red River Parish in Louisiana, a small parish of 8,621 individuals, had the highest rate of heart failure admissions (0.247).

<Insert Figure 14>

<Insert Figure 15>

<Insert Figure 16>

<Insert Figure 17>

<Insert Figure 18>

Similar to what was done at the state level, an exploratory spatial regression model using a first-order Queen contiguity criterion to evaluate the importance of geography wase performed using rolled, Z-scaled, county-level independent variables on the county-level admission rate variable (admissions per population in each county). The final model included average profit margin for facilities in the geography, total ER visits, total surgeries, total staffed beds, total physicians, proportion Medicare patients, proportion voluntary non-profits, proportion with graduate or major medical school affiliations, and proportion that were short-term acute care hospitals (STACs). Results of the regression are in Table 6, and the residual map is shown in Figure 19. The regression accounted for only a small fraction of the sum of the squares (R^2 = .155).

<Insert Table 6 About Here>

-

<Insert Figure 19>

The residual map is not suggestive of spatial autocorrelation given the residual dispersion by county. Moran's I again indicated no significant spatial correlation (observed = 0.019, expected=-0.001, p=.140). Future explanatory models can omit spatial correlation.

3.7. Explanatory Models

Several models were leveraged to explain the number of diagnoses admissions by facility level (the unit of observation in the dataset). The importance of these models is that we might estimate demand based on workload, technical, financial, and geospatial variables. A discussion of data preparation and analysis follows.

3.6.1. Box-Cox Multivariate Transformations

To meet required regression assumptions, multivariate transformation using Box-Cox methods was conducted on location-transformed variables. The location transform was necessary to ensure that all variables were positive definite. Box-Cox methods search for the optimal power transform of all variables simultaneously such that the assumption of multivariate normality cannot be rejected. A logarithmic transform is defined as the power of zero. In order to ensure that all possible transformations are feasible, the data must be positive definite. Thus each variable that was non-positive definite had the absolute value of the minumum added to each observations plus .01. Doing so ensured a positive definite location transform. These transformations are necessary only for non-tree models. (Tree models are location-scale invariant.) To prevent bias from being induced into the unknown test set, the transformations are completed only on the training set. The optimal powers found from the optimization associated with Box-Cox multivariate analysis are then applied to the test set. See Table 7 for the optimal powers.

<Insert Table 7 About Here>

3.6.2. Regression Models

Using the postive definite, Box-Cox transformed data, a regression model was fit hierarchially using the following blocks (in order): technical, workload, financial, geo-spatial. The multivariate transformation assumes that at least some independent variables cannot be fully observed or that we have incomplete observations on variables that might be fully observed. Thus, the transformations from the Box-Cox methods attempt to achieve multivariate normality rather than univariate normality. Hierarchical models attempt to fit obvious (known) variable blocks first followed by those of mmost interest. In our case, all blocks were statistically relevant to the analysis (see Table 8).

<Insert Table 8 About Here>

Linear regression on the training set resulted in a reasonable fit that accounted for R^2=.750 or 75% of the sum of squared variability. No collinearity problems were present after transformation. Performance on the training set was insightful; however, the proof of model explanatory power rests in the training set estimates of the test set values. Applying the parameter estimates generated from the training set to the test set resulted in an R^2 of .749, barely any loss.

Given the model's ability to predict, the linear regression model was re-run on the entirety of the dataset after re-estimating the Box-Cost transformations, transformations which were only slight different in magnitude than those produced by the training set. The results again produce R^2=0.749. The actual versus predicted plot is shown in Figure 20.

<Insert Figure 20>

Further, we evaluated the coefficients and directions of those coefficients for forecasting the transformed dependent variable based on the number of variables included in the model. The top 1- variables in the regression model with their associated R^2 are shown in Table 9. Discharges, Medicare percentage, and hospital type are the primary variables of interest.

<Insert Table 9 About Here>

Outside of simple linear regression, we explored constrained regression techniques (lasso, ridge, and Elastic Net). Lasso regression was inferior to linear regression in terms of R^2 (.651 vs. .750) on the test set. Ridge regression and Elastic Net were also unable to beat linear regression in terms of effects with both R^2 nearly identical to that of lasso, 0.651 and 0.681, respectively. In this case, the linear regression model was not overfit.

3.6.3. Tree Ensemble Models

Several tree regressor models were built and compared on an 80% training set. There was no need to use the transformed data for tree models, as these are location / scale invariant. These tree models included a bagging regressor (BR), a random forest regressor (RFR), an extra trees regressor (ETR), a gradient boosted regressor (GBR), and an extreme-gradient boosted model (XGR). Tree models are atheoretic, as each tree developed may be different from the previous one. When ensembled, variable importances emerge that determine which items are most important to determining how to classify or regress in a nonlinear fashion (piecewise). The number of trees used for each estimator was tuned along with the maximum depth of the trees (number of branches). A pseudo-random number ensured that any model improvements were not due to the random number stream. The results of these models on the unseen test set are shown in Table 6. Most importantly, all of these models account for more variance than regression models. The models predict at 97.1% and above in terms of variability capture. (See Table 10.)

<Insert Table 10 About Here>

Because of the tight congruence of these models, we ensembled the estimates of the number of DRGs forecast by each to produce importance statistics. The variables of most importance includes discharges, Medicare percentage, acute days, affiliated physicians, staffed beds, employees psychiatric hospital status, ER visits, medical school affiliation status, Puerto Rico status and surgeries. See Table 11.

<Insert Table 11 About Here>

When comparing the regression models with the ensembled forests, we see that the first two terms are congruent (discharges and Medicare percent). Interestingly, no financial models are in the top 10 effect sizes of the regression or tree models. Facility technical and workload variables are the most important determinants of heart failure. In the tree models, there were piecewise linear effects identified for states that were not seen in the regression models.

4. Discussion

4.1. Review of Findings

With Figure 2 (DRGs per year), we can see that the number of DRGs for heart failure is increasing over time. We do not have sufficient data or monthly data to run time series analyses such as exponential trend seasonality and auto-regressive integrated moving average models. Even without those models, it is clear that there appears to be an increase in heart failure admission diagnoses and a change in intensity from 2016. What is most interesting is that intensity changes are largely in the North Central while current incidence rates are highest East of the Texas panhandle.

Further, we see variables that explain the number of DRGs of a facility over time. Some of these are logically associated with the size of facility (e.g., number of discharges). One of these is logically associated with age (Medicare, available to those 65 and older.) However, the tree model ensembles suggest a significant geographic component for explaining heart failure. Specifically, Tennessee, Puerto Rico, North Carolina, New York, South Dakota, Hawaii, Arizona, Michigan, Washington D.C., Florida, and California. Further, there is an effect by year noticed in the ensemble of tree models, as 2016 is much lower than 2018.

Considering our findings from a financial perspective, our results clearly indicate there has been a significant shift in cardiology diagnoses since 2016. As we note, it is clear that DRG 291, Heart Failure and Shock with Major Complication of Comorbidity (MCC), counts and costs have increased nonlinearly. DRG 292, Heart Failure and Shock with Complication or Comorbidity (CC), has seen a small drop and DRG 293, Heart Failure and Shock without Complication or Comorbidity (CC) / Major Complication or Comorbidity, is flat. A DRG is determined by the principal diagnosis, the principal procedure, if any, and certain secondary diagnoses identified by CMS as comorbidities and complications (CCs) and major comorbidities and complications (MCCs) [53]. A comorbidity is a condition that existed before admission. A complication is any condition occurring after admission, not necessarily a complication of care [54]. Although heart failure DRGs represented the largest cause of hospitalizations among Medicare beneficiaries and were among the costliest to Medicare prior to 2016, the results of our study now suggest that total cost estimates for these three DRGs in 2018 are now nearly $61 billion more than 2016 [55-57]. DRG 291, the most expensive DRG, is associated with $91 billion cost increases from 2016.

4.2. Limitations and Future Work

This study is limited in that only three complete years of data were available. As more data become available, the analysis will be expanded. Further, the study does not consider sub-DRGs, which might provide additional value in understanding the cost structure, particularly since procedures such as Extracorporeal Membrane Oxygenation (ECMO) are highly costly yet coded across multiple DRGs.

While it is likely that many individuals receiving care in a geographic area are from outside that county or state, the majority are likely to receive care near the vicinity of the admission, particularly since heart failure is a medical emergency. Further, the intent of the study is to explain admissions and their associated locations. For public health professionals interested in where heart failures (rather than admissions) occur, the state level geographical analysis would be more reflective as it reduces bias associated with facility locations.

Although our research has demonstrated substantial reliability in the explanatory factors associated with the longitudinal growth trajectory, it does not explain the reasons why we see such substantial growth in DRG 291 versus DRGs 292 and 293. Given our study results, there are several potential drivers that could meaningfully contribute to the growth in DRG 291 from 2016 to 2018. First, there may have been a significant increase in patients with cardiac conditions with additional major comorbidities. This cannot be simply dismissed given the rapid increase in Medicare eligible beneficiaries – by some estimates as many as ten thousand per day – and the prevalence of obesity, coronary obstructive pulmonary disease, and other age and lifestyle related conditions [58-60]. However, given the relatively flat or declining rate in DRGs 292 and 293, we do not believe this is the only driver of our findings. Our findings support other predictions that soon patient demand will outpace the supply [61,62].

Second, up until October 2018, all extracorporeal membrane oxygenation (ECMO) cases were assigned to DRG 003, which typically reimburses at a rate of roughly $100,000 per case [63]. In fiscal year 2019, which started in October 2018, that reimbursement methodology changed so that every ECMO case would no longer be assigned to DRG 003. Rather, the DRG assigned depends on the path of the cannulation. If the ECMO patient is accessed centrally, DRG 003 is still applied. However, if cannulated peripherally, then it falls into another (lower-paying) DRG [64,65]. Although there is only a three month overlap of this change and our study dataset, there is high likelihood this additional volume is reflected in our study in 2018.

Third, since 2010 and the passage of the Affordable Care Act, many cardiologists have sought hospital employment versus private practice. The uncertainty of continued healthcare reform efforts, burdensome electronic health record costs, declining CMS reimbursement rates in physician professional fees for non-invasive testing procedures (e.g., electrocardiograms, nuclear stress tests, etc.), and younger clinicians' different expectations related to work and personal life balance have all combined to prompt cardiology groups to seek ways to stay financially viable. Today more than 70 percent of cardiologists are employed by hospitals or health systems [66,67]. Hospitals, in turn, seek to maximize utilization and reimbursement from the highly resource intensive cardiology service lines. Prior research from the National Bureau of Economic Research found that hospitals responded to price changes by up-coding patients to diagnosis codes associated with large reimbursement increases. These authors indicate hospitals do not alter their treatment or admissions policies based on diagnosis-specific prices; however, they employ sophisticated coding strategies in order to maximize total reimbursement [68,69].

Fourth, we suspect the recent transition from ICD-9 to ICD-10 that occurred in October 2015 is a contributing factor. Starting on October 1, 2015, there were 68,069 valid ICD-10-CM diagnosis codes, representing a nearly 5-fold increase from the 14,025 valid ICD-9-CM diagnosis codes. ICD-10-CM diagnosis codes are structured differently from ICD-9-CM codes and provide more detail [49]. This code expansion allows providers the ability to capture the severity and specificity of the condition in much greater detail – which may prompt increased use of DRG 291.

As we look at the number of times many of the codes are being assigned to any particular patient, we see a significant change in how physicians are diagnosing. Previously, we had an ICD-9 diagnosis code with some generic areas that covered many patients. A very general and generic set of heart failure codes existed under 428.x in ICD-9. There was little specificity as to sidedness of the issue or specifics of the disease. ICD-10 codes allow a very specific diagnosis per codes, and these codes will continue to change over time due to physicians' adaptation of coding in this manner. For example, the I50.8xx codes did not exist in 2016, but they have been used since 2017, with another change adding more sub-codes in 2018.

Today, we have very specific codes for very specific diseases and processes which go on within the heart, to include acute on chronic concerns as well. The adjustment to ICD-10 codes has undoubtedly created a learning phase for practitioners on determining the appropriate codes as well as when and how to use them.

We would expect to see some elevation from year to year with the growth of the Baby Boomer population coming into healthcare, without an age adjustment to the population. This is shown in the numbers from 2016-2018 with total admissions diagnoses increasing from 5.39M to 5.61M to 5.69M. However, how the diagnosis codes are being applied shows variation from year to year, to include some years of negative numbers in several codes. Many of the negative values for codes are for “unspecified” types of heart disease. This shows that we are moving away from generic diagnoses and towards diagnoses based in specificity instead, which is one of the purposes of moving to ICD-10.

One could draw a conclusion of upcoding: a monetary free-for-all, assigning diagnoses based on what pays the most. However, in many cases the physician is not billing based on a diagnosis code, but on the level of the visit and the type. This is obviously dependent upon insurance types, contracts, and other inputs outside the discussion level of this paper.

Of curious note, we are seeing an interesting trend looking at the GIS information included in this paper to where heart failure diagnoses are being seen. In the areas which are surrounding oil and gas pipelines, we have seen a growth in the numbers of heart failure diagnoses in those areas. For our purposes here, the conclusion is only empirical, however there is a significant change in the heat maps in the areas surrounding pipelines. If the reader will overlay the route of the Keystone Pipeline from Canada to Galveston, Texas, you will not a curious overlap with incidence of heart failure. One author also noted an increased use of methamphetamine and cocaine by oil field workers [70]. It is certainly beyond the scope of this research, but it might be something to consider for future research because a consequence of the use of these illicit drugs are differing heart disorders, to include heart failure.

5. Conclusions

The policy implications of this analysis are several. First, clearly the need to continue to focus on a population health approach to reduce obesity rates across the country is needed, focusing specifically on the geographic states identified with the highest incidence and prevalence across the study timeline. The large increase in the DRGs 291 – 293 show that shifting funding to prevention from chronic disease management certainly has the financial evidence to support this approach. The argument is certainly made that education is not sufficient to change lifestyle and behaviors contributing to the rise of heart disease shown here, so it is time to begin exploring a punitive annual health assessment requirement for high-risk individuals who fail to make significant risk factor changes. While a punitive health assessment might incentivize behavioral modifications and might result in lower costs, there is also the possibility that these modifications will possibly require medication and surgical interventions. Using such a strategy alone is not likely to produce the results required. The health administrator will certainly need to analyze both the volume and scope of services within these analyzed DRGs to ensure the evident increase in demand indicated will be available, specifically in the identified high incidence geographic areas. In Certificate of Need (CON) states, this analysis will be beneficial in getting the CON approved based on the increased demand. Evidence shows that CON states for cardiac services, of which most of the high incidence and prevalence states in the study are, have higher mortality rates for cardiac services [71]. Another significant potential policy implication is a continued re-evaluation of the need for CONs in general, as multiple researchers are showing it is in question if they are still needed in today’s healthcare environment, and potentially are leading to restriction of services that are in increasing demand and lead to higher mortality [72] .

Abbreviations

Emergency Room

Complication of Comorbidity

CMS. Centers for Medicare & Medicaid
CON. Certificate of Need

DRG. Diagnostic Related Group

ECMO. Extracorporeal membrane oxygenation

GIS. Geographical Information System

HFpEF. Heart Failure preserved Ejection Fraction

HFrEF. Heart Failure reduced Ejection Fraction

ICD. International Classification of Disease Version (-version)

MCC. Major Complication of Comorbidity

Declarations

Ethics approval and consent to participate. NA

Consent for publication. NA

Availability of data and materials. The dataset(s) supporting the conclusions of this article are available in the Definitive Healthcare repository, https://www.definitivehc.com/" For Python and R code, please contact Dr. Larry Fulton at lf25@txstate.edu.

Competing interests. The authors have no competing interests.

Funding. The authors received no funding for this work.

Authors' contributions.All authors have read and agree to the published version of the manuscript. Conceptualization, S.K. and LF.; methodology, L.F.; software, L.F.; validation, M.B, M.M., B.B..; formal analysis, L.F..; writing—original draft preparation, S.K..; writing—review and editing, L.F., M.B, B.B., M.M."

Acknowledgements. NA

References

1. Fulton, A.N.; Minoo, T.; Lawrence. Health Disparities and Cardiovascular Disease. *Healthcare* **2020**, *8*, 65, doi:10.3390/healthcare8010065.
2. Sanchis-Gomar, F.; Perez-Quilis, C.; Leischik, R.; Lucia, A. Epidemiology of coronary heart disease and acute coronary syndrome. *Annals of Translational Medicine* **2016**, *4*, 7.
3. CDC.gov. Heart Disease Facts | cdc.gov. Availabe online: https://www.cdc.gov/heartdisease/facts.htm (accessed on 26 May 2020).

4. Steenman, M.; Lande, G. Cardiac aging and heart disease in humans. *Biophysical Reviews* **2017**, *9*, 131-137, doi:10.1007/s12551-017-0255-9.

5. Lakatta, E.G.; Levy, D. Arterial and Cardiac Aging: Major Shareholders in Cardiovascular Disease Enterprises. *Circulation* **2003**, *107*, 346-354, doi:doi:10.1161/01.CIR.0000048893.62841.F7.

6. Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; Das, S.R.; Ferranti, S.d.; Després, J.-P.; Fullerton, H.J., et al. Heart Disease and Stroke Statistics&#x2014;2016 Update. *Circulation* **2016**, *133*, e38-e360, doi:doi:10.1161/CIR.0000000000000350.

7. Census.gov. Population Clock. Availabe online: https://www.census.gov/popclock/population_widget_200x402.php?component=density&no_scode#us (accessed on 26 May 2020).

8. Pencina, M.J.; Navar, A.M.; Wojdyla, D.; Sanchez, R.J.; Khan, I.; Elassal, J.; D'Agostino, R.B.; Peterson, E.D.; Sniderman, A.D. Quantifying Importance of Major Risk Factors for Coronary Heart Disease. *Circulation* **2019**, *139*, 1603-1611, doi:doi:10.1161/CIRCULATIONAHA.117.031855.

9. Fodor, J.G.; Tzerovska, R. Coronary heart disease: is gender important? *The Journal of Men's Health & Gender* **2004**, *1*, 32-37, doi:https://doi.org/10.1016/j.jmhg.2004.03.005.

10. Berkowitz, S.A.; Berkowitz, T.S.Z.; Meigs, J.B.; Wexler, D.J. Trends in food insecurity for adults with cardiometabolic disease in the United States: 2005-2012. *PLOS ONE* **2017**, *12*, e0179172, doi:10.1371/journal.pone.0179172.

11. Dalen, J.E.; Alpert, J.S.; Goldberg, R.J.; Weinstein, R.S. The epidemic of the 20(th) century: coronary heart disease. *Am J Med* **2014**, *127*, 807-812, doi:10.1016/j.amjmed.2014.04.015.

12. Jones, D.S.; Podolsky, S.H.; Greene, J.A. The Burden of Disease and the Changing Task of Medicine. *New England Journal of Medicine* **2012**, *366*, 2333-2338, doi:10.1056/NEJMp1113569.

13. Cole, H.M.; Fiore, M.C. The War Against Tobacco: 50 Years and Counting. *JAMA* **2014**, *311*, 131-132, doi:10.1001/jama.2013.280767.

14. NIH.gov. Morbidity & Mortality: 2012 Chart Book on Cardiovascular,Lung, and Blood Diseases. Availabe online: https://www.nhlbi.nih.gov/files/docs/research/2012_ChartBook (accessed on 26 May 2020).

15. Prevalence of coronary heart disease--United States, 2006-2010. *MMWR Morb Mortal Wkly Rep* **2011**, *60*, 1377-1381.

16. Reynolds, I.; Page, R.L.; Boxer, R.S. Cardiovascular Health and Healthy Aging. In *Healthy Aging: A Complete Guide to Clinical Management*, Coll, P.P., Ed. Springer International Publishing: Cham, 2019; 10.1007/978-3-030-06200-2_5pp. 31-51.

17. CDC.gov. Heart Failure. Availabe online: https://www.cdc.gov/heartdisease/heart_failure.htm?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fdhdsp%2Fdata_statistics%2Ffact_sheets%2Ffs_heart_failure.htm (accessed on 26 May 2020).

18. Aamc.org. Number of People per Active Physician by Specialty, 2015. Availabe online: https://www.aamc.org/data-reports/workforce/interactive-data/number-people-active-physician-specialty-2015 (accessed on 26 May 2020).

19. Heidenreich, P.A.; Trogdon, J.G.; Khavjou, O.A.; Butler, J.; Dracup, K.; Ezekowitz, M.D.; Finkelstein, E.A.; Hong, Y.; Johnston, S.C.; Khera, A., et al. Forecasting the Future of Cardiovascular Disease in the United States. *Circulation* **2011**, *123*, 933-944, doi:doi:10.1161/CIR.0b013e31820a55f5.

20. Tsao, C.W.; Lyass, A.; Enserro, D.; Larson, M.G.; Ho, J.E.; Kizer, J.R.; Gottdiener, J.S.; Psaty, B.M.; Vasan, R.S. Temporal Trends in the Incidence of and Mortality Associated With Heart Failure With Preserved and Reduced Ejection Fraction. *JACC: Heart Failure* **2018**, *6*, 678-685, doi:https://doi.org/10.1016/j.jchf.2018.03.006.

21. Kan, H.J.; Kharrazi, H.; Chang, H.-Y.; Bodycombe, D.; Lemke, K.; Weiner, J.P. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PLOS ONE* **2019**, *14*, e0213258, doi:10.1371/journal.pone.0213258.

22. Rajalaxmi, A.S.A.; Rajalaxmi, R.R.; Abdullah, A.S.; R, R. A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier. In Proceedings of IJCA Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls, 22/4/2012; pp. 22-25.

23. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random Forests. In *Ensemble Machine Learning: Methods and Applications*, Zhang, C., Ma, Y., Eds. Springer US: Boston, MA, 2012; 10.1007/978-1-4419-9326-7_5pp. 157-175.

24. Concannon, T.W.; Kent, D.M.; Normand, S.-L.; Newhouse, J.P.; Griffith, J.L.; Ruthazer, R.; Beshansky, J.R.; Wong, J.B.; Aversano, T.; Selker, H.P. A Geospatial Analysis of Emergency Transport and Inter-Hospital Transfer in ST-Segment Elevation Myocardial Infarction. *American Journal of Cardiology* **2008**, *101*, 69-74, doi:10.1016/j.amjcard.2007.07.050.

25. Sun, W.; Gong, F.; Xu, J. Individual and contextual correlates of cardiovascular diseases among adults in the United States: a geospatial and multilevel analysis. *GeoJournal* **2019**, 10.1007/s10708-019-10049-7, doi:10.1007/s10708-019-10049-7.

26. Bhunia, G.S.; Shit, P.K. Spatial Database for Public Health and Cartographic Visualization. In *Geospatial Analysis of Public Health*, Springer International Publishing: Cham, 2019; 10.1007/978-3-030-01680-7_2pp. 29-57.

27. Chakraborty, J. Revisiting Tobler's First Law of Geography: Spatial Regression Models for Assessing Environmental Justice and Health Risk Disparities. In *Geospatial Analysis of Environmental Health*, Maantay, J.A., McLafferty, S., Eds. Springer Netherlands: Dordrecht, 2011; 10.1007/978-94-007-0329-2_17pp. 337-356.

28. Ramos, A.C.V.; Yamamura, M.; Arroyo, L.H.; Popolin, M.P.; Chiaravalloti Neto, F.; Palha, P.F.; Uchoa, S.A.d.C.; Pieri, F.M.; Pinto, I.C.; Fiorati, R.C., et al. Spatial clustering and local risk of leprosy in São Paulo, Brazil. *PLOS Neglected Tropical Diseases* **2017**, *11*, e0005381, doi:10.1371/journal.pntd.0005381.

29. Brownwright, T.K.; Dodson, Z.M.; van Panhuis, W.G. Spatial clustering of measles vaccination coverage among children in sub-Saharan Africa. *BMC Public Health* **2017**, *17*, 957, doi:10.1186/s12889-017-4961-9.

30. Wende, M.E.; Stowe, E.W.; Eberth, J.M.; McLain, A.C.; Liese, A.D.; Breneman, C.B.; Josey, M.J.; Hughey, S.M.; Kaczynski, A.T. Spatial clustering patterns and regional variations for food and physical activity environments across the United States. *International Journal of Environmental Health Research* **2020**, 10.1080/09603123.2020.1713304, 1-15, doi:10.1080/09603123.2020.1713304.

31. Ma, L.-G.; Chen, Q.-H.; Wang, Y.-Y.; Wang, J.; Ren, Z.-P.; Cao, Z.-F.; Cao, Y.-R.; Ma, X.; Wang, B.-B. Spatial pattern and variations in the prevalence of congenital heart disease in children aged 4–18 years in the Qinghai-Tibetan Plateau. *Science of The Total Environment* **2018**, *627*, 158-165, doi:https://doi.org/10.1016/j.scitotenv.2018.01.194.

32. MacQuillan, E.L.; Curtis, A.B.; Baker, K.M.; Paul, R.; Back, Y.O. Using GIS Mapping to Target Public Health Interventions: Examining Birth Outcomes Across GIS Techniques. *Journal of Community Health* **2017**, *42*, 633-638, doi:10.1007/s10900-016-0298-z.

33. Fulton, L.; Dong, S.; Zhan, B.; Kruse, C.S.; Stigler-Granados, P. Geospatial-Temporal and Demand Models for Opioid Admissions, Implications for Policy. *Journal of Clinical Medicine* **2019**, *8*, 993, doi:10.3390/jcm8070993.

34. Fulton, L.; Kruse, C.S. Hospital-Based Back Surgery: Geospatial-Temporal, Explanatory, and Predictive Models. *J Med Internet Res* **2019**, *21*, e14609, doi:10.2196/14609.

35. DHC.com. Definitive Healthcare. Availabe online: https://www.definitivehc.com/ (accessed on 26 May 2020).

36. Bureau, U.S.C. County Population by Characteristics: 2010-2018. U.S. Census Bureau: Washington, DC, USA, 2020.

37. CMS.gov. ICD-10-CM/PCS MS-DRG v37.0 Definitions Manual. Availabe online: https://www.cms.gov/icd10m/version37-fullcode-cms/fullcode_cms/P0140.html (accessed on 26 May 2020).

38. Mahara, G.M.; Wang, C.; Yang, K.; Chen, S.; Guo, J.; Gao, Q.; Wang, W.; Wang, Q.; Xiuhua, G. The Association between Environmental Factors and Scarlet Fever Incidence in Beijing Region: Using GIS and Spatial Regression Models. *International Journal of Environmental Research and Public Health* **2016**, *13*, 1083, doi:10.3390/ijerph13111083.

39. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association* **1937**, *32*, 675-701, doi:10.2307/2279372.

40. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*; Springer: New York, NY, 2009.

41. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2005**, *67*, 301-320, doi:10.1111/j.1467-9868.2005.00503.x.

42. Biscay Lirio, R.; Valdés Sosa, P.A.; Pascual Marqui, R.D.; Jiménez-Sobrino, J.C.; Alvarez Amador, A.; Galán Garcia, L. Multivariate Box-Cox transformations with applications to neurometric data. *Computers in Biology and Medicine* **1989**, *19*, 263-267, doi:https://doi.org/10.1016/0010-4825(89)90013-9.

43. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J., et al. API design for machine learning software: experiences from the scikit-learn project. In Proceedings of European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (2013), 2013/09/01.

44. Team, P.C. *Python: A dynamic, open source programming language.*, 2015.

45. Team, R.C. *R: A language and environment for statistical computing*, 2018.

46. Bls.gov. Producer Price Index by Industry: General Medical and Surgical Hospitals (PCU622110622110) | FRED | St. Louis Fed. Availabe online: https://fred.stlouisfed.org/series/PCU622110622110 (accessed on 30 May 2020).

47. Ahip.org. National Comparisons of Commercial and Medicare Fee-for-Service Payments to Hospitals. American Health Insurance Programs Data Brief. Availabe online: https://www.ahip.org/wp-content/uploads/2016/02/HospitalPriceComparison_2.10.16.pdf (accessed on 30 May 2020).

48. Di Lorenzo, P. *usmap: US Maps Including Alaska and Hawaii. R package version 0.5.0.*, R Statistical Software: 2019.

49. Cdc.gov. Adult Obesity Prevalence Maps | Overweight & Obesity | CDC. Availabe online: https://www.cdc.gov/obesity/data/prevalence-maps.html (accessed on 30 May 2020).

50. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, 2002.

51. Tennekes, M. tmap: Thematic Maps in R. *Journal of Statistical Software* **2018**, *84*, 1-39.

52. Moran, P.A.P. Notes on Continuous Stochastic Phenomena. *Biometrika* **1950**, *37*, 17-23, doi:10.2307/2332142.

53. Lisa Roat, R.C.C. Understanding the Impact of ICD-10 on DRGs - ICD10monitor. Availabe online: https://www.icd10monitor.com/understanding-the-impact-of-icd-10-on-drgs (accessed on 30 May 2020).

54. Pinson, R. The ABCs of DRGs. Availabe online: https://acphospitalist.org/archives/2019/05/coding-corner-the-abcs-of-drgs.htm (accessed on 30 May 2020).

55. Fitch, K.; Pelizarri, P.; Pyenson, B. The high cost of heart failure for the Medicare population: An actuarial cost analysis. Availabe online: https://us.milliman.com/en/insight/the-high-cost-of-heart-failure-for-the-medicare-population-an-actuarial-cost-analysis (accessed on 30 May 2020).

56. Ahrq.gov. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013 #204. Availabe online: https://www.hcup-us.ahrq.gov/reports/statbriefs/sb204-Most-Expensive-Hospital-Conditions.jsp (accessed on 30 May 2020).

57. Kilgore, M.; Patel, H.K.; Kielhorn, A.; Maya, J.F.; Sharma, P. Economic burden of hospitalizations of Medicare beneficiaries with heart failure. *Risk Manag. Healthc. Policy* **2017**, *10*, 63-70, doi:10.2147/RMHP.S130341.

58. Kuehn, B. Obesity Rates Increasing. *JAMA* **2018**, *320*, 1632-1632, doi:10.1001/jama.2018.15094.

59. MEDPAC.gov. Medicare Payment Advisory Commission. Report to the Congress: Medicare and the Health Care. Delivery System. Availabe online: http://www.medpac.gov/docs/default-source/reports/chapter-2-the-next-generation-of-medicare-beneficiaries-june-2015-report-.pdf (accessed on 19 December 2019).

60. Maio, S.; Baldacci, S.; Carrozzi, L.; Pistelli, F.; Angino, A.; Simoni, M.; Sarno, G.; Cerrai, S.; Martini, F.; Fresta, M., et al. Respiratory symptoms/diseases prevalence is still increasing: a 25-yr population study. *Respir. Med.* **2016**, *110*, 58-65, doi:10.1016/j.rmed.2015.11.006.

61. Guilford-Blake, R. Clinician shortage ahead? Cardiology's workforce prepares for a pair of silver tsunamis. Availabe online: https://www.cardiovascularbusiness.com/clinician-shortage-ahead-cardiologys-workforce-prepares-pair-silver-tsunamis (accessed on 30 May 2020).

62. Sauer, J. Cardiology Workforce Analysis. Availabe online: https://www.medaxiom.com/clientuploads/documents/Workforce_Analysis.pdf (accessed on 30 May 2020).

63. Recker, S. ECMO Programs: A Financial Synopsis. *CathLab Digest* **2019**, *27*, 1.

64. Sts.org. Changes to ECMO MS-DRG Assignment Impacts Hospital Payment | STS. Availabe online: https://www.sts.org/advocacy/changes-ecmo-ms-drg-assignment-impacts-hospital-payment (accessed on 30 May 2020).

65. Rose, R.A.; Combs, P.; Piech, R.; LaBuhn, C.; Jeevanandam, V.; Song, T. The CMS Changes to a US ECMO Reimbursement: The Financial Impact upon ECMO Programs. *The Journal of Heart and Lung Transplantation* **2019**, *38*, S261, doi:10.1016/j.healun.2019.01.650.

66. Sobal, L. Has Employment of Cardiologists Been a Successful Strategy? – Part 1 - American College of Cardiology. Availabe online: http://www.acc.org/membership/sections-and-councils/cardiovascular-management-section/section-updates/2019/11/06/09/49/has-employment-of-cardiologists-been-a-successful-strategy-part-1 (accessed on 30 May 2020).

67. Wann, S. Consolidation and hybridization in the health care enterprise: How are cardiologists affected? Page 3. *Cardiology Today* **2018**, *Online*.

68. Dafny, L.S. How Do Hospitals Respond to Price Changes? *American Economic Review* **2005**, *95*, 1525-1547, doi:10.1257/000282805775014236.

69. Moore, B.J.; McDermott, K.W.; Elixhauser, A. ICD-10-CM Diagnosis Coding in HCUP Data: Comparisons With ICD-9-CM and Precautions for Trend Analyses. Availabe online: https://www.hcup-us.ahrq.gov/datainnovations/ICD-10_DXCCS_Trends112817.pdf (accessed on 30 May 2020).

70. Farrell, P. Methamphetamine fuels the West's oil and gas boom — High Country News – Know the West. Availabe online: https://www.hcn.org/issues/307/15811 (accessed on 30 May 2020).

71. Ho, V.; Ku-Goto, M.-H.; Jollis, J.G. Certificate of Need (CON) for Cardiac Care: Controversy over the Contributions of CON. *Health Services Research* **2009**, *44*, 483-500, doi:10.1111/j.1475-6773.2008.00933.x.

72. Mitchell, M. Do Certificate-of-Need Laws Still Make Sense in 2019? *Managed Healthcare Executive* **2019**, *Online*.

Tables

Table 1. Independent variables

Technical Variables	Defined	Measurement
Staffed Beds	Number of staffed beds operated by hospital	Integer
Affiliated Physicians	Number of physicians affiliated with hospital	Integer
Employees	Number of direct employees of hospital	Integer
% Medicare	Percent of patients reimbursing via Medicare	Ratio
% Medicaid	Percent of patients reimbursing via Medicaid	Ratio
Diagnostic-Related Groups	DRG 291, DRG 292, DRG 293	Categorical
Ownership	Hospital Ownership	Categorical
Medical School Affiliation	None, Limited, Major, Graduate Affiliation	Categorical
Hospital Type	Children, Critical Access, Long-Term, Psychiatric, Rehab, Short-Term	Categorical
Workload Variables	**Defined**	**Measurement**
Discharges	Number of patients discharged from admission	Integer
ER Visits	Number of emergency room visits	Integer
Surgeries	Number of surgeries performed	Integer
Acute Days	Number of acute bed days of hospital	Integer
Financial Variables	**Defined**	**Measurement**
Net Income	Profit minus loss	Ratio
Operating Profit Margin	Profit divided by revenue	Ratio
Cash on Hand	Cash available to the organization	Ratio
Equity	Assets minus liabilities	Ratio
Geospatial Variables (and Time Window)	**Defined**	**Measurement**
State	Indicator variables for hospital’s state	Dichotomous
County	Indicator variables for county in states	Dichotomous
Urban / Rural	Indicator variable for metropolitan status	Dichotomous
Year	Indicator variables for year of observation (2016 through 2018)	Dichotomous

Table 2. Descriptive statistics for the study (dollars in millions)

n=40,523 hospital observations		Mean	SD	Median	Min	Max
Quantitative Technical Variables	Number DRGs	1,635.62	3330.63	383.00	11	57,461
	Staffed Beds	146.39	171.98	87	2	2,753
	Affiliated Physicians	231.37	352.66	104	1	4328
	Employees	1,005.53	1679.14	437	4	26,491
	Percent Medicare	0.45	0.19	0.42	0	.98
	Percent Medicaid	0.09	.09	.06	0	.87
Workload Variables	Discharges	6,996.58	9,881.39	2,825	1	129,339
	ER Visits	32,865.58	33,893.77	25,236	0	543,457
	Surgeries	6,348.76	7,965.82	4,490	0	130,741
	Acute Days	34,181.38	51,725.31	14,051	5	701,074
Financial Variables	Net Income ($ in M)	$17.23	$117.65	$2.04	-$1.21	$3,31
	Cash on Hand ($ in M)	$20.28	$120.24	$1.99	-$2.51	$3.88
	Profit Margin	-0.03	1.25	-0.02	-15.45	62.07
	Equity ($ in M)	$174.11	$625.76	$33.94	-$3.25	$10.24

Table 3. Estimated total costs for heart failure by DRG in thousands, linear extrapolation method

DRG	2016	2017	2018
DRG 291	$12,780	$13,155	$13,243
DRG 292	$8,934	$9,245	$9,257
DRG 293	$5,788	$5,891	$5,998

Table 4. Estimated total costs for heart failure by DRG in thousands, medical inflation rate method

DRG	2016	2017	2018
DRG 291	$12,058	$12,273	$12,582
DRG 292	$8,267	$8,414	$8,626
DRG 293	$5,693	$5,795	$5,491

Table 5. Results of the state-level regression

Variable (By Geographic Region)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.000	0.095	0.000	1.000
Mean Profit Margin	-0.262	0.119	-2.195	0.034
Total Surgeries	1.917	0.507	3.785	<0.001
Total Acute Days	-2.295	0.955	-2.404	0.028
Total Staffed Beds	2.136	0.928	2.303	0.026
Total Physicians	-1.778	0.575	-3.090	0.004
Proportion Medicare	0.652	0.121	5.396	<0.001
Proportion Voluntary NFP	0.285	0.128	2.234	0.031
Proportion Major Affiliation Med. School	0.364	0.157	2.325	0.025
Proportion Urban	0.529	0.153	3.466	0.001
Residual standard error: 0.679 on 41 degrees of freedom				
Multiple R-squared: 0.622, Adjusted R-squared: 0.539				
F-statistic: 7.49 on 9 and 41 DF, p-value: < 0.001				

Table 6. Regression table for spatial analysis

Variables (by Geographical Unit)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.000	0.186	0.000	1.000
Mean Profit Margin	0.046	0.189	2.458	0.014
Total ER Visits	-0.197	0.088	-2.236	0.026
Total Surgeries	0.421	0.078	5.384	<0.001
Total Staffed Beds	0.292	0.106	2.744	0.006
Total Physicians	-0.465	0.087	-5.345	<0.001
Proportion Medicare Patients	0.119	0.022	5.351	<0.001
Proportion Voluntary NFP	0.099	0.019	5.181	<0.001
Proportion Graduate / Major Med School Affiliation	0.165	0.020	8.193	<0.001
Proportion Short-Term Acute Care Hospitals	0.308	0.022	13.965	<0.001
Residual standard error: 0.919 on 2421 degrees of freedom				
Multiple R-squared: 0.158, Adjusted: 0.155				
F-statistic: 50.7 on 9 and 2421 DF, p-value:<0.001				

Table 7. Optimal power transformations, Box-Cox methods

Variable	Power
Admissions for Diagnoses	-.0673
Discharges	0.290
ER Visits	0.417
Surgeries	0.364
Acute Days	0.238
Net Income	0.541
Operating Profit Margin	0.530
Cash on Hand	0.919
Equity	0.344
Staffed Beds	0.170
Affiliated Physicians	0.226
Employees	0.086
Medicare	0.788
Medicaid	0.032

Table 8. Hierarchical analysis suggests all blocks are important

Block	Residual Degrees of Freedom (df)	Residual Sum of Squares	df	Sum of Squares	F	Pr(>F)
Technical	32404	53.933				
Workload	32400	44.398	4	9.595	1853.223	<.0001
Financial	32396	43.882	4	0.516	99.609	<.0001
Geospatial	32343	41.864	53	2.018	29.416	<.0001

Table 9. Most important variables and associated parameter estimates based on number of variables in the model

# Variables:	1	2	3	4	5	6	7	8
Intercept	0.790	0.775	0.762	0.760	0.809	0.831	0.831	0.831
Discharges	-0.011	-0.011	-0.011	-0.009	-0.010	-0.010	-0.010	-0.010
DRG 293		0.056	0.069	0.070	0.070	0.071	0.071	0.071
DRG 292			0.026	0.026	0.026	0.027	0.027	0.027
Acute-Care Hospital				-0.025	-0.033	-0.051	-0.050	-0.048
Medicare					-0.065	-0.068	-0.066	-0.065
Critical Access						-0.025	-0.025	-0.025
Major Med. School							0.014	0.013
Voluntary Non-Profit								-0.006
R^2	0.545	0.680	0.703	0.719	0.733	0.741	0.744	0.745

Table 10. Coefficients of determination for the five tree ensemble models

Model	R^2
Extreme Gradient Boosting	0.975
Extra Trees	0.975
Gradient Boosting	0.974
Random Forest	0.971
Bagging	0.970

Table 11. The variables and the importance factors associated with them, averaged over all tree ensemble models

Variable	Importance
Discharges	0.331
DRG__DRG291	0.287
Medicare	0.063
Acute_Days	0.035
Year__Y16	0.031
ER_Visits	0.029
Hospital_Type__Short Term Acute Care Hospital	0.026
State__VA	0.023
Affiliated_Physicians	0.023
State__MD	0.017
Hospital_Type__Rehabilitation Hospital	0.017
Staffed_Beds	0.016
State__FL	0.015
Surgeries	0.014
DRG__DRG292	0.011
State__OK	0.011
Medical_School_Affiliation__Major	0.010
State__IL	0.010
State__NE	0.010
State__OH	0.010
DRG__DRG293	0.010
Employees	0.009
State__NC	0.009
State__MI	0.009
Equity	0.008
Medicaid	0.007
Cash_on_Hand	0.007
Profit_Margin	0.006
Net_Income	0.006
State__CT	0.003
Year__Y18	0.002
Year__Y17	0.002

Figures

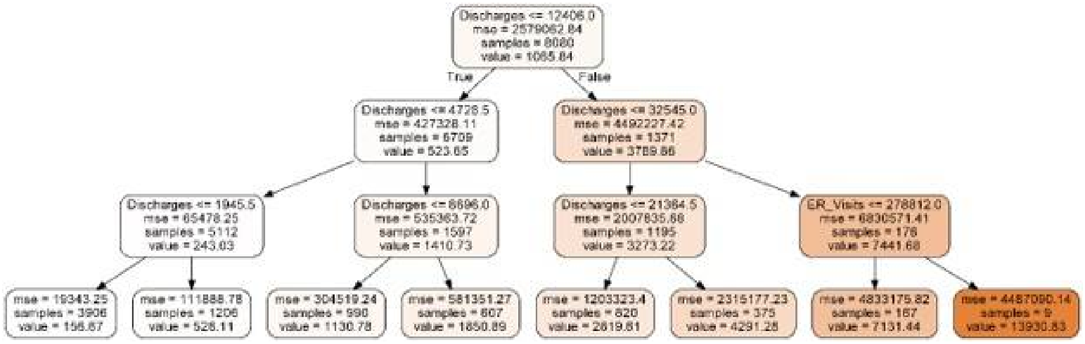


Figure 1

An example of a tree model to classify opioid admissions.

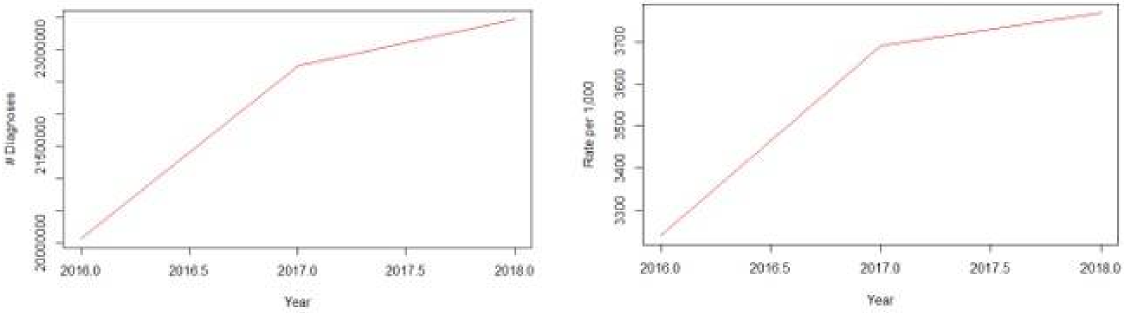


Figure 2

Number and rates of DRGs as a function of year

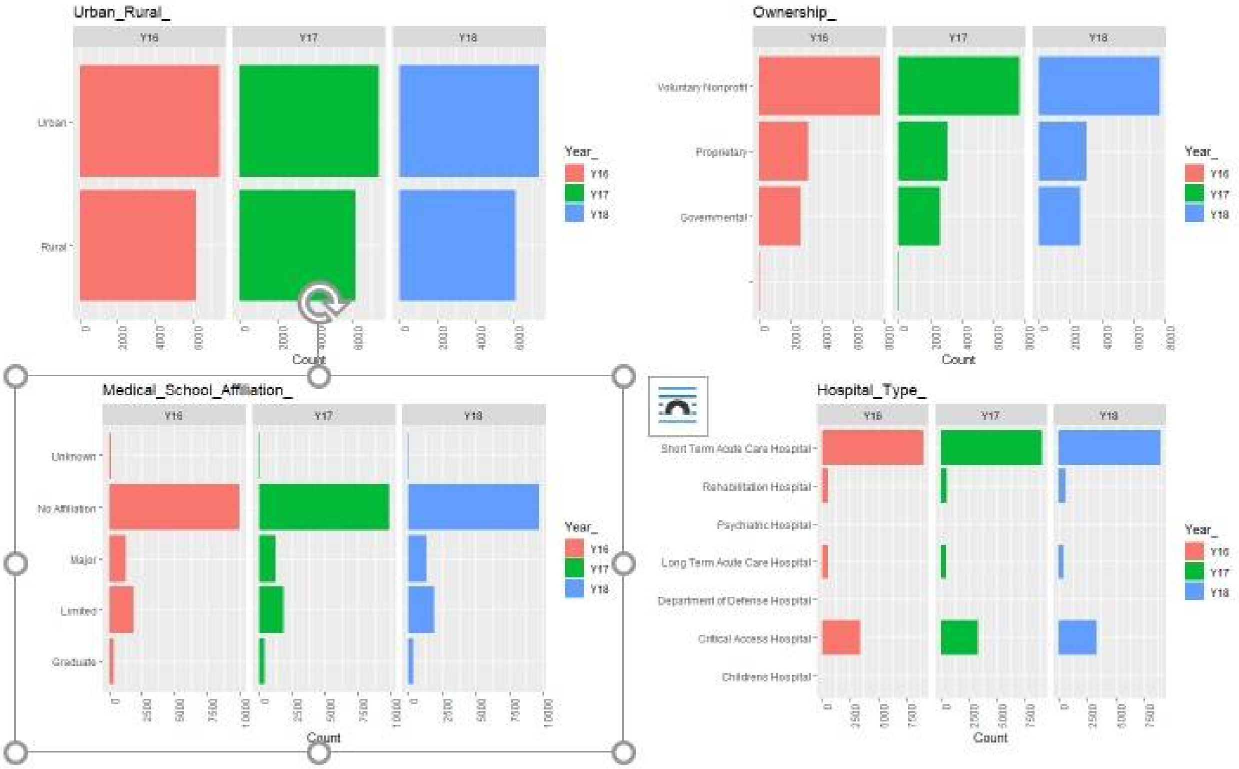


Figure 3

Categorical variables by year

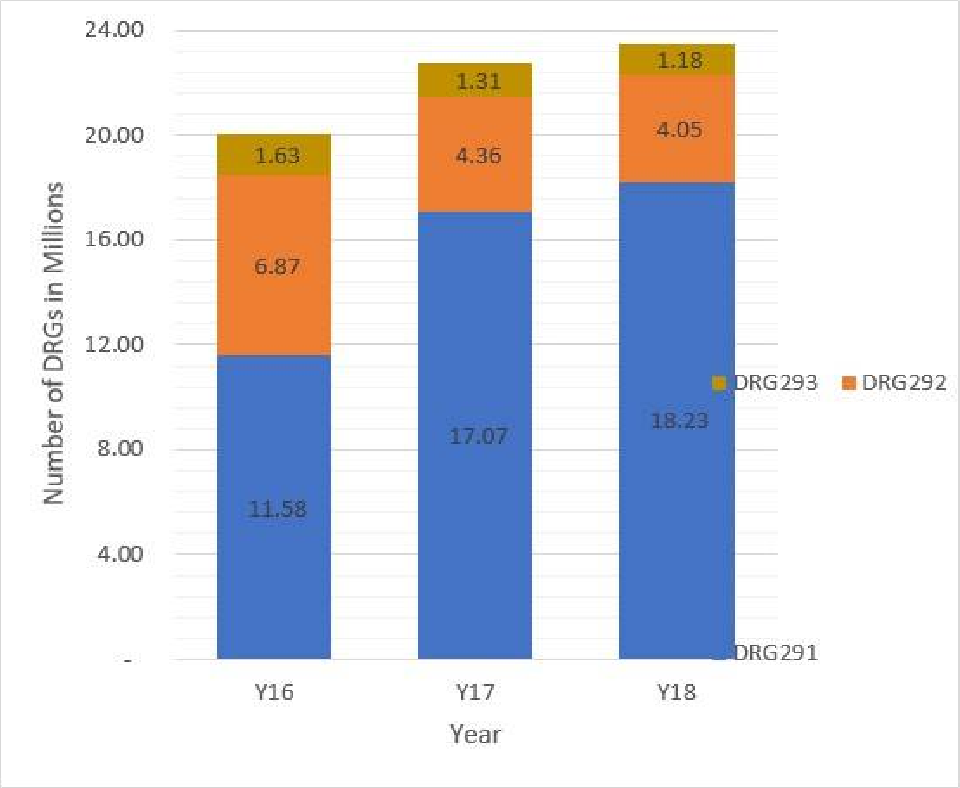


Figure 4

Number of DRGs by type (left axis) and cost estimates by DRG type and total, 2016 through 2018

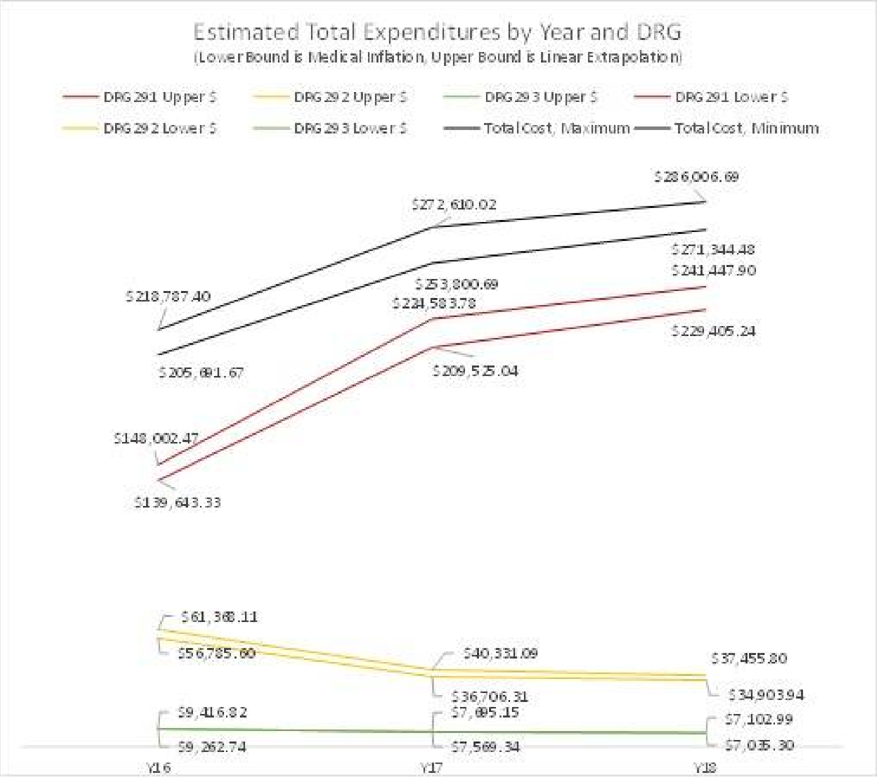


Figure 5

Associated cost estimates in billions (total and by DRG) per year

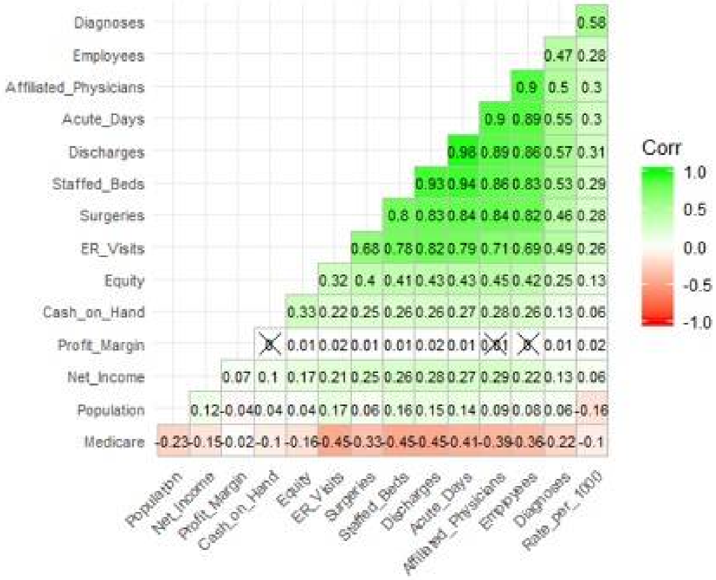


Figure 6

Hierarchical clustered correlation of quantitative variables

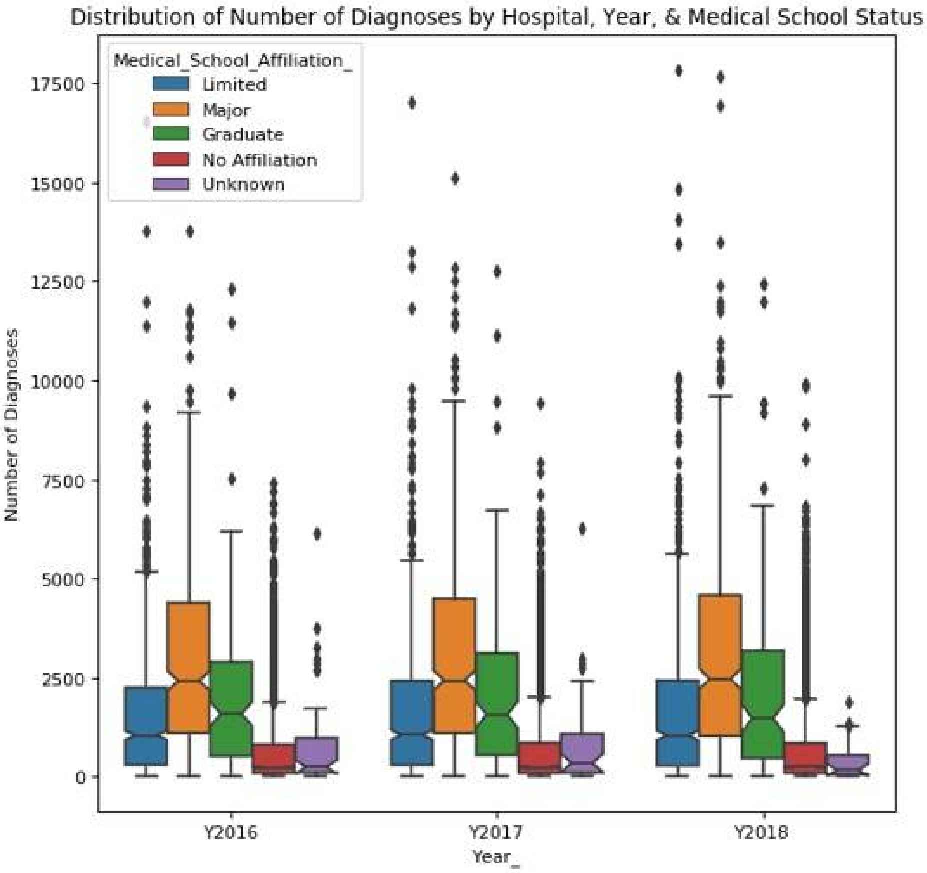


Figure 7

Number of diagnoses by year by medical school affiliation

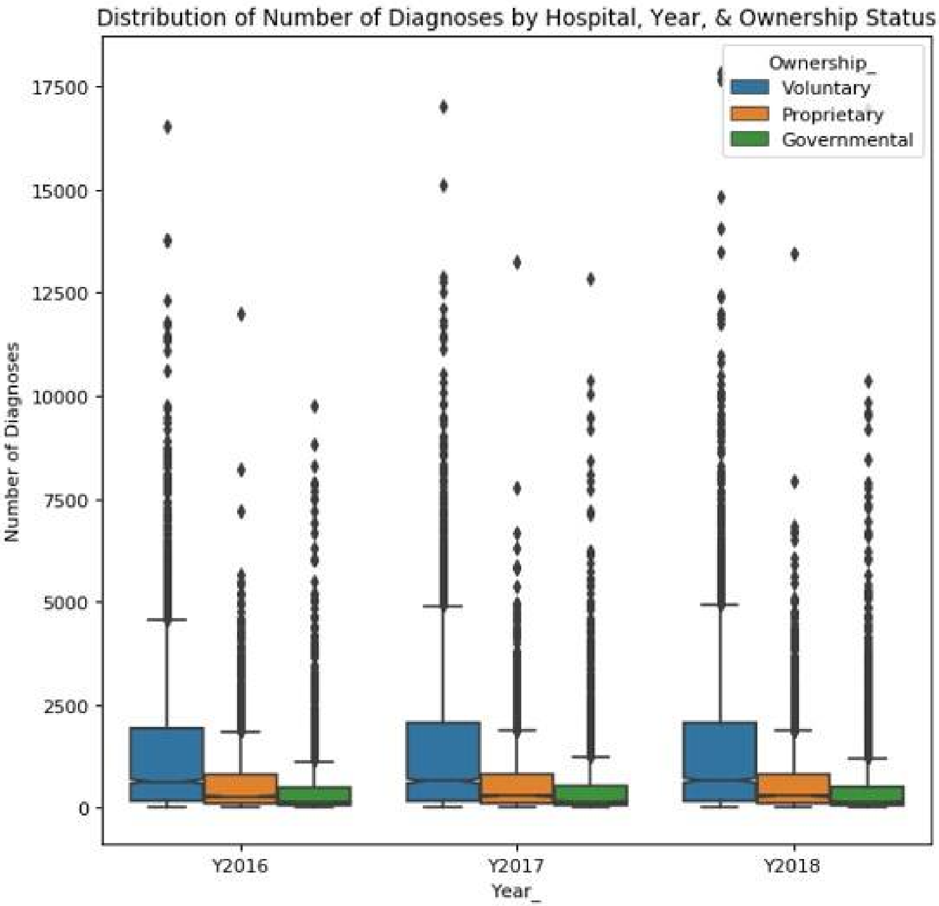


Figure 8

Number of diagnoses by year by type of hospital

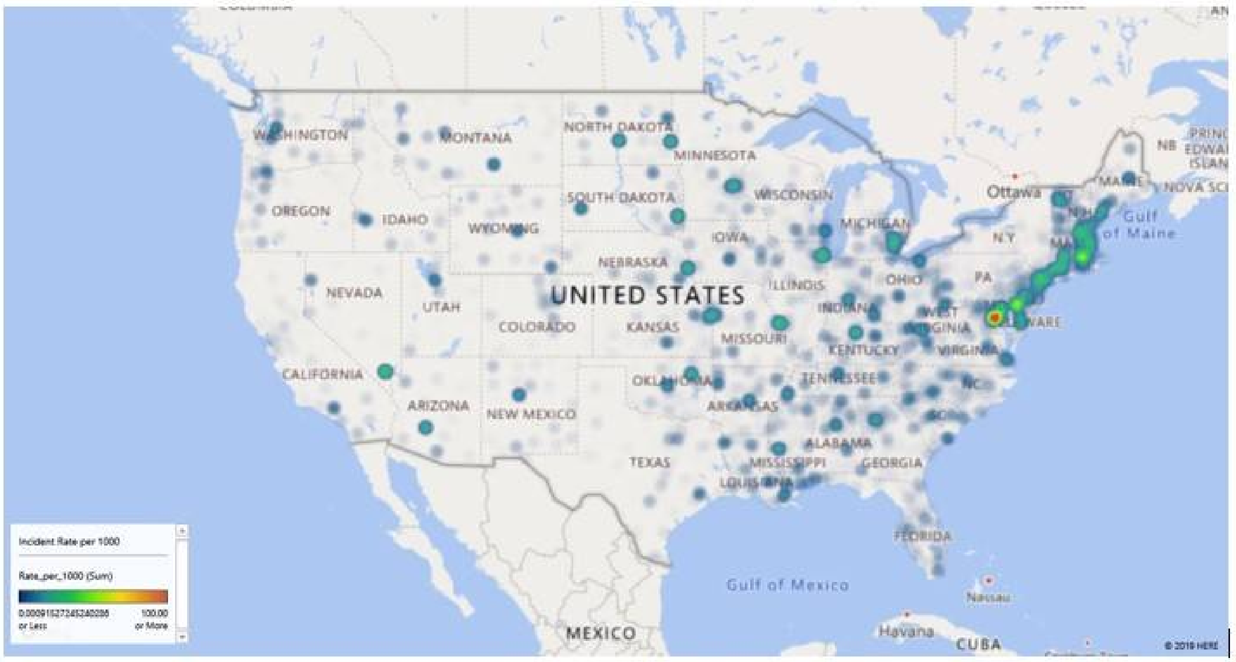


Figure 9

DRG rates per 1000, 2016

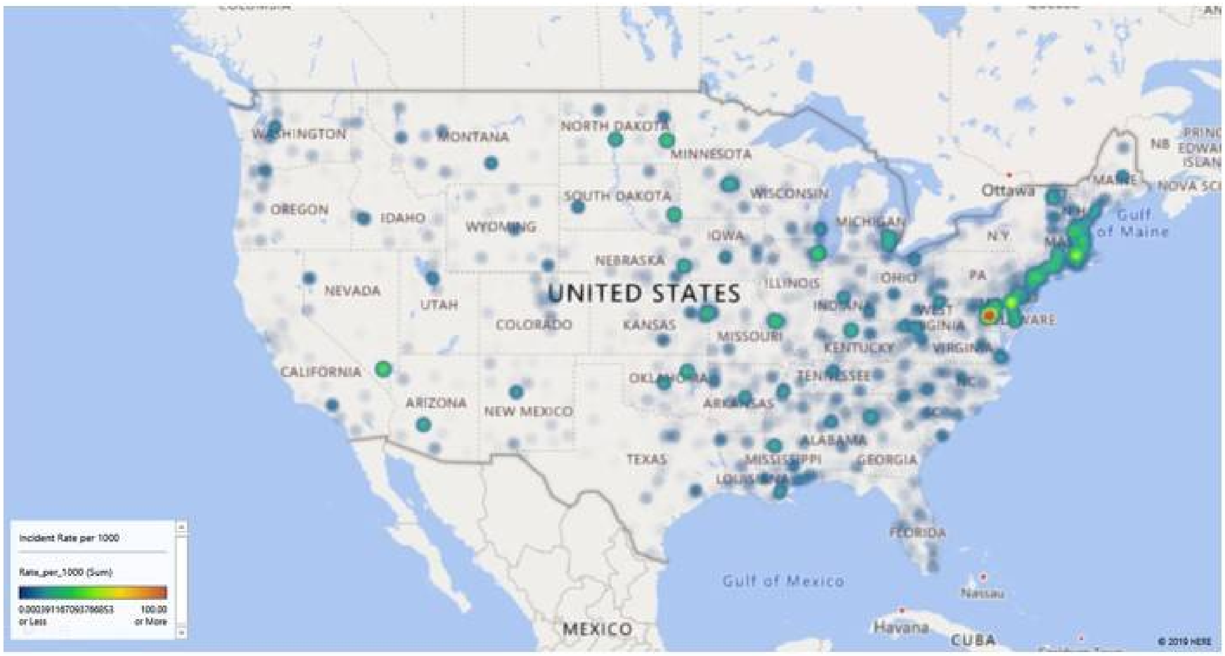


Figure 10

DRG rates per 1000, 2017

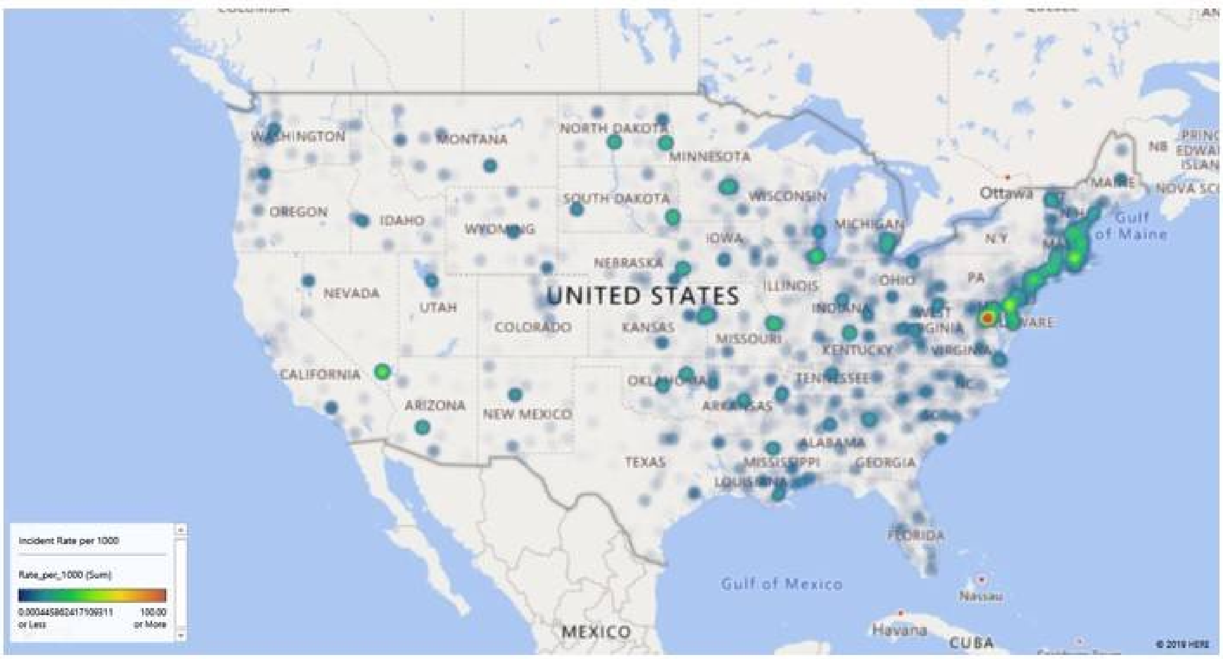
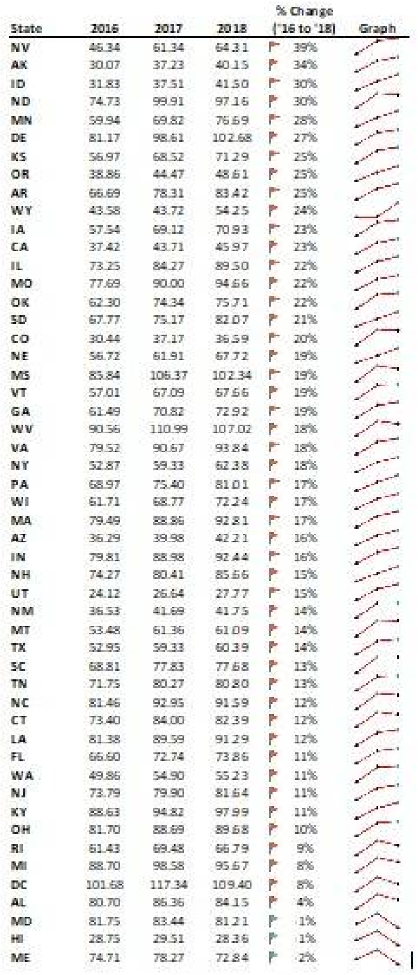


Figure 11

DRG rates per 1000, 2018



State	2016	2017	2018	% Change ('16 to '18)	Graph
NV	46.34	61.34	64.31	⚑ 39%	
AK	30.07	37.23	40.15	⚑ 34%	
ID	31.83	37.51	41.50	⚑ 30%	
ND	74.73	99.91	97.16	⚑ 30%	
MN	59.94	69.82	76.69	⚑ 28%	
DE	81.17	98.61	102.68	⚑ 27%	
KS	56.97	68.52	71.29	⚑ 25%	
OR	38.86	44.47	48.61	⚑ 25%	
AR	66.69	78.31	83.42	⚑ 25%	
WY	43.58	43.72	54.25	⚑ 24%	
IA	57.34	69.12	70.93	⚑ 23%	
CA	37.42	43.71	45.97	⚑ 23%	
IL	73.25	84.27	89.50	⚑ 22%	
MO	77.69	90.00	94.66	⚑ 22%	
OK	62.30	74.34	75.71	⚑ 22%	
SD	67.77	75.17	82.07	⚑ 21%	
CO	30.44	37.17	36.59	⚑ 20%	
NE	56.72	61.91	67.72	⚑ 19%	
MS	85.84	106.37	102.34	⚑ 19%	
VT	57.01	67.09	67.66	⚑ 19%	
GA	61.49	70.82	72.92	⚑ 19%	
WV	90.56	110.99	107.02	⚑ 18%	
VA	79.52	90.67	93.84	⚑ 18%	
NY	52.87	59.33	62.38	⚑ 18%	
PA	68.97	75.40	81.01	⚑ 17%	
WI	61.71	68.77	72.24	⚑ 17%	
MA	79.49	88.86	92.81	⚑ 17%	
AZ	36.29	39.98	42.21	⚑ 16%	
IN	79.81	88.98	92.44	⚑ 16%	
NH	74.27	80.41	85.66	⚑ 15%	
UT	24.12	26.64	27.77	⚑ 15%	
NM	36.53	41.69	41.75	⚑ 14%	
MT	53.48	61.36	61.09	⚑ 14%	
TX	52.95	59.33	60.39	⚑ 14%	
SC	68.81	77.83	77.68	⚑ 13%	
TN	71.75	80.27	80.80	⚑ 13%	
NC	81.46	92.95	91.59	⚑ 12%	
CT	73.40	84.00	82.39	⚑ 12%	
LA	81.38	89.59	91.29	⚑ 12%	
FL	66.60	72.74	73.86	⚑ 11%	
WA	49.86	54.90	55.23	⚑ 11%	
NJ	73.79	79.90	81.64	⚑ 11%	
KY	88.63	94.82	97.99	⚑ 11%	
OH	81.70	88.69	89.68	⚑ 10%	
RI	61.43	69.48	66.79	⚑ 9%	
MI	88.70	98.58	95.67	⚑ 8%	
DC	101.68	117.34	109.40	⚑ 8%	
AL	80.70	86.36	84.15	⚑ 4%	
MD	81.75	83.44	81.21	⚑ -1%	
HI	28.75	29.51	28.36	⚑ -1%	
ME	74.71	78.27	72.84	⚑ -2%	

Figure 12

Diagnoses per 1,000 by year by state

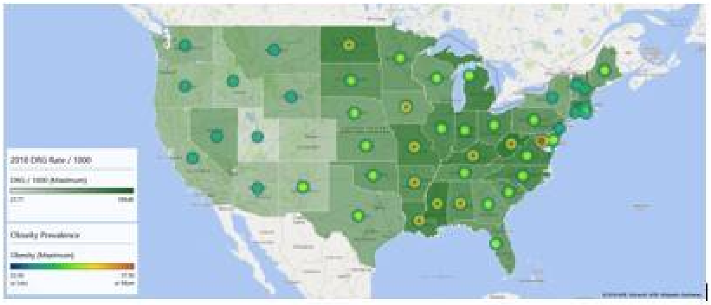


Figure 13

Map of DRG Rates / 1000 versus obesity prevalence

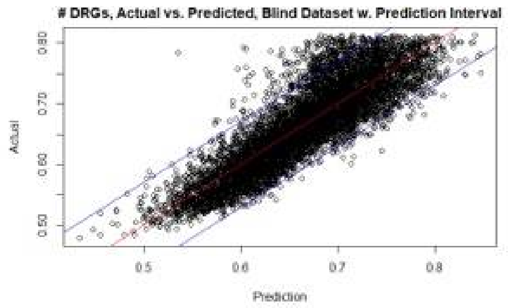


Figure 14

Plot of actual test-set data versus predictions from the training set