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

Posted Date: October 22nd, 2020

DOI: https://doi.org/10.21203/rs.2.21707/v3

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. The number of diagnoses is forecast using regression (constrained and unconstrained) and ensemble (random forests, extra trees regressor, gradient boosting, and bagging) techniques at the hospital unit of analysis. Descriptive maps of heart failure diagnosis-related groups (DRGs) depict areas of high incidence. State and county level spatial and non-spatial regression models of heart failure admission rates are performed. Expenditure forecasts were calculated for 2016 through 2018.

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 since 2016. The best predictive model for forecasting the number of diagnoses at the hospital unit of analysis was an extremely randomized tree ensemble (predictive R^2 = 0.86 applied to a 20% unobserved test set.) The important variables in this model included workload metrics and hospital type. State level spatial lag models using 1st order Queen's criteria were best at estimating heart failure admission rates (R^2 =.816). At the county level, OLS was preferred over any GIS model based on a statistically insignificant Moran's I and resultant R^2; however, none of the traditional models performed well (R^2=.169 for the OLS). Gradient boosted tree models were able to predict 36% of the total Sum of Squares; however, and the most important factors were facility workload, mean cash-on-hand of the hospitals in the county, and mean equity of those hospitals.. Online interactive maps at the state and county levels are provided.

Conclusions. Heart failure and associated expenditures are increasing. 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. Models like the ones presented here that reliably forecast demand are needed to inform healthcare leaders.

1. Introduction

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.

Geospatial analyses are becoming increasingly important and relevant in the area of public health [24]. With visualization and analytical capabilities, GIS in public health has become widespread [25-33]. A review of geospatial methods used for public health as well as the historical development is available from Saran, Singh, Kumar, & Chauhan [24].

Specific 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 [34] as well as individual and contextual correlates of cardiovascular disease [35]. Spatial analysis in the area of public health is conducted at the worldwide, country, and regional levels of analysis [36]. 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 [36]. Spatial regression techniques such as simultaneous autoregressive (SAR) models are often used to document health risks [37], and spatial clustering has been used for leprosy in Brazil [38], measle vaccination in sub-Saharan Africa [39], as well as food and physical activity in the United States [40]. Spatial recognition has been used to identify congenital heart disease in youth aged 4-18 in China as well [41]. Geospatial mapping has been used for describing birthing incidence [42], the opioid epidemic [43], evaluating back surgery growth over time [44], 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 by state and county and to build predictive models that forecast hospitals' number of heart failure DRGs using technical, workload, financial, and geospatial variables. Analysis and forecasting of the raw numbers of heart failure DRGs provides for financial and demand estimation based on empirical inflationary pressures and existing / forecast disease rate increases. Understanding patterns is important to both policymakers, epidemiologists, and health administrators alike for cost control and planning efforts. Finally, the demand and supply analysis highlight potential shortfalls that may require redress.

2. Methods

2.1. Data

Data from this study come from the Definitive Healthcare dataset [45]. 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) [45]. State and county-level population data for rate calculations were from the Census Bureau [7, 46]. 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 admissions as defined by Diagnostic-Related Groups 291, 292, and 293 [47]. 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 l 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.

Table 1. Independent variables

Technical Variables	Defined	Measurement
% 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
Affiliated Physicians	Number of physicians affiliated with hospital	Integer
Employees	Number of direct employees of hospital	Integer
Staffed Beds	Number of staffed beds operated by hospital	Integer
Surgeries	Number of surgeries performed	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

2.3. Models for Number of Heart Failure Admissions

2.3.1. Train and Test Sets

For the non-spatial model exploration, 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 32,206 and 8,051, respectively. Models were built on the training set and evaluated on the test set. The splitting of the data occurred prior to any imputation or transformations, so that no information would be leaked from one set to the other. The primary model selection metric for non-spatial models was R^2, the proportion of the sum of squares accounted for by variables in the model.

2.3.2 Imputation, Transformation, and Scaling

Very little data were missing (2%). Observations with 33% or more missing data were deleted. Imputation was conducted separately for the total data, the training data, and the test data. Six workload-related variables (discharges, emergency room visits, surgeries, staffed beds, affiliated physicians, and employees) were highly collinear and replaced with a single principal component that accounted for 84.8% and 84.9% of the variability in the total and training datasets, respectively. The workload variable for the test set was built using the linear combination estimated on the training set to avoid information leakage between the two data sets. Data were then scaled between zero and unity, as some models are not scale invariant.

2.3.3. Explanatory Analysis for the Number of Heart Failure Diagnoses

Linear regression, lasso regression, robust regression, Elastic Net regression, random forests, extra trees random forests, extreme gradient-boosted random forests, and bagging regressors estimate the DRG heart failure admissions. To

investigate the bias-variance trade-off [48], we built multiple models on an 80% training and evaluated on a 20% test set. 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) [48]. Elastic Net combines both Lasso and Ridge penalty functions [49]. While coefficients are easily interpreted in regression-type models, the data typically require scaling and transformations. 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, principal component analysis is performed.

2.3.4. Tree Models

Random forests are an ensemble of de-correlated tree models. Figure 1 is an example of a tree with three branches that includes a random subset of candidate features (variables). 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). In a random forest, every tree produces a separate forecast. All trees produced are than averaged to produce the estimate. Trees are "pruned," to prevent overfitting [48].

Extremely randomized regression trees (extra trees) add randomness by generating random split locations for features and using the best threshold as the splitting rule [50]. These models typically result in less variance but higher bias.

Gradient boosting is an ensemble of weighted trees composed by iteratively assigning weights to trees that reduce prediction error [51]. 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* [48].

A bagging tree regressor (or bootstrap aggregation) uses random subsets of the data to generate estimates, which are then aggregated to form a final prediction [52]. A good implementation and discussion of bagging regressors is available from the Python scikit-learn module [53].

2.4. Geospatial Analysis, State and County Heart Failure Admission Rates

Hospital data were aggregated at the state and county levels and then merged with geospatial data. All states had admissions, so no data were missing. Some counties had no admissions likely due either to lack of hospital facilities or small populations. A single principal component was used for the 6 workload variables as before, accounting for 98% of the variability at the state level and 97% of the variability at the county level.

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 [46].

The determination about whether to use spatial regression / error models (see Mahara et al. [54]) or simple spatial mapping was informed by residual diagnostics of regression models (i.e., Moran's I global test of residuals with post-hoc Lagrangian multiplier diagnostics as required). In all cases, choropleth maps of regression residuals are mapped to provide a visual indicator of outliers [55]. Spatial models are compared to non-spatial models for both coefficients and performance metrics (i.e., R^2). Spatial contiguity was modeled using both Rook (edge borders considered neighbors) and Queen (edge and vertex borders considered neighbors) criterion. Only first order Rook and Queen criteria were evaluated,

as distance effects are expected to fade. Further, we used the row-standardized sums for weighting neighbors and a zero-policy that allowed for weight vectors of zero length for areas with unconnected neighbors.

2.5. Changes in DRGs

The significance of changes for 2016 to 2018 (DRG rates) are also 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 [56].

2.6. Software

All analysis was performed in Anaconda Python Release 3.8 [57], R Statistical Software, and Microsoft Excel 2016 [29]. Python was used primarily for tree models, while R provided regression analysis and geospatial analyses. The primary geospatial packages included *tmap* [58], *sf* [59], *sp* [60], and *spatialregression* [60]. Online interactive maps were done in the R package, *leaflet* [61].

3. Results

3.1. Descriptive Statistics-Quantitative Data

Descriptive statistics for all data are freely available online at https://rpubs.com/R-Minator/heart [62]. Roll-up for the quantitative data are provided in Table 2. The average hospital observation during any given year had about 1,600 observations of DRG 291, 292, and 293 (median of 383). That same hospital had about 147 staffed beds (median of 86), 7 thousand discharges (median of 2.8 thousand), and about 6.4 thousand surgeries (median of 4.5 thousand). The average hospital had positive income (in millions) of $17.3 (median of $2.03), significant cash-on-hand ($20.3 thousand, median of $1.95 thousand), and positive equity. The typical hospital had just over 1,000 employees (median of 436) with 232 affiliated physicians (median of 104) and was reimbursed 45% by Medicare (median of 42%). Only 9% reimbursement was from Medicaid (median of 6%).

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

n=40,257	Mean	SD	Median	Min	Max
Number DRGs	1,640.258	3,334.942	385	11	57,461
Staffed Beds	146.507	172.468	86	2	2,753
Affiliated Physicians	231.786	353.461	104	1	4,328
Employees	1,008.034	1,683.991	436	4	26,491
Percent Medicare	0.448	0.186	0.422	0	0.983
Percent Medicaid	0.087	0.091	0.063	0	0.869
Discharges	7,014.259	9,908.036	2,811	1	129,339
ER Visits	32,864.497	3,3976.188	25,085	0	543,457
Surgeries	6,349.317	7,987.273	4,464	0	130,741
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

Year over year, both DRGs and rates of DRGs per 1000 population increased as illustrated in Figure 2 and Figure 3, respectively. 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.

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 1.7 million, 1.6 million, and 1.5 million, 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, 103, and 94, respectively. Utah, Hawaii, and Colorado had the smallest average rates, 26, 29, and 35, respectively.

Most hospitals were in urban settings (58%). Fifty-two percent were voluntary non-profits with 29% proprietary and 18.7% governmental. The vast majority (75%) had no affiliation with a medical school and were short-term care facilities (60%). Nearly no hospitals were classified as Department of Defense (DoD) or children's hospitals.

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.

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

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 [63]. 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.

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

Both estimates are reasonably 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 [64]. Figure 4 illustrates the number of DRGs by year, while Figure 5 shows the associated aggregate cost estimates.

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) illustrates tight relationships among many variables. This type of 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 and staffed beds are most closely associated with the number of diagnoses, our primary variable of interest. More importantly, the workload variables appear to have significant collinearity that must be addressed for regression-based models.

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 hospitals with major medical school affiliations experience 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).

3.5. Explanatory Models for Heart Failure Diagnoses, Hospital Unit of Analysis

3.5.1. Regression Models

Linear, lasso, and elastic net regression evaluated the number of diagnoses as a function of all other variables. Models built on the training set and applied to the training and test sets resulted in predicted R^2 values of 0.501, 0.328, and 0.417 (training) and 0.454, 0.323, 0.348 (test) for the OLS, lasso, and elastic net models, respectively. The OLS model predicted better than the constrained regression models and did not overfit. Table 5 provides the coefficient estimates for all variables after fitting on the entire dataset

$$(F_{73,40183} = 534, p < .001, R^2 = 0.492).$$

Table 5. Results of regression analyses for the number of diagnoses, hospital unit of analysis

	Linear		Lasso	Elastic Net	Variable	Linear		Lasso	Elastic Net
l	-0.439	***	-0.298	-0.323	State_MA	0.010	**	0.000	0.000
ne	-0.043	***	0.000	0.000	State_MD	0.013	***	0.000	0.000
rgin	0.026	***	0.000	0.000	State_ME	-0.002		0.000	0.000
Hand	-0.082	*	0.000	0.000	State_MI	0.013	***	0.000	0.000
	0.012	***	0.000	0.000	State_MN	0.003		0.000	0.000
are	0.029	**	0.000	0.000	State_MO	0.005		0.000	0.000
aid	-0.002	***	0.000	0.000	State_MS	0.003		0.000	0.000
ry Ownership	0.003		0.000	0.000	State_MT	0.004		0.000	0.000
it Ownership	0.006	***	0.000	0.000	State_NC	0.016	***	0.000	0.000
Med Sch Aff	0.002	***	0.000	0.000	State_ND	0.004		0.000	0.000
ed Sch Aff	-0.009		0.000	0.000	State_NE	0.001		0.000	0.000
Sch Aff	0.000	***	0.000	0.000	State_NH	0.002		0.000	0.000
n Med Sch Aff	-0.005		0.000	0.000	State_NJ	0.012	***	0.000	0.000
ccess Hospital	0.074		0.000	0.000	State_NM	-0.002		0.000	0.000
pital	0.000	***	0.000	0.000	State_NV	0.007		0.000	0.000
spital	0.048	***	0.000	0.000	State_NY	-0.005		0.000	0.000
spital	0.067	***	0.000	0.000	State_OH	0.007	*	0.000	0.000
ospital	0.058	***	0.000	0.000	State_OK	0.001		0.000	0.000
spital	0.084	***	0.000	0.006	State_OR	-0.002		0.000	0.000
	0.005		0.000	0.000	State_PA	0.002		0.000	0.000
	0.003		0.000	0.000	State_RI	0.002		0.000	0.000
	-0.002		0.000	0.000	State_SC	0.006		0.000	0.000
	0.000		0.000	0.000	State_SD	-0.002		0.000	0.000
	-0.003		0.000	0.000	State_TN	0.003		0.000	0.000
	0.013	**	0.000	0.000	State_TX	0.004		0.000	0.000
	0.006		0.000	0.000	State_UT	-0.004		0.000	0.000
	0.018	**	0.000	0.000	State_VA	0.014	***	0.000	0.000
	0.006		0.000	0.000	State_VT	-0.001		0.000	0.000
	0.009	**	0.000	0.000	State_WA	0.005		0.000	0.000
	-0.003		0.000	0.000	State_WI	0.003		0.000	0.000
	0.001		0.000	0.000	State_WV	0.002		0.000	0.000
	0.000		0.000	0.000	State_WY	0.002		0.000	0.000
	0.009	**	0.000	0.000	Urban	0.004	***	0.000	0.000
	0.006		0.000	0.000	Year 2017	0.003	***	0.000	0.000
	0.002		0.000	0.000	Year 2018	0.004	***	0.000	0.000
	0.005		0.000	0.000	DRG 292	-0.040	***	0.000	-0.016
	0.006		0.000	0.000	DRG 293	-0.056	***	-0.013	-0.029

*p<.05, **p<.01, ***p<.001

Very few coefficients are recommended by the lasso and elastic net models. The lasso model suggest that the workload principal component and DRG 293 are important predictors, both of which are associated with reduced diagnoses ceterus parabus. Elastic net was similar in recommending inclusion of workload as well as DRG 292 and DRG 293, all associated with reduced diagnoses. The OLS model had a larger array of variables that were statistically significant, and the coefficients of the largest magnitude for the min-max scaled variables were associated with workload (-0.439), Short-Term Acute Care hospitals (STAC, 0.084), cash-on-hand (0.082), and Critical Access Hospitals (CAH, 0.082). When evaluated by categorical groups, the most significant variables were workload (0.302 additional R^2), DRGs (0.162 additional R^2), and hospital type (0.011 additional R^2).

3.5.2. Tree Ensemble Models

Random forests, extra trees regression, gradient boosting, and bagging regressors after some hyperparameter tuning on the training set predicted heart failure diagnoses on the test set with reasonable accuracy (R^2 = 0.829, 0.862, 0.821, and 0.830, respectively) . 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. All models accounted for more variance than any regression model.

The best performing tree ensemble was the extra trees regression. This model ensembled 50 trees and resulted in variable importance shown in Figure 9. Similar to the regression models, hospital type (STAC / LTAC), workload (PC), and DRG (DRG 293) were important along with the state of Utah.

The conclusion for both the regression and tree models is that hospital-level diagnoses by DRG are forecastable and that workload along with hospital type are important in doing so. Further, the models indicated that geography might be important, as individual state variables and urban / rural status were important to the OLS and the tree models. These models were evaluated on the hospital unit of analysis for raw diagnoses numbers. Rate-based admission models were then evaluated for the states and counties.

3.6. State Level Geospatial Analysis

A descriptive analysis of heart failure from 2016-2018 using geographical informat systems was conducted to evaluate regional differences. Primarily, we were interested in rates per standardized unit in the population of the geographical area. Populations were based on Census Bureau estimates for each geographic region [7, 46]. The state level analysis was limited in that only 50 states and Washington, D.C. were included (n=51).

Results of the state GIS analysis are available here: https://rpubs.com/R-Minator/HeartState [65]). 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 experienced the highest average admission rate for diagnoses of heart failure (109.5 per 1000), which might be due to the large presence of military and veteran care facilities) followed by West Virginia (102.8 per 1000), Mississippi (98.2 per 1000), Michigan (94.3 per 1000), Delaware (94.2 per 1000), Kentucky (93.8 per 1000), North Dakota (90.6 per 1000), North Carolina (88.7 per 1000), Virginia (88.0 per 1000), and Missouri (87.5 per 1000) . Of interest is that previous studies indicate these states also see many admissions due to the opioid crisis [43].

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.

Further, evaluating obesity prevalence intensity from the Centers for Disease Control and Prevention (CDC) shows significant correlation between obesity and DRGs per 1000 [66]. A Spearman's test for correlation of obesity prevalence and 2018 DRGs per 1000 was statistically significant with r=0.689, S=6,867.7, p<.001.

Ordinary Least Square regression was performed on the state admission rate as a function of the quantitative, aggregated variables. While the model was statistically significant and accounted for reasonable variability

$$(F_{10,40} = 4.5, p < 0.001, R^2 = 0.529),$$

only the proportion Medicare was significant at the 0.05 level. 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 [58] as well as residuals associated separate linear models for all included variables is available as an interactive GIS map

here: https://rpubs.com/R-Minator/heart [62]. The spatial residuals shows some spatial correlation. The visual check was confirmed by a global test for both Queen and Rook neighbors suggest that spatial relationships exist, (I=0.309, p<.001 and I=0.306, p<.001, respectively) [67]. Lagrangian Multiplier Diagnostics (non-robust and robust) suggested that the preferred models would be spatial lag rather than spatial error, as robust tests for error models were insignificant while lag models remained significant (see [62]).

Generalized spatial two-stage least squares estimated Queen, and Rook models, while a comparison linear model was estimated in traditional fashion. R^2 for OLS, Queen, and Rook models were 0.529, 0.816, and 0.809, respecctively. Queen and Rook models performed better on the state-aggregated data. The coefficient results of the spatial models were nearly identical to each other, while the OLS was obviously needed geospatial data to improve its performance (see Table 6). The geographic component for Queen and Rook models were statistically significant along with mean profit margin and the proportion of facilities with major medical school affiliation were important in predicting diagnoses rates.

Table 6. Results of the state-level regression

Variable	Linear Model		Queen Model		Rook Model	
Rho			0.993	***	0.993	***
(Intercept)	-0.221		-0.101		-0.123	
Income	-0.055		0.011		0.01	
Profit Margin	-0.418		-0.458	**	-0.458	**
Cash on Hand	-0.162		0.015		0.023	
Equity	0.183		0.060		0.049	
% Medicare	0.842	***	0.221		0.24	
% Medicaid	-0.163		0.058		0.061	
% Non-Profit	0.129		-0.128		-0.122	
% Med School	0.386		0.398	***	0.408	***
% STAC	0.483	**	-0.016		-0.015	
Workload	-0.004		-0.162		-0.152	

*p<.05, **p<.01, ***p<.001

3.7. County-Level Spatial Analysis

3.7.1. Interactive, Online Maps

The average three-year heart failure admissions per 1000 county population are shown in the interactive map online [68]. These county maps show that the admissions are generally (as expected) in large metropolitan areas, e.g., Dekalb, Illinois (0.65 per 1000). There are exceptions, however. For example, Montour, Pennsylvania is a small county that is home to a large Geisinger facility and thus has a higher than expected admission rate (1.00 per 1000).

The top ten counties for average rates per 1000 over three-years were Winchester, Virginia (3.33 per 1000); Norton, Virginia (3.21 per 1000), Montour, Pennsylvania (3.01 per 1000); Fredericksburg, Virginia (2.13 per 1000); DeKalb, Illinois (1.95 per 1000); Harrisonburg, Virginia (1.58 per 1000); Petersburg, Virginia (1.57 per 1000); Boyd, Kentucky (1.45 per 1000); St. Francois, Missouri (1.34 per 1000); and Adams, North Dakota (1.21 per 1000). Of interest is that half of these counties are in the state of Virginia, perhaps due to the large military and veteran medical centers located in the area. Many of these counties (e.g., Winchester) are small but have large healthcare facilities.

3.7.2. Regression Models, County-Level of Analysis

Similar to what was done at the state level, an exploratory spatial regression model using first-order Queen and Rook contiguity criterion to evaluate the importance of geography was performed using rolled, Z-scaled, county-level

independent variables on the county-level admission rate variable (admissions per population in each county). Moran's I global test suggested that the OLS model was probably sufficient (I=0.02, p=0.100); however, we explored further with Lagrangian multiplier diagnostics. The robust from of these statistics slightly favored a lag model. Results of the regression are in Table 7, and the residual maps for the global model and the individual variables are available online [62] The OLS, Queen, and Rook regression models accounted for only a small fraction of the sum of the squares (R^2 = .169, R^2 = .132, R^2 = .132, respectively).

Table 7. Regression table for county analysis

Variable	Linear Model		Queen Model		Rook Model	
Rho	-0.539	***	-0.538	***		
(Intercept)	0.019	0.048	***	0.047	***	
Income	0.010	0.015	0.015			
Profit Margin	-0.007	*	-0.002	-0.002		
Cash on Hand	-0.063	*	-0.058	***	-0.057	***
Equity	0.090	*	0.081	***	0.081	***
% Medicare	0.049	**	0.050	***	0.050	***
% Medicaid	0.012	*	-0.001	-0.001		
% Non-Profit	0.013	**	0.016	***	0.016	***
Mean Affiliated Providers	0.045	**	0.044	***	0.044	***
% STAC	0.041	**	0.044	***	0.044	***
Workload	0.084	*	0.079	***	0.079	***

Moran's I favors the linear model, but all coefficients are similar.
*p<.05, **p<.01, ***p<.001

Most variables in all models are statistically significant largely due to the sample size, but the coefficients are of small magnitude. Every variable except for the proportion Medicaid was statistically significant in the best-fiting OLS model, and yet the magnitude of the coefficients across the three models (OLS, Queen, Rook) was quite similar. Profit margin, and cash-on-hand were negatively associated with admission rates in the OLS model, ceterus parabus. All other variables had positive coefficients in the OLS model. Interpretation of directionality must be done cautiously, as the variables act together in prediction.

Interactive maps of the admission rate, model residuals (OLS, Queen, and Rook), as well as residuals for individual variables are provided online [62]. The residual maps are not suggestive of spatial autocorrelation given the residual dispersion by county. Future explanatory models can omit spatial correlation.

Given the small contribution of the OLS, Queen and Rook models to estimating county-level admission rates, ensemble models were investigated at the county level. With 2,431 valid observations, sufficient power existed to split the data into training and test sets (80% / 20%). Results of hyperparameter tuned models suggested that extreme gradient boosting was the best model as the predictive R^2 for random forests, extra trees, extreme gradient boosting, and bagging regressors was 0.242, 0.292, 0.264, and 0.130, respectively. The first three models performed much better on the 20% withhold set than regression models. The variable importance analysis suggested that workload, cash-on-hand, and mean equity were the most important variables with importance scores of 0.35, 0.13, and 0.13, respectively.

4. Discussion

4.1. Review of Findings

From Figure 2, 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.

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) [69]. A comorbidity is a condition that existed before admission. A complication is any condition occurring after admission, not necessarily a complication of care [70]. 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 [71-73]. DRG 291, the most expensive DRG, is associated with $91 billion cost increases from 2016.

Tree models at the hospital unit of analysis were capable of capturing close to 90% of the variability on a 20% withhold test set. The variables of most importance to this prediction were consistent with the variables found through regression modeling. Specifically, hospital type and workload variables captured by a single principal component re important in predicting the number of diagnoses by facility. This finding is particularly useful in that localities and states may forecast expected costs associated with the ever-increasing number of heart failure admissions with hospital-level granularity.

For the state level GIS analysis, OLS as well as 1st order Queen and Rook models were estimated, and the coefficients were both stable and congruent across models. Facility ownership and hospital type were important predictors, and the 1st-order Queen provided the best variance capture (R^2=.816). No sophisticated predictive models were available due to the small samples size (50 states plus the District of Columbia).

At the county level, no regression models performed well, although the OLS model captured more variance than the 1st order Queen and Rook models. In fact, the best performing model after hyperparameter tuning was an extreme gradient boosted tree ensemble that captured 36% of the variability on an unseen test set.

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 treatment 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 [74-76]. 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 [77, 78].

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 [79]. 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 [80, 81]. 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 [82, 83]. 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 [84, 85].

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 [86]. 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 [87]. 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 [88] .

List Of Abbreviations

ER. Emergency Room

CC. 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/ [45]. All Python and R code are available here: https://rpubs.com/R-Minator/heart [62], https://rpubs.com/R-Minator/HeartState [65], and https://rpubs.com/R-Minator/HeartCounty [68].

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. Niakouei A, Tehrani M, Fulton L: **Health Disparities and Cardiovascular Disease**. In: *Healthcare: 2020*: Multidisciplinary Digital Publishing Institute; 2020: 65.
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**(13):7.
3. **Heart Disease Facts | cdc.gov** [https://www.cdc.gov/heartdisease/facts.htm]
4. Steenman M, Lande G: **Cardiac aging and heart disease in humans**. *Biophysical Reviews* 2017, **9**(2):131-137.
5. Lakatta EG, Levy D: **Arterial and Cardiac Aging: Major Shareholders in Cardiovascular Disease Enterprises**. *Circulation* 2003, **107**(2):346-354.

6. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, Das SR, Ferranti Sd, Després J-P, Fullerton HJ *et al*: **Heart Disease and Stroke Statistics—2016 Update**. *Circulation* 2016, **133**(4):e38-e360.
7. **Population Clock** [https://www.census.gov/popclock/population_widget_200x402.php?component=density&no_scode#us]
8. Pencina MJ, Navar AM, Wojdyla D, Sanchez RJ, Khan I, Elassal J, D'Agostino RB, Peterson ED, Sniderman AD: **Quantifying Importance of Major Risk Factors for Coronary Heart Disease**. *Circulation* 2019, **139**(13):1603-1611.
9. Fodor JG, Tzerovska R: **Coronary heart disease: is gender important?** *The Journal of Men's Health & Gender* 2004, **1**(1):32-37.
10. Berkowitz SA, Berkowitz TSZ, Meigs JB, Wexler DJ: **Trends in food insecurity for adults with cardiometabolic disease in the United States: 2005-2012**. *PLOS ONE* 2017, **12**(6):e0179172.
11. Dalen JE, Alpert JS, Goldberg RJ, Weinstein RS: **The epidemic of the 20(th) century: coronary heart disease**. *Am J Med* 2014, **127**(9):807-812.
12. Jones DS, Podolsky SH, Greene JA: **The Burden of Disease and the Changing Task of Medicine**. *New England Journal of Medicine* 2012, **366**(25):2333-2338.
13. Cole HM, Fiore MC: **The War Against Tobacco: 50 Years and Counting**. *JAMA* 2014, **311**(2):131-132.
14. **Morbidity & Mortality: 2012 Chart Book on Cardiovascular,Lung, and Blood Diseases** [https://www.nhlbi.nih.gov/files/docs/research/2012_ChartBook]
15. **Prevalence of coronary heart disease–United States, 2006-2010**. *MMWR Morb Mortal Wkly Rep* 2011, **60**(40):1377-1381.
16. Reynolds I, Page RL, Boxer RS: **Cardiovascular Health and Healthy Aging**. In: *Healthy Aging: A Complete Guide to Clinical Management*. edn. Edited by Coll PP. Cham: Springer International Publishing; 2019: 31-51.
17. **Heart Failure** [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]
18. **Number of People per Active Physician by Specialty, 2015** [https://www.aamc.org/data-reports/workforce/interactive-data/number-people-active-physician-specialty-2015]
19. Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, Finkelstein EA, Hong Y, Johnston SC, Khera A *et al*: **Forecasting the Future of Cardiovascular Disease in the United States**. *Circulation* 2011, **123**(8):933-944.
20. Tsao CW, Lyass A, Enserro D, Larson MG, Ho JE, Kizer JR, Gottdiener JS, Psaty BM, Vasan RS: **Temporal Trends in the Incidence of and Mortality Associated With Heart Failure With Preserved and Reduced Ejection Fraction**. *JACC: Heart Failure* 2018, **6**(8):678-685.
21. Kan HJ, Kharrazi H, Chang H-Y, Bodycombe D, Lemke K, Weiner JP: **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**(3):e0213258.
22. Rajalaxmi ASA, Rajalaxmi RR, Abdullah AS, R R: **A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier**. In: *IJCA Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls: 22/4/2012 2012*: International Journal of Computer Applications IJCA; 2012: 22-25.
23. Cutler A, Cutler DR, Stevens JR: **Random Forests**. In: *Ensemble Machine Learning: Methods and Applications*. edn. Edited by Zhang C, Ma Y. Boston, MA: Springer US; 2012: 157-175.
24. Saran S, Singh P, Kumar V, Chauhan P: **Review of Geospatial Technology for Infectious Disease Surveillance: Use Case on COVID-19**. *Journal of the Indian Society of Remote Sensing* 2020, **48**(8):1121-1138.
25. Ahmad R, Ali WNWM, Nor ZM, Ismail Z, Hadi AA, Ibrahim MN, Lim LH: **Mapping of mosquito breeding sites in malaria endemic areas in Pos Lenjang, Kuala Lipis, Pahang, Malaysia**. *Malaria Journal* 2011, **10**(1):361.

26. Hanafi-Bojd AA, Vatandoost H, Oshaghi MA, Charrahy Z, Haghdoost AA, Zamani G, Abedi F, Sedaghat MM, Soltani M, Shahi M *et al*: **Spatial analysis and mapping of malaria risk in an endemic area, south of Iran: A GIS based decision making for planning of control**. *Acta Tropica* 2012, **122**(1):132-137.
27. Kolivras KN: **Mosquito Habitat and Dengue Risk Potential in Hawaii: A Conceptual Framework and GIS Application**. *The Professional Geographer* 2006, **58**(2):139-154.
28. Martin C, Curtis B, Fraser C, Sharp B: **The use of a GIS-based malaria information system for malaria research and control in South Africa**. *Health & Place* 2002, **8**(4):227-236.
29. Nykiforuk CIJ, Flaman LM: **Geographic Information Systems (GIS) for Health Promotion and Public Health: A Review**. *Health Promotion Practice* 2009, **12**(1):63-73.
30. Rasam ARA, Noor AMM, Ahmad N, Ghazali R: **MyGeoHealth: GIS-based cholera transmission risk system in Sabah, Malaysia**. In: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications: 4-6 March 2011 2011*; 2011: 474-479.
31. Zhang W, Wang L, Fang L, Ma J, Xu Y, Jiang J, Hui F, Wang J, Liang S, Yang H *et al*: **Spatial analysis of malaria in Anhui province, China**. *Malaria Journal* 2008, **7**(1):206.
32. Zhang Z, Jin J-m, Liu F: **The application of geographic information system (GIS) in the field of public health**. In: *2010 Second IITA International Conference on Geoscience and Remote Sensing: 28-31 Aug. 2010 2010*; 2010: 442-445.
33. Feinhandler I, Cilento B, Beauvais B, Harrop J, Fulton L: **Predictors of Death Rate during the COVID-19 Pandemic**. *Healthcare* 2020, **8**.
34. Concannon TW, Kent DM, Normand S-L, Newhouse JP, Griffith JL, Ruthazer R, Beshansky JR, Wong JB, Aversano T, Selker HP: **A Geospatial Analysis of Emergency Transport and Inter-Hospital Transfer in ST-Segment Elevation Myocardial Infarction**. *American Journal of Cardiology* 2008, **101**(1):69-74.
35. 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.
36. Bhunia GS, Shit PK: **Spatial Database for Public Health and Cartographic Visualization**. In: *Geospatial Analysis of Public Health.* edn. Cham: Springer International Publishing; 2019: 29-57.
37. 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.* edn. Edited by Maantay JA, McLafferty S. Dordrecht: Springer Netherlands; 2011: 337-356.
38. Ramos ACV, Yamamura M, Arroyo LH, Popolin MP, Chiaravalloti Neto F, Palha PF, Uchoa SAdC, Pieri FM, Pinto IC, Fiorati RC *et al*: **Spatial clustering and local risk of leprosy in São Paulo, Brazil**. *PLOS Neglected Tropical Diseases* 2017, **11**(2):e0005381.
39. Brownwright TK, Dodson ZM, van Panhuis WG: **Spatial clustering of measles vaccination coverage among children in sub-Saharan Africa**. *BMC Public Health* 2017, **17**(1):957.
40. Wende ME, Stowe EW, Eberth JM, McLain AC, Liese AD, Breneman CB, Josey MJ, Hughey SM, Kaczynski AT: **Spatial clustering patterns and regional variations for food and physical activity environments across the United States**. *International Journal of Environmental Health Research* 2020:1-15.
41. 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.
42. MacQuillan EL, Curtis AB, Baker KM, Paul R, Back YO: **Using GIS Mapping to Target Public Health Interventions: Examining Birth Outcomes Across GIS Techniques**. *Journal of Community Health* 2017, **42**(4):633-638.
43. Fulton L, Dong S, Zhan B, Kruse CS, Stigler-Granados P: **Geospatial-Temporal and Demand Models for Opioid Admissions, Implications for Policy**. *Journal of Clinical Medicine* 2019, **8**(7):993.

44. Fulton L, Kruse CS: **Hospital-Based Back Surgery: Geospatial-Temporal, Explanatory, and Predictive Models**. *J Med Internet Res* 2019, **21**(10):e14609.
45. **Definitive Healthcare** [https://www.definitivehc.com/]
46. Bureau USC: **County Population by Characteristics: 2010-2018**. In. Washington, DC, USA: U.S. Census Bureau; 2020.
47. **ICD-10-CM/PCS MS-DRG v37.0 Definitions Manual** [https://www.cms.gov/icd10m/version37-fullcode-cms/fullcode_cms/P0140.html]
48. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition**. New York, NY: Springer; 2009.
49. Zou H, Hastie T: **Regularization and variable selection via the elastic net**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005, **67**(2):301-320.
50. Geurts P, Ernst D, Wehenkel L: **Extremely randomized trees**. *Machine learning* 2006, **63**(1):3-42.
51. Prettenhofer P, Louppe G: **Gradient boosted regression trees in scikit-learn**. 2014.
52. Breiman L: **Bagging predictors**. *Machine learning* 1996, **24**(2):123-140.
53. 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: *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (2013)*. vol. 1309: @arxiv; 2013.
54. Mahara GM, 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**(11):1083.
55. **Finding Outliers in Models of Spatial Data** [http://www.stat.rice.edu/~scottdw/DG/dgo03.pdf]
56. 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**(200):675-701.
57. Team PC: **Python: A dynamic, open source programming language.** In.; 2015.
58. Tennekes M: **tmap: Thematic Maps in R.** *Journal of Statistical Software* 2018, **84**(6):1-39.
59. Pebesma E: **Simple Features for R: Standardized Support for Spatial Vector Data**. *The R Journal* 2018, **10**(1):439-446.
60. Bivand R, Pebesma E, Gomez-Rubio V: **Applied Spatial Data Analysis with R**. In., 2d edn. New York, NY: Springer; 2013.
61. **leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library** [https://CRAN.R-project.org/package=leaflet]
62. **Heart Failure Analysis** [https://rpubs.com/R-Minator/heart]
63. **Producer Price Index by Industry: General Medical and Surgical Hospitals (PCU622110622110) | FRED | St. Louis Fed** [https://fred.stlouisfed.org/series/PCU622110622110]
64. **National Comparisons of Commercial and Medicare Fee-for-Service Payments to Hospitals. American Health Insurance Programs Data Brief** [https://www.ahip.org/wp-content/uploads/2016/02/HospitalPriceComparison_2.10.16.pdf]
65. **Heart Failure Analysis, State Analysis** [https://rpubs.com/R-Minator/HeartState]
66. **Adult Obesity Prevalence Maps | Overweight & Obesity | CDC** [https://www.cdc.gov/obesity/data/prevalence-maps.html]
67. Moran PAP: **Notes on Continuous Stochastic Phenomena**. *Biometrika* 1950, **37**(1/2):17-23.
68. **Heart Failure Analysis, County Analysis** [https://rpubs.com/R-Minator/HeartCounty]
69. **Understanding the Impact of ICD-10 on DRGs - ICD10monitor** [https://www.icd10monitor.com/understanding-the-impact-of-icd-10-on-drgs]
70. **The ABCs of DRGs** [https://acphospitalist.org/archives/2019/05/coding-corner-the-abcs-of-drgs.htm]

71. **The high cost of heart failure for the Medicare population: An actuarial cost analysis** [https://us.milliman.com/en/insight/the-high-cost-of-heart-failure-for-the-medicare-population-an-actuarial-cost-analysis]
72. **National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013 #204** [https://www.hcup-us.ahrq.gov/reports/statbriefs/sb204-Most-Expensive-Hospital-Conditions.jsp]
73. Kilgore M, Patel HK, Kielhorn A, Maya JF, Sharma P: **Economic burden of hospitalizations of Medicare beneficiaries with heart failure**. *Risk Management and Healthcare Policy* 2017, **10**:63-70.
74. Kuehn B: **Obesity Rates Increasing**. *JAMA* 2018, **320**(16):1632-1632.
75. **Medicare Payment Advisory Commission. Report to the Congress: Medicare and the Health Care. Delivery System** [http://www.medpac.gov/docs/default-source/reports/chapter-2-the-next-generation-of-medicare-beneficiaries-june-2015-report-.pdf]
76. 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.
77. **Clinician shortage ahead? Cardiology's workforce prepares for a pair of silver tsunamis** [https://www.cardiovascularbusiness.com/clinician-shortage-ahead-cardiologys-workforce-prepares-pair-silver-tsunamis]
78. **Cardiology Workforce Analysis** [https://www.medaxiom.com/clientuploads/documents/Workforce_Analysis.pdf]
79. Recker S: **ECMO Programs: A Financial Synopsis**. *CathLab Digest* 2019, **27**(9):1.
80. **Changes to ECMO MS-DRG Assignment Impacts Hospital Payment | STS** [https://www.sts.org/advocacy/changes-ecmo-ms-drg-assignment-impacts-hospital-payment]
81. Rose RA, 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**(4):S261.
82. **Has Employment of Cardiologists Been a Successful Strategy? – Part 1 - American College of Cardiology** [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]
83. Wann S: **Consolidation and hybridization in the health care enterprise: How are cardiologists affected? Page 3**. *Cardiology Today* 2018, **Online**.
84. Dafny LS: **How Do Hospitals Respond to Price Changes?** *American Economic Review* 2005, **95**(5):1525-1547.
85. **ICD-10-CM Diagnosis Coding in HCUP Data: Comparisons With ICD-9-CM and Precautions for Trend Analyses** [https://www.hcup-us.ahrq.gov/datainnovations/ICD-10_DXCCS_Trends112817.pdf]
86. **Methamphetamine fuels the West's oil and gas boom — High Country News – Know the West** [https://www.hcn.org/issues/307/15811]
87. Ho V, Ku-Goto M-H, Jollis JG: **Certificate of Need (CON) for Cardiac Care: Controversy over the Contributions of CON**. *Health Services Research* 2009, **44**(2p1):483-500.
88. Mitchell M: **Do Certificate-of-Need Laws Still Make Sense in 2019?** *Managed Healthcare Executive* 2019, **Online**.

Figures

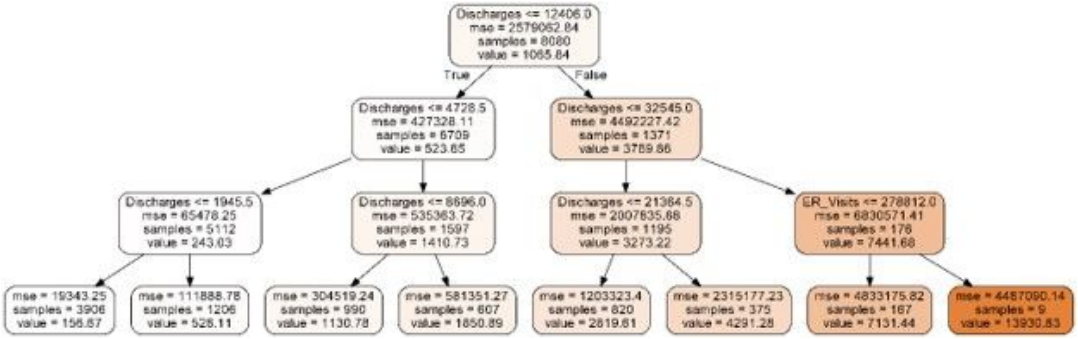

Figure 1

An example of a tree model to classify opioid admissions.

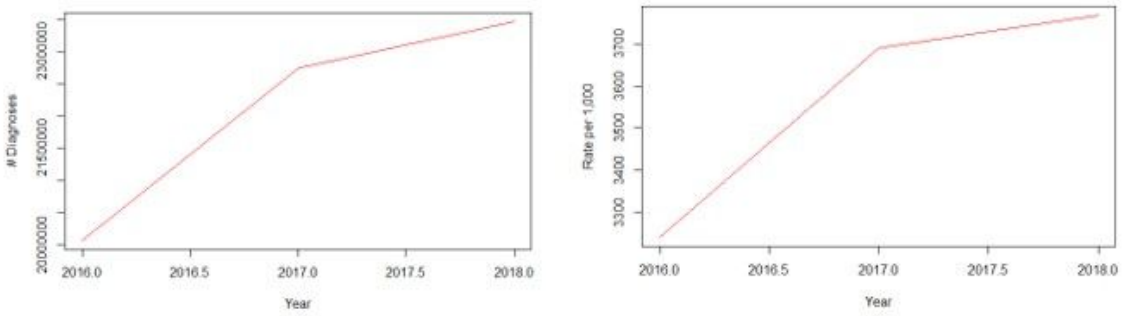

Figure 2

Number and rates of DRGs as a function of year

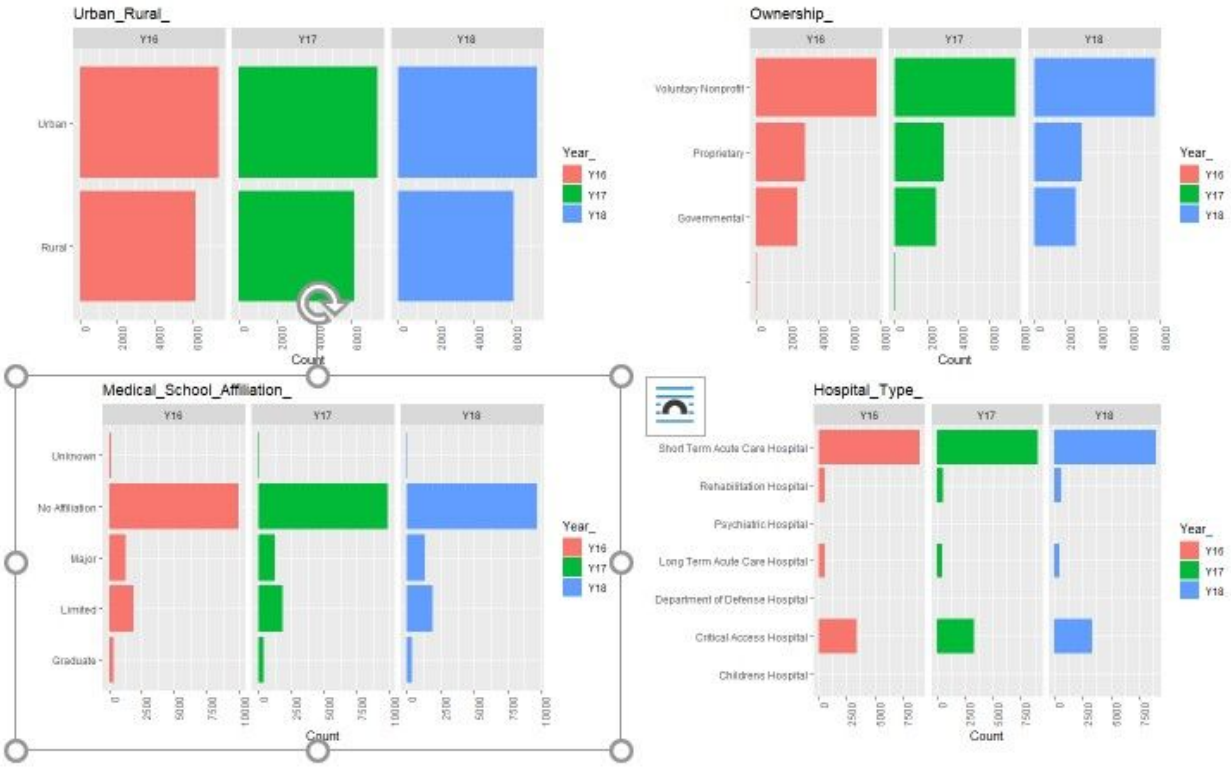

Figure 3

Categorical variables by year

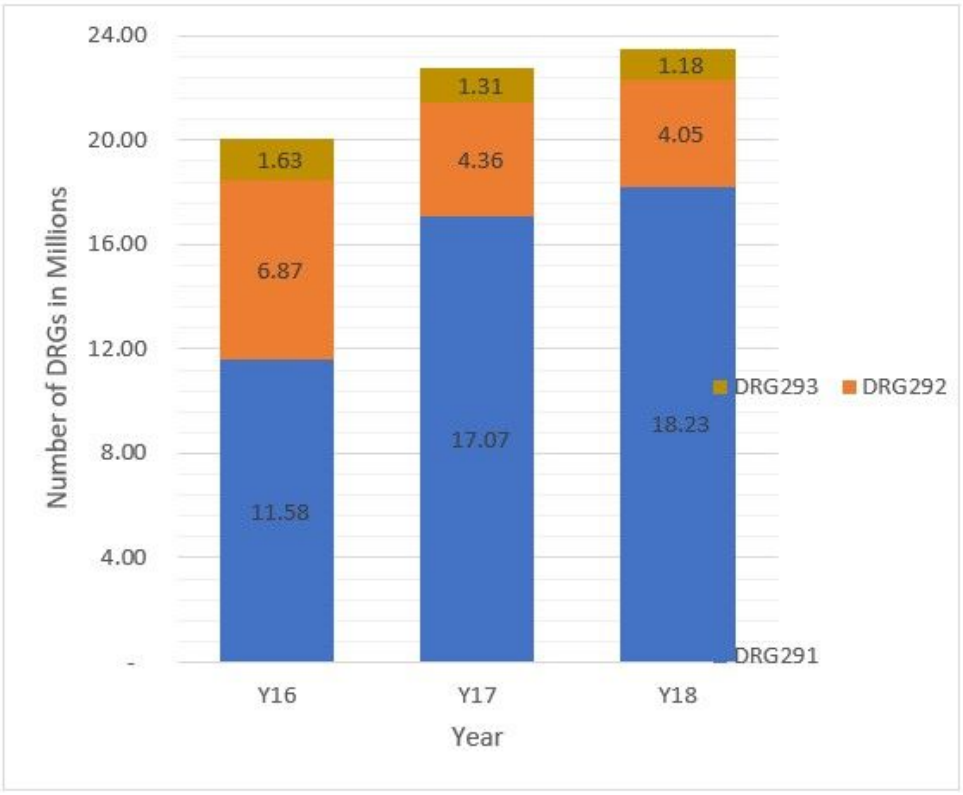

Figure 4

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

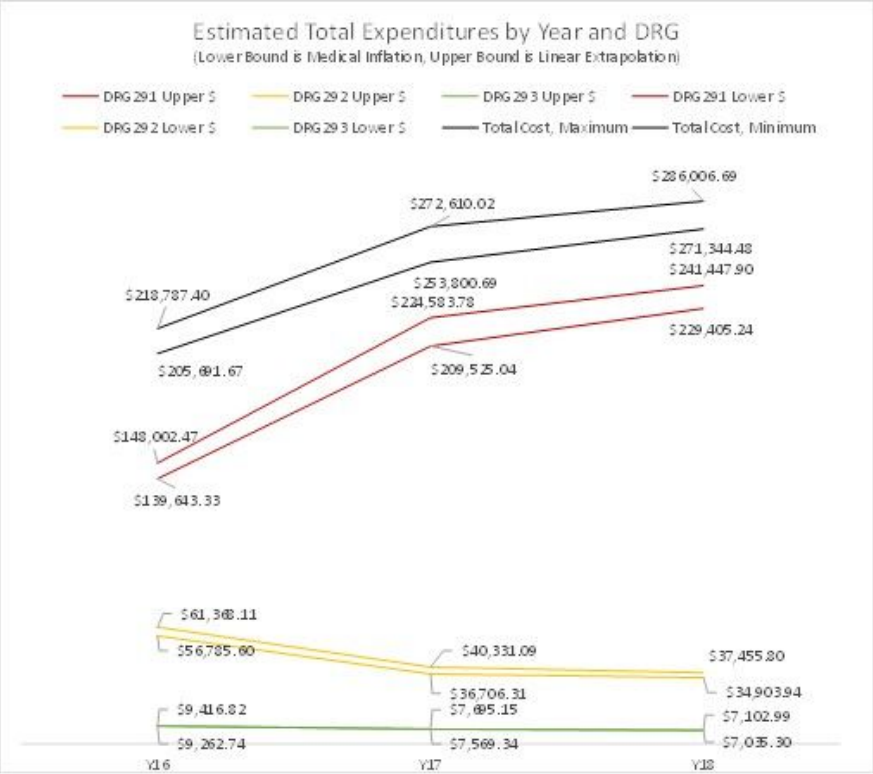

Figure 5

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

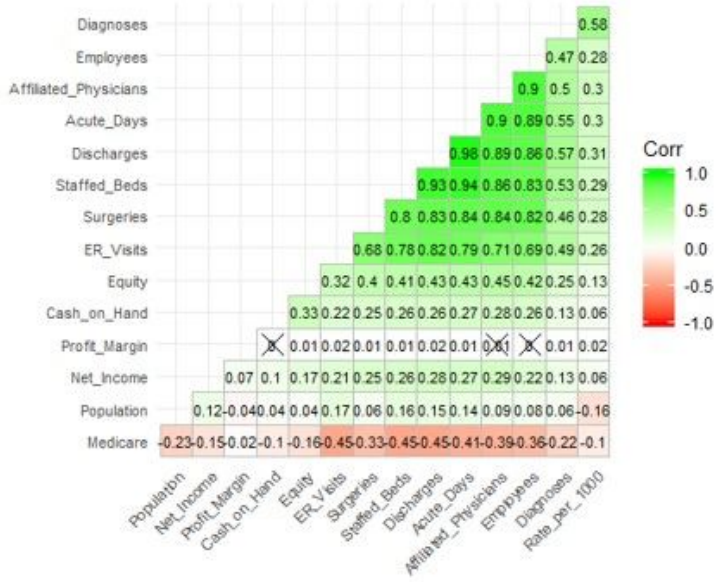

Figure 6

Hierarchical clustered correlation of quantitative variables

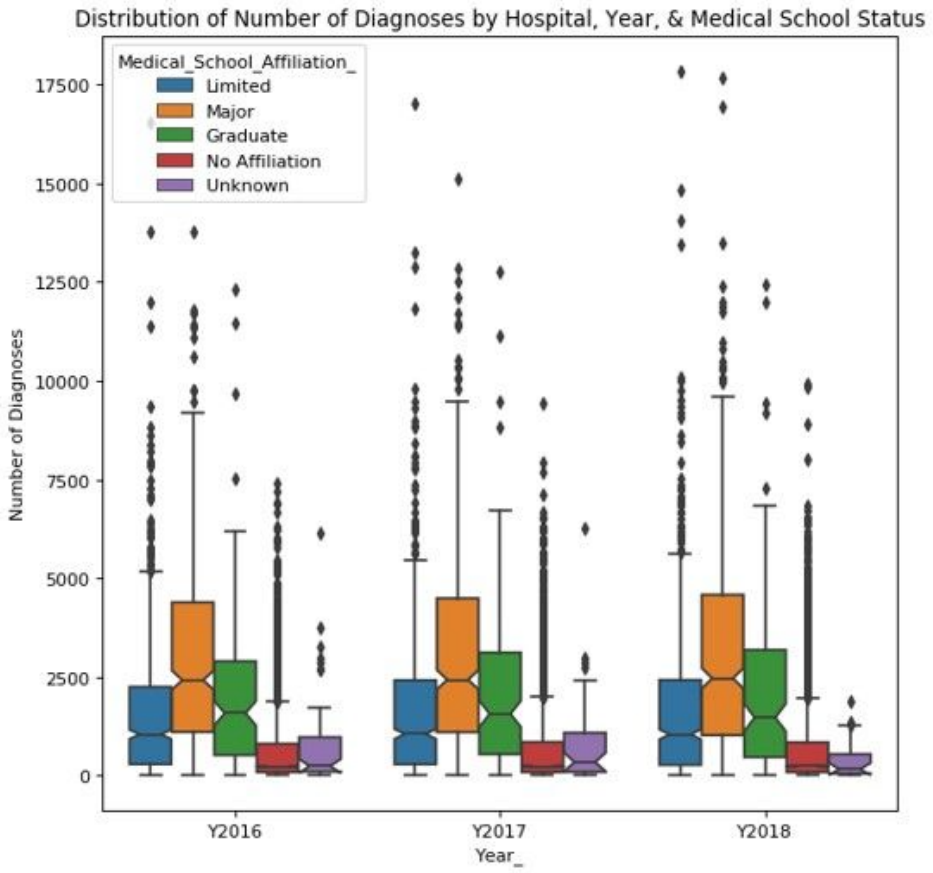

Figure 7

Number of diagnoses by year by medical school affiliation

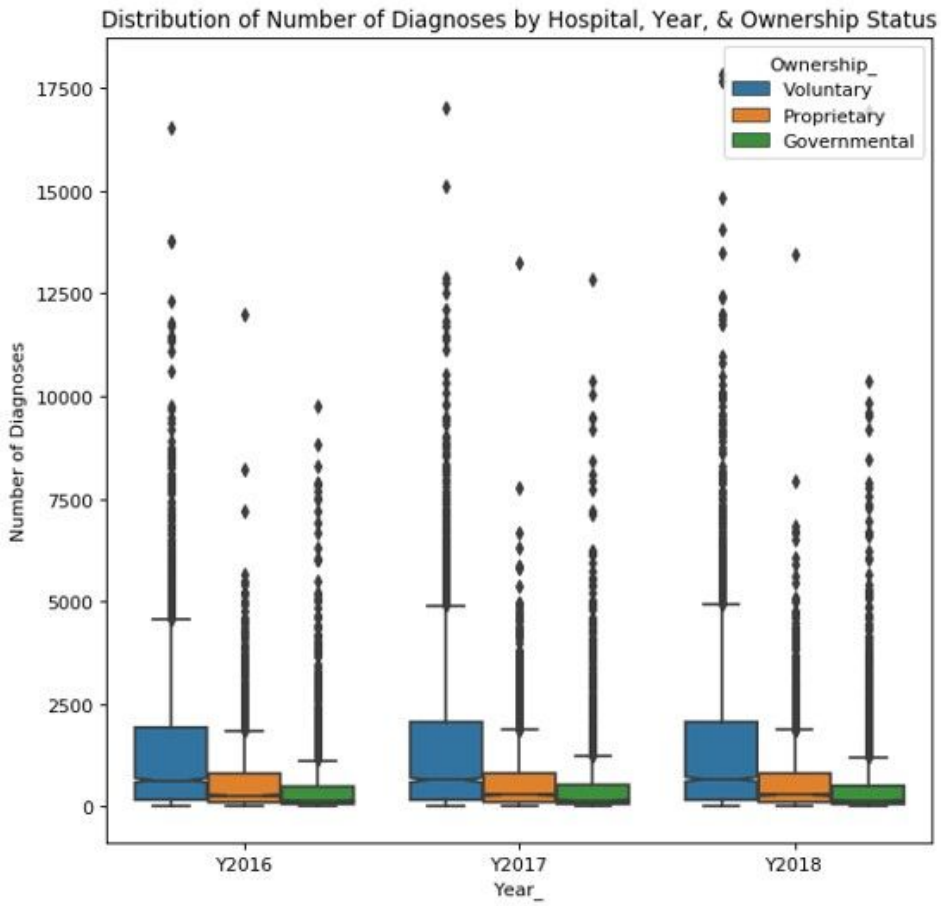

Figure 8

Number of diagnoses by year by type of hospital

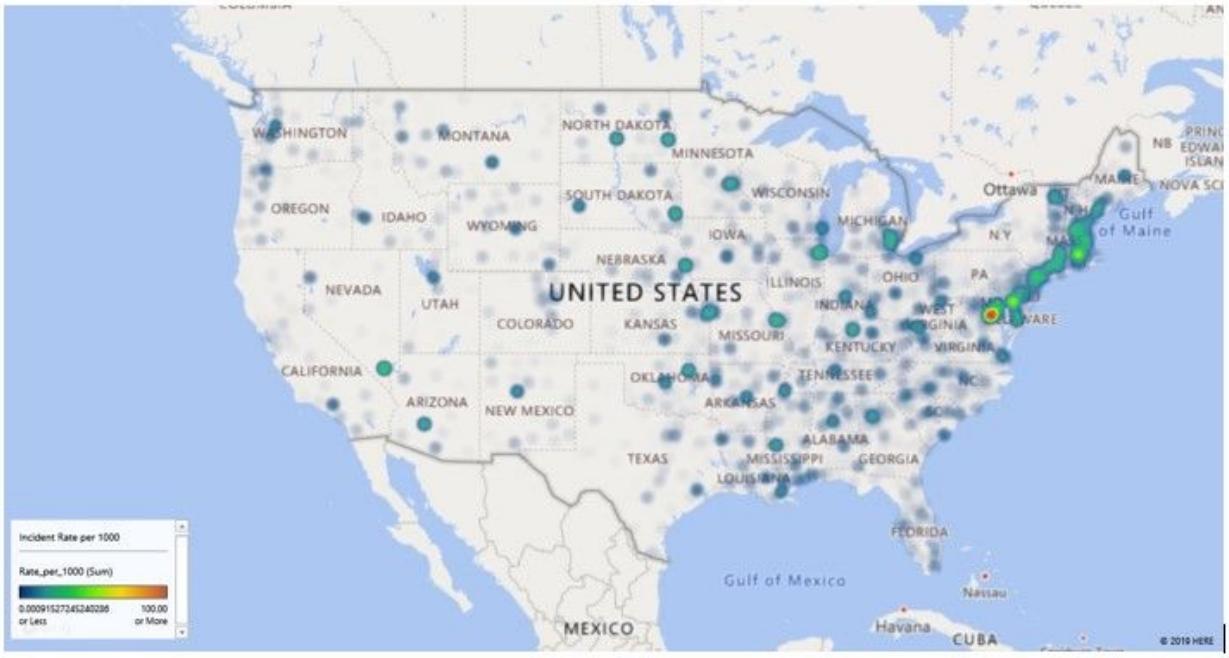

Figure 9

DRG rates per 1000, 2016

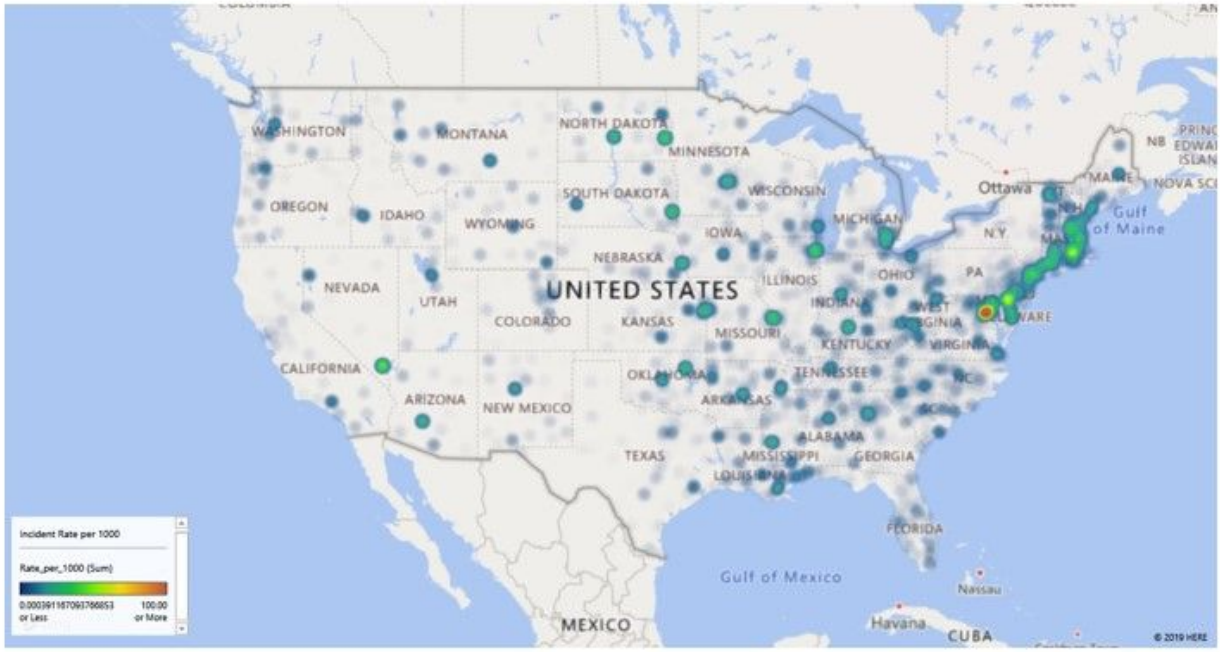

Figure 10

DRG rates per 1000, 2017

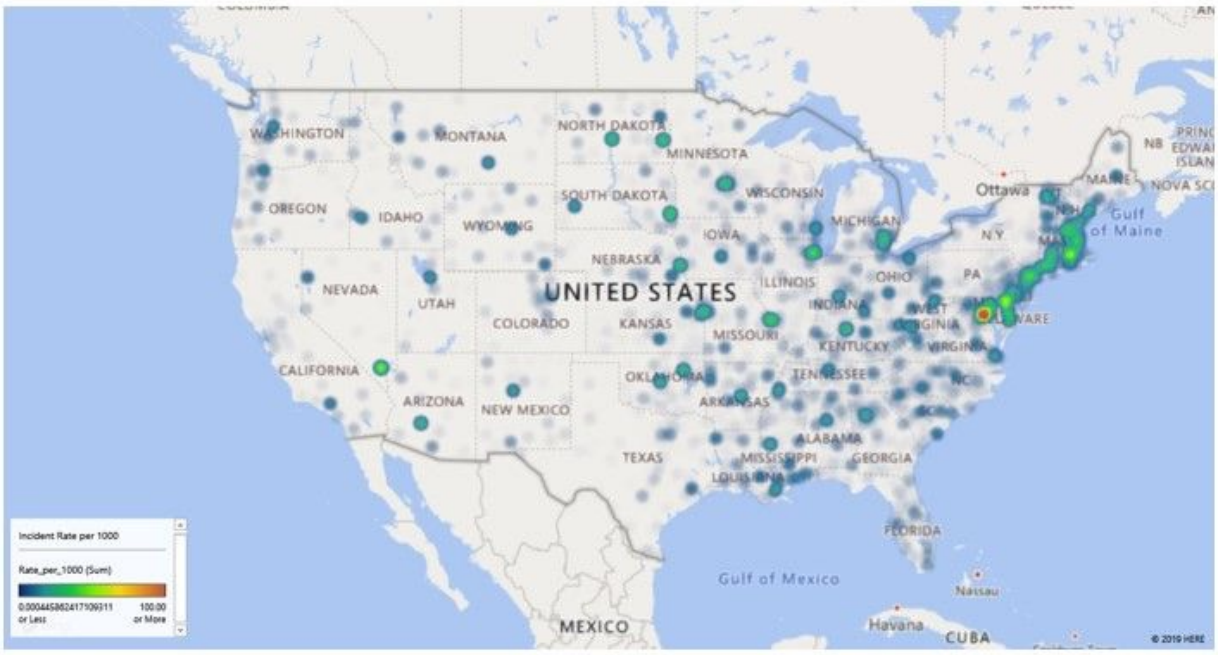

Figure 11

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	57.54	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

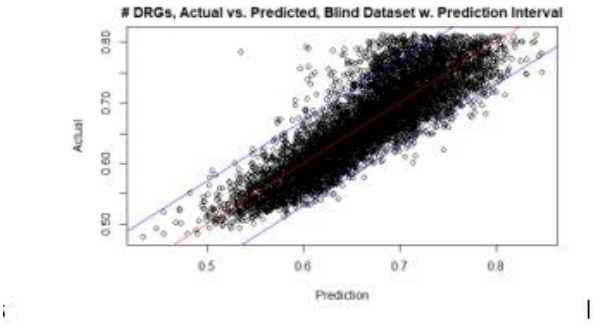

Figure 14

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