Analyzing the Effects of Non-Medical Factors[1] on Medical Care Outcomes

Himanshu Chauhan*, Kate McArdle*, Radhika Sundar†
* Dept. of Electrical and Computer Engineering
†Division of Statistics and Scientific Computation
The University of Texas at Austin
{himanshu, kate.mca, radhika.sundar}@utexas.edu

Abstract

In this paper, we analyze the effects of demographic factors on two medical care outcomes: length of stay and total charges. We use the Patient Discharge Data from the state of California and apply classification techniques to predict above or below length of stay, and charges. We also analyze patterns of heart failure occurrences across three major counties. Our results indicate that even though there are significant variations in medical outcomes across different demographies, predicting outcome values based only on demographic features is a hard problem.

I. Introduction

The area of healthcare analytics has gained vital social and economic importance in recent past. With advancements in technology, capturing, storing, and analyzing medical data has become easier and faster. In addition, applications of data mining and machine learning techniques have led to advancements in multiple areas of medical sciences. The scale of healthcare related data has already reached huge proportions, and with analysis of such data new insights being gained continuously. The increasing costs of providing healthcare and providing access to health insurance and medical treatment necessitate careful study of the available data so that targeted health policies can be formulated. It is important for policymakers and insurance companies to know if certain demographic sections are being poorly served by the healthcare system. The health care system needs to become more efficient and cost-effective, especially in terms of reducing the burden of cost on economically weak and indigent sections. For example, if there are geographic areas where access to health insurance is limited, or hospital infrastructure is poor, then these issues could be addressed by appropriate policy measures. Other situations that warrant study are regional imbalances in healthcare access and outcomes, inequalities in availability of medical care, and prevalence of disease conditions among different races in the population.

In order to take such remedial steps, one must first carefully study the empirical data on patients' healthcare outcomes and the impact of demographics on such outcomes. There is much insight to be gained on disease patterns among different demographic sections, and such findings could be invaluable to securing a healthy future for society. Our contributions to the existing body of healthcare analytics research are:

- We examine the relationships between demographics and total charges incurred by heart failure patients from the California population in the years 2009-2001.
- We assess if demographics can be predictors of a patient's length of stay in the hospital and total charges incurred for patients with various diseases in the same population.

The rest of the paper is organized as follows: Section II describes our data source in detail, and Section III provides an overview of the related work. Section IV formulates the two problems that we address and discusses the applied methods. Section V presents the insights generated by our analysis of the data and evaluates the performance of our predictive models. Section VI discusses some additional approaches that we explore to improve the performance of our predictive models, and Section VII gives the concluding remarks.

II. Data Source

We use the Patient Discharge Data Set from the California Office of Statewide Health Planning and Development [1] for the years $2009-2011$. This data set contains about 11.8 million records over these three years. Each record includes 128 features, including demographic and medical information. In order to protect patient confidentiality in the case of unique records, certain variables are masked by replacing those values with a $*$. Candidate variables for masking include some demographic variables and the hospital's identification number. The variables are masked in a pre-defined order as described in the documentation for the dataset. In Section IV-C we discuss our approach to handling records with masked features.

A. Predictor Features

The predictor features we are interested in represent a patient's demographics and we provide a summary of them here.

Age Range (20 categories): The age range of the patient, broken into 20 five-year increments. The patient's actual age is recorded in a separate continuous-valued variable in addition to this variable and the 5-category variable. Because the actual age variable is masked 48% of the time, we use the categorical versions of age.

Age Range (5 categories): The age range of the patient, broken into five 17-year categories.

Gender: One of five different levels: Invalid, Male, Female, Other, and Unknown. The 'Other' category includes transgendered patients and patients of undeterminable gender, while 'Unknown' indicates that the medical record does not list a gender.

Patient County: The geographic county in California where the patient resides, or a value denoting the patient resides outside of California.

Source of Admission: This variable indicates the place from where the patient arrived at the hospital (e.g. home, residential care facility, another hospital, etc) as well as if the patient was admitted by the Emergency Room.

Expected Source of Payment (Payer Category): The entity expected to pay the largest share of the patient's bill. This value is generally a form of insurance (e.g. Medicare, Medi-Cal, or Private Coverage), Self Pay, or another organization (such as a government program).

Do-Not-Resuscitate (DNR) Order (Pre-Hospital Care & Resuscitation): Variable recording if there is an order directing not to resuscitate the patient in the event of cardiac or pulmonary arrest.

Principal Diagnosis: The diagnosis established to be the main reason the patient was admitted to the hospital. We use this field to filter out the records by principal diagnosis. These diagnoses are coded according to the ICD-9-CM standards of coding medical diagnoses.

Other Diagnoses (up to 24 listed): Other medical conditions the patient already had, developed during the hospital stay, or were relevant to treatment.

Principal Procedure: The procedure performed as the definitive treatment for the primary diagnosis, coded according to ICD-9-CM.

Other Procedures (up to 20 listed): All other procedures performed on the patient during the visit of record, coded according to ICD-9-CM.

B. Outcome Variables

We are interested in two outcomes related to medical care that a patient receives, which we discuss here.

Length of Stay: Total number of days from the patient's admission date to discharge date.

Total Charges: The total charges billed for all services. Invalid and unknown charges are listed as $0, and charity care is listed as $1. Charges exceeding 7-digits are denoted by $$9,999,999$.

III. Related Work

A wide variety of data mining techniques has been used to analyze hosptial discharge data. Bindman et. al. [2] analyze California hospital discharge data to examine whether higher hospital admission rates for chronic medical conditions such as asthma, hypertension, congestive heart failure, etc. in low-income communities result from community differences in access to care, prevalence of the diseases, propensity to seek care, or physician admitting style. Geary et. al. [3] analyze medical discharge data in the British armed forces to study if military service has a stronger detrimental effect on females than males. Their findings indicate that that for all disease and injury categories of medical discharge there is a statistically significant excess in females.

Green et. al. [4] analyze the results of an audit on discharge abstract records of Medicare-aged patients hospitalized in California, and study if hospital discharge data are accurate for evaluating effectiveness of care. Jencks et. al. [5] use medical discharge data of New York and California to assess the meaning of hospital-associated death rates. Specifically, they study whether mortality within 30 days of hospital admission (30-day mortality) is more informative than inpatient mortality and whether detailed assessment of additional discharge diagnoses helps in understanding death rates. Demlo et. al. in [6] focus on improving the quality of hospital discharge data. They show that routine use of discharge summary and operative reports to abstract information on patient disposition, principal diagnosis, and principal procedures are influential approaches for improving the quality of discharge data. They also offer additional recommendations for developmental activities intended to improve the designation of diagnosis, diagnostic classification schemes and hospital medical records systems.

Many other research efforts have applied data mining techniques to study the problems related to heart failures. Vijiyirani et. al. [7] review some common classification techniques for heart disease prediction such as random forests and decision trees. Dangare and Apte in [8] employ neural networks to predict heart disease occurrence; they find gender to be significant in addition to medical factors. Other studies on heart-related medical conditions focus on medical factors. Soni et. al. in [9] perform a comparative study of classification methods for predicting heart disease. They present their findings that decision trees are more effective than K-Nearest Neighbor and clustering-based classification methods. Xing et. al. in

[10] use support vector machines (SVMs), neural networks, and decision trees to predict survival rates in coronary heart disease. They use cross-validation to train their models and report the comparative predictive power of the different models.

Data on many diseases besides heart-related conditions has been mined. A study by Banta et. al. [11] on sepsis in-hospital mortality uses staged logistic regression to predict mortality and find that patient demographics, health status, and hospital characteristics are important predictors. Sharma in [12] compares models based on tree boosting and multilayer perceptron neural networks in oral cancer prediction. This study finds that the single most important variable, according to both models, is the presence of a particular lymph node. Renal disease was studied by Yang et. al. in [13]. Here the authors develop models to predict survival likelihoods. The methodologies employed are a standard statistical approach using Cox proportional hazards regression and a data mining approach using SVMs. Their principal finding is that when the features are combined in a certain way, the SVM model outperforms the regression model. A data mining study on breast cancer diagnosis by Shukla et. al. in [14] shows that neural networks can be used for early detection of cancer. The authors use a feed-forward neural network trained using three different algorithms: back propagation, radial basis function networks, and learning vector quantization. A study by Liu et. al. [15] focuses on geriatric patients' lengths of stay in hospitals and uses Naïive Bayes classifiers and decision tree methods. The focus of their study is geriatric patients who require longer durations of hospital stay.

Diabetes is a focus of much previous work. Breault et. al. in [16] perform a data mining study on a diabetes data warehouse. They detail the challenges of dealing with such data and converting it into a file suitable for data mining purposes. They employ the classification tree approach in CART (Classification and Regression Tree). Their results show that belonging to a young age group is significant as a predictor for diabetes in the presence of certain other predictors. This unexpected finding demonstrates the novel associations that data mining-based analysis can produce. Al Jarullah in [17] discusses the use of decision trees in predicting Type II diabetes. Similarly, Zorman et. al. in [18] use decision trees in combination with association rules to examine a diabetes database. Based on their results, the authors explain that the sets of rules obtained by decision trees are smaller than those obtained using association rules.

Groseli [19] gives an overview of problems that arise in applying data mining to medical databases. The author also discusses protecting confidentiality of records. Finally, Cios and Moore in [20] discuss the unique aspects of medical data mining and knowledge discovery. They also present the ethical, legal and security-related issues that arise in handling medical data.

Most of these research efforts focus on analyzing patterns of diseases and drawing insights using medical factors, with demographic factors having an auxiliary role. However, we focus our analysis primarily on demographic factors as predictors of medical care outcomes.

IV. Methods

We study two problems on the patient discharge data. The first problem focuses on gaining insights from the data, especially in terms of differences for occurrences and treatment of a particular disease across different locations. The problem formulation is given by the following:

[Problem 1**] Insight Mining**: For the three counties - Los Angeles, San Francisco, and Santa Clara, explore patterns of disease occurrence, treatment expenses, and type of insurance.

For this problem, we restrict our attention to the primary diagnosis of heart failure. To perform a per capita comparison of disease occurrences across the three counties, we use county population information, available at [21]. In addition, to identify the largest hospitals, we use the number of discharges per year (provided in the dataset). The pre-processing and filtering steps performed for this study are described later in this section.

Our second goal is to use demographic information for predicting length of stay (LOS), and total charges (CHG).

[Problem 2**] Outcome Predictions**: *For a given record within a primary diagnosis category, predict the two outcomes - length of stay (LOS) and total charges (CHG) - using the patient's demographics.*

As discussed in Section II, a significant proportion of the features in the data is categorical, with no ordinal relationship. With most of the features being purely categorical, the problem of finding separation boundaries is a challenging problem. In addition, both our response variables, *LOS* and *CHG*, are continuous. Prediction of continuous values based on categorical features is a much harder problem.

To reduce the hardness of the problem, we transform the problem into a binary classification problem of predicting two classes, for both of the response variables. We define these classes in the same manner for each of the response variables:

- $Class0$: The response variable value is less than or equal to the expected value (mean).
- $Class1$: The response variable value is greater than the expected value (mean).

Hence, for response variable LOS, if the length of stay for a record is less than or equal to the mean length of stay it is assigned label $Class0$, otherwise it is assigned label $Class1$. Similarly, for response variable CHG of any record if the total charges paid are less than or equal to the mean value of charges then it is assigned label $Class0$, otherwise class label $Class1$.

For the scope of this paper, we restrict our analysis to two primary diagnoses: heart failure and cardiac arrhythmia. Our predictions for records with primary diagnosis of cardiac arrythmia are performed for a specific subset of the records for which procedure code 37.34 (ICD-9 code) was performed. This procedure is one of the most common procedures performed (in the dataset) for the cardiac arrythmia cases, and pertains to 'excision or destruction of other lesion or tissue of heart.' Restricting our focus to a specific procedure allows us to contrast the predication accuracy results for heart failure data - on which no procedure restriction was applied - with the same predictive models. This comparison can also be helpful in assessing the degree of information gain provided by the additional restriction.

We also perform predictive analysis on an extended set of primary diagnoses: such as diabetes, hypertension, lung cancer, and schizophrenia. Detailed results of studies on these diagnoses are presented in the Appendix of the extended report available at [22].

As discussed in the results section (Section V-B1), using only demographic information for outcome predictions with our models leads to somewhat poor accuracies for the binary classification problem. To improve our prediction accuracies, we use three additional features:

- Hospital Size: The dataset [1] provides total number of discharges for each hospital. For each record, we use the normalized value of the average number of dischages (over the years 2009-2011) for its hospital.
- Number of Other Diagnoses: Each record lists additional diagnoses (up to 24) present for any patient. It is natural to assume that the presence of additional medical conditions should be an important factor in the treatment of patient. We compute only the the total number of additional diagnoses, and use it as a feature for our models.
- Numer of Other Procedures: Irrespective of the primary procedure performed, there may be additional procedures performed on a patient during her stay. We use the number of such additional procedures as a feature.

A. Classification Methods

For the binary classification problem defined in Problem 2, we apply three different classification techniques. The primary intuition behind the application of multiple classification techniques is the assumption that the prediction results from different methods may not be the same. Due to the differences in the fundamental modeling principles of the methods, different methods may perform differently on same set of records. We apply the following three classification methods:

- **Logistic Regression**: This technique has been used extensively for the purpose of binary classification [23]. Based on our initial analysis the two classes are not distributed equally across the features. Also, we do not assume that the errors are normally distributed. With these observations, the underlying assumption of 'non-homoscedasticity' in the logistic regression technique suits our purpose for the binary classification.
- **Naïve Bayesian Classifier**: As most of the features in the dataset are categorical with no ordinal relationship between their values, this classification technique is used in a manner similar to document classification approach [24]. To avoid the case of probabilities of zero occurrences, we applied smoothing for missing features in the training set.
- **Generalized Boosting**: We apply generalized boosting methods on decision trees for binary classification. This technique is helpful in identifying the most important features and their relative importance in comparison to each other. In addition, the boosting [25] approach allows improvement on overall predictions even when accuracies of individual decision tree are somewhat low. The parametric settings used for this method are listed in Section V-A.
- **Ensemble**: We build four ensemble models to combine the predictions of the individual models previously described. Each of the four ensemble models runs the naïve Bayes, logistic regression, and GBM models exactly as before. The ensemble models differ only in the combiner step, making class predictions as follows:
  - Ensemble 1: Assign the class label that is the majority vote of the three individual models.
  - Ensemble 2: Assign label $Class1$ if any of the three models predict $Class1$; otherwise assign label $Class0$.
  - Ensemble 3: Average over each model's probabilistic estimate for label $Class1$; assign label $Class1$ if this average probability is greater than or equal to 0.5; otherwise assign label $Class0$.
  - Ensemble 4: Select the maximum probabilitic estimate for label $Class1$ if any model indicates that such probability is greater than or equal to 0.5, and select the minimum probabilitic estimate for label $Class0$ otherwise. Assign label $Class1$ if the selected probability is greater than or equal to 0.5; assign label $Class0$ otherwise.

Note that the second and fourth ensembles produce the same final class predictions. The second ensemble is convenient in case we later decide to add models that do not provide class probabilities, while the fourth ensemble allows us to assess the ensemble's Rate of Change (ROC) curve and lift charts.

B. Key Challenges with Data

We now discuss some challenges posed by the dataset.

Masked variables and missing data: Multiple demographic features were subject to masking in order to protect the identity of the patients. Hence, the patients could not be tracked across visits or readmissions. Not being able to track patient readmissions, rules out the possibility of addressing some interesting research challenges related to disease relapse rates. Also, there were a large number of missing data for some demography variables.

Large number of categorical predictors: Excluding a handful of features, such as length of stay, patient age and total charges, most of the other features in the dataset are categorical. Thus, the scope of dimensional transformations is quite limited and is mostly restricted to to binning strategies.

Skewness in outcome variables: The outcome variables that are of interest to us exhibit either a wide spread in their valuesor are too concetrated towards one end of their range. For example, total charges for the principal diagnosis heart failure, range from less than $\$100$ to more than $8,000,000$, even when charities or invalid charges are excluded. Also, the a significant proportion of records have length of stay values less than 10 days, where as the maximum value for this variable is more than one year.

C. Pre-Processing & Filtering of Records

With our restricted focus on two diseases - heart failure and cardiac arrythmia - we filter the records based on the primary disagnosis code (identified from [26]). For heart failure, the ICD-9 code is 428, and for cardiac arrythmia it is 427. Once all records for the particular primary diagnosis code have been selected, we apply a feature-specific filtering to only retain the features of our interest. This pruning is helpful for retaining records for which other features, not in our desired feature set, might be missing. For the insight mining problem, we select the feature of interest one at a time and then discard missing values.

For the purpose of predictive analysis, we remove all records that have missing values for any of the features of our interest. Given that the size of the dataset is large, we expect that complete removal of such records does not result in significant loss of information. For predicting class labels for total charges (CHG), we remove records that have length of stay greater that 365 days, as such records are very few, and are not representative of the larger set of records. We also discard records that have total charges of 0 or 1 dollars. The reason for this pruning is that a charge of 0, indicates an unknown or invalid reported value; where as charge of 1 is reported for cases of charity care. We then build three models separately on this filtered data set.

V. Evaluation

We now present the experimental evaluation of our approach.

A. Experimental Setup

To assess the approach described in the previous section, we use the R programming language and its built-in packages to train and test Naïve Bayes, Logistic Regression, GBM, and Ensemble models. We train and test each model separately on two subsets of the dataset, one for heart failure patients and one for cardiac arrhythmia patients who underwent procedure 37.34. After preprocessing as described in Section IV-C, the heart failure dataset has $232,596$ records and the cardiac arrhythmia dataset has $11,738$ records. We use random $70-30$ train/test splits. To ensure repeatability across experiments, we set a specific seed before drawing each split.

We implement naïve Bayes models using the R package `e1071`, assigning a Laplace smoothing value of one. To implement logistic regression, we use the `cv.glmnet` function found in the R package `glmnet`. In our implementation, we use 10-fold cross validation for logistic regression with a 'lasso' penalty. To predict class labels, we use the lasso penalty that corresponds to the lowest mean cross-validated error. The lasso penalty values used for predicting length of stay (*LOS*) and total charges (*CHG*) are presented in Table I for the two diagnoses of our focus. For the input features that are categorical (age, race, source of admission, payment category, and do-not-resuscitate order), we apply `model.matrix` on these features in the train and test datatsets before calling `cv.glmnet`. This function converts the specified features into dummy variables, with one variable for each possible value of each categorical feature, and each record is populated with a '1' in the columns corresponding to the feature values present in that record and a '0' otherwise. To implement GBM, we use the API provided by the `gbm` package. To train the GBM model, we use a maximum of 200 trees, 0.05 shrinkage rate, interaction depth of two, bag fraction of 0.5, train fraction of 0.7, minimal total weight of 50 for each node, and 3-fold cross-validation. For

Diagnosis	$\lambda_{LOS} \times 10^{-4}$	$\lambda_{CHG} \times 10^{-4}$
Heart Failure	0.8	0.7
Cardiac Arrhythmia	3.2	1.0

TABLE I: Lasso Penalty Values for Logistic Regression

class label predictions, we restrict the number of trees to the value returned by the 'best-iteration' from cross-validation. Finally, we build four ensemble models as described in Section IV-A to combine the predictions of the individual models.

We next use these models for the binary classification problem of predicting *LOS* and, separately, *CHG*, for the records in our test datasets.

B. Analysis & Results

We first present the insights gained from the analysis of the heart failure records in the dataset. The evaluation of our binary classification approach using the predictive modeling is presented later.

1) Insights from Heart Failure Data: For the three counties - Los Angeles, San Francisco, and Santa Clara - the total number of heart failures, and per capita rate of heart failures are reported in Table II. Observe that per capita rate of heart failures in Santa Clara county is significantly lower than that in Los Angeles, and San Francisco counties.

	Los Angeles	San Francisco	Santa Clara
Total	73,392	5,823	8,833
Per Capita	0.0074	0.0070	0.0048

TABLE II: Heart Failure across LA, San Francisco, and Santa Clara

Table III presents a comparison of mean values of charges for heart failure records across all hospitals, against those for the three largest hospitals in California. For the computation of these mean values we ignore the records where the total charges are one million dollars or more. The presented values suggest that there are considerable differences in the expected values of payment across the hospitals. Possible reasons for these differences would be the locations of hospitals (rural versus urban), whether or not they are specialized (for intensive care or special surgeries), and overall hospital size.

Fig. 1 presents the comparison of mean charges over all the heart failure records across various types of insurance categories. As seen in the plot, the mean charges paid for private insurance, and self pay

Hospital	Mean value of Charges ($)
Overall mean	62,329
Cedars Sinai - LA	151,201
Centinela Hospital - LA	56,682
St Agnes - Bay Area	41,259

TABLE III: Mean Heart Failure Charges across Hospitals

categories are significantly higher in Santa Clara county, in comparison to the rest of the two counties. Also, the mean charges paid for workers-compensation category in Los Angeles county are significantly higher than those in the other two counties.

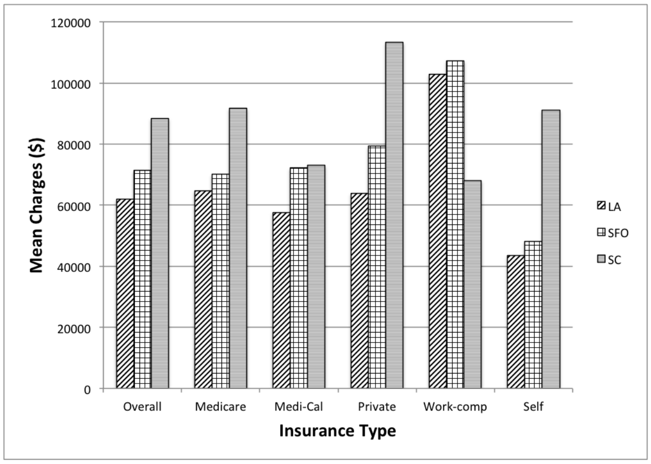


Fig. 1: Mean Charges for Heart Failure across Insurance Categories

We now present the results of the binary classification problem for the two data subsets of focus: one corresponding to heart failure patients, and the other corresponding to cardiac arrhythmia patients who underwent principal procedure with ICD-9 code 37.34. Results for experiments run on additional diagnoses can be found in the Appendix of the extended report [22].

We use three measures to evaluate our results on the test datasets: accuracy of predictions, rate of change (ROC) curve and area under the ROC (AUROC), and lift curve. For each dataset, we calculate the proportion of the majority class in the test set and use this prior probability as a baseline accuracy. For example, of the 3521 records in the cardiac arrhythmia test set, 2688 records have label $Class0$, so the baseline accuracy is 2688/3521, or 76.3%.

Table IV displays the accuracy and AUROC values obtained with the baseline and with our classification

Model	**% Accuracy**	**AUROC**
Baseline	73.4	-
Demographics only	73.0	0.65
Naïve Bayes	79.5	0.79
Logistic Regression	79.7	0.79
Gradient Boosting	81.2	0.83
Ensemble	80.6	0.81

TABLE IV: Accuracy and AUROC for Heart Failure Charge Predictions

models, for the heart failure dataset. Our initial goal was to use only demographics to predict *CHG* and *LOS*, and for this purpose our input features are age, gender, race, patient's geographic county, source of admission, and payment category. However, our models with only these input features have similar prediction accuracies as the baseline; one such set of results is displayed in the second entry of Table IV. Given this poor performance, we augment the input feature set with approximations for medical information (described in more detail in Section IV): number of additional diagnoses present, number of procedures performed, and hospital size. To maintain relatively low model complexity, we also remove from the feature set those features which have low relative importance as given by the GBM analysis: gender, race, and patient county. In the rest of this section, all results presented are based on experiments using this augmented input feature set. We discuss approaches to further improve the input feature set in Section VI.

In Table IV, it is evident that the GBM model outperforms all other models, in both accuracy and AUROC. However, the accuracy obtained with GBM is 81.2%, which is only 7.8 percentage points higher than the baseline accuracy. The naïve Bayes and logistic regression models provide similar results, with accuracies around six percentage points higher than the baseline accuracy. The ensemble results presented represent that of the best-performing ensemble for this setting, which is Ensemble 2 as described previously. This ensemble's accuracy and AUROC (and thus those of all each ensemble) are lower than those obtained by the GBM-based model. While the motivation for creating an ensemble is to improve the performance over all individual models, the fact that none of our ensembles perform better than GBM can likely be explained by a lack of diversity among the individual models.

Fig. 2 shows the ROC and lift curves for our models' *CHG* predictions for heart failure patients. These curves demonstrate the similarity of the models: although the curves for GBM are slightly better than those for the other models, the difference in the curves is insignificant. In particular, assessing the lift

Model	% Accuracy	AUROC
Baseline	63.8	-
Naïve Bayes	69.7	0.71
Logistic Regression	70.2	0.72
Gradient Boosting	71.1	0.73
Ensemble	70.5	0.73

TABLE V: Accuracy and AUROC for Heart Failure Length of Stay Predictions

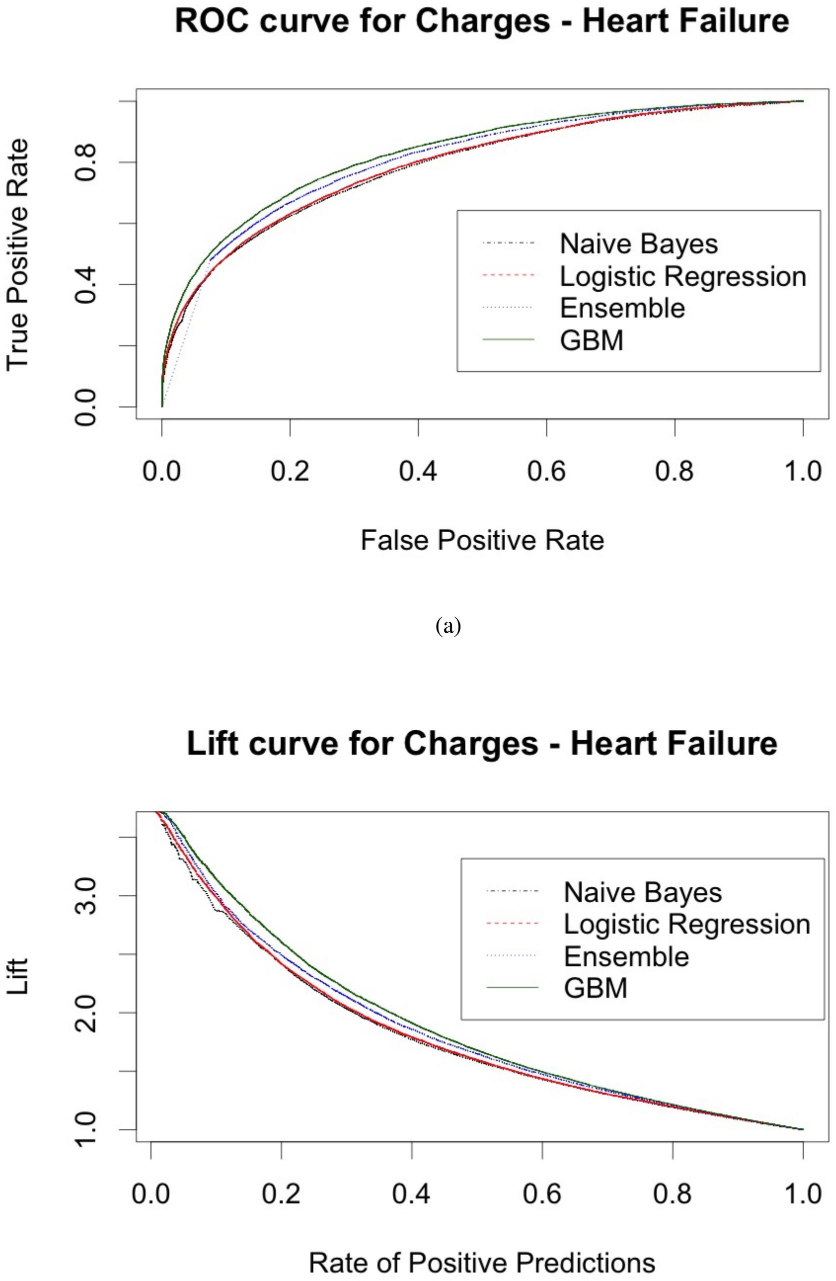
curve, one can see that with the top 10% of positive predictions, the lift ranges only from about 3.1 with naïve Bayes to about 3.5 with GBM.

Table V presents the models' accuracy and AUROC values for predicting *LOS* for heart failure patients, and Fig. 3 displays the ROC and lift curves generated by the same models. The improvements in accuracy over the baseline are in similar proportion as for those of predicting *CHG*, while the AUROC is lower in comparison. Based on these results, it can be said that these models are less successful at predicting *LOS* than they are for predicting *CHG* for heart failure patients.

Accuracy and AUROC values for predicting *CHG* and *LOS* for cardiac arrhythmia patients who underwent procedure 37.34 are shown in Table VI and Table VII, respectively. Fig. 4 and Fig. 5 display the ROC and lift curves generated by these models for *CHG* and *LOS*, respectively. For these cardiac arrhythmia patients, our models achieve the highest accuracy and AUROC for predicting *LOS*. This finding is in contrast to our results for heart failure patients, in which we were more successful at predicting *CHG*. It is important to note that the largest gain in accuracy is found in predicting *CHG* for cardiac arrhythmia patients, using GBM. This model provides a gain of 20 percentage points, while no other model (for either diagnosis or response variable) achieves an accuracy gain of greater than 10 percentage points. Additionally, the AUROC obtained by using GBM to predict *CHG* for cardiac arrhythmia patients is high, at 0.87. The significant improvement provided by GBM over the other models is apparent in Fig. 4. This relative improvement differs from our other prediction results, in which all of the models perform similarly.

VI. Discussion

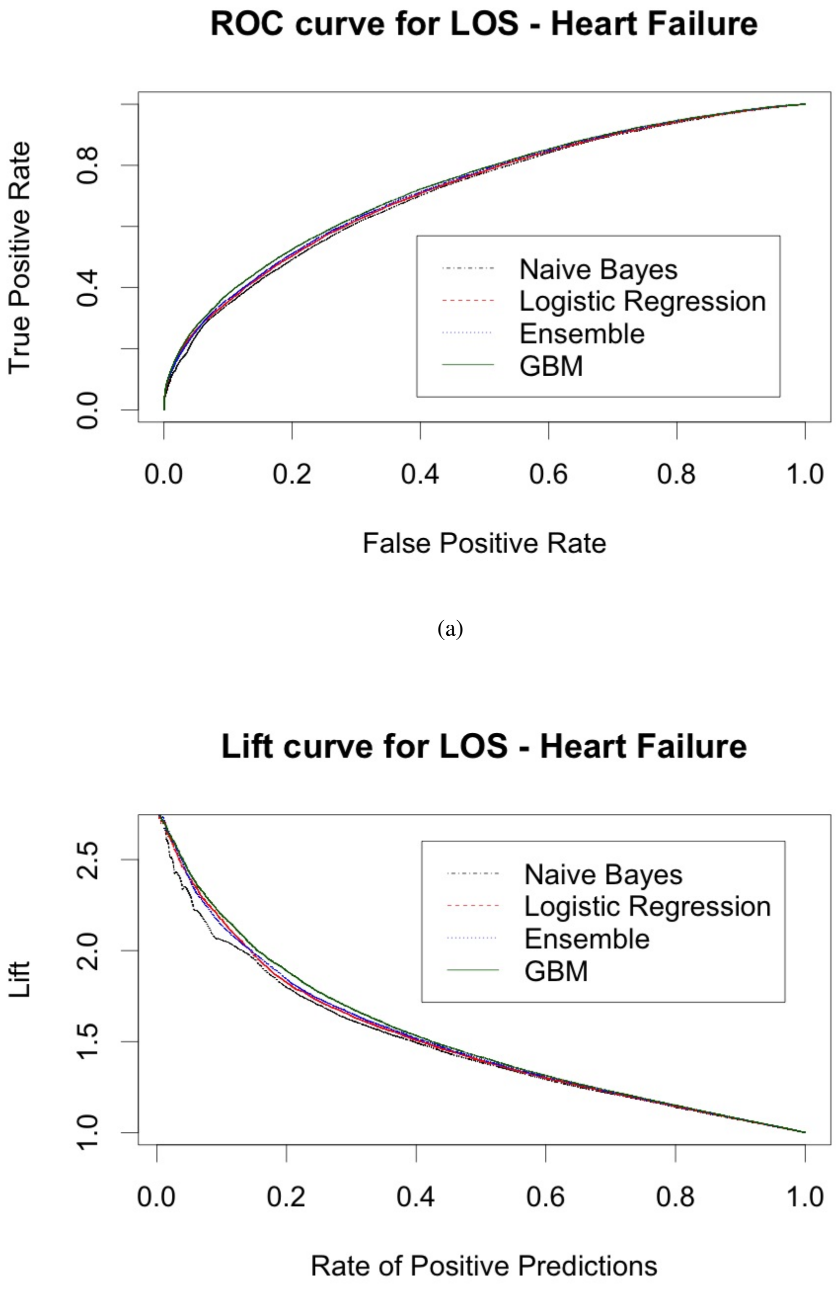
Our results for the binary classification of *CHG* and *LOS* for heart failure and cardiac arrhythmia patients indicate that there is room for improvement. In particular, the low accuracies suggest that the problem we are trying to solve - predicting medical care outcomes based on demographic information -



(a)

(b)

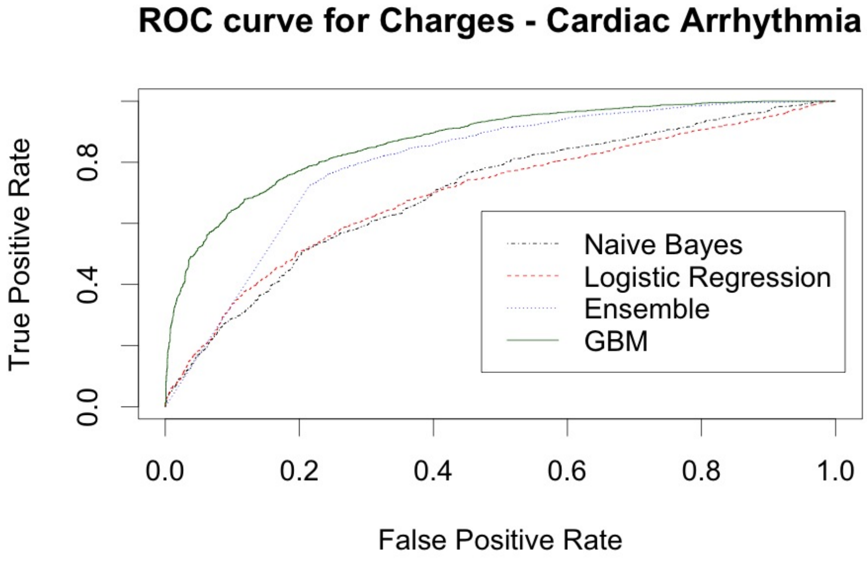
Fig. 2: ROC and Lift Curves for Heart Failure Charge Predictions



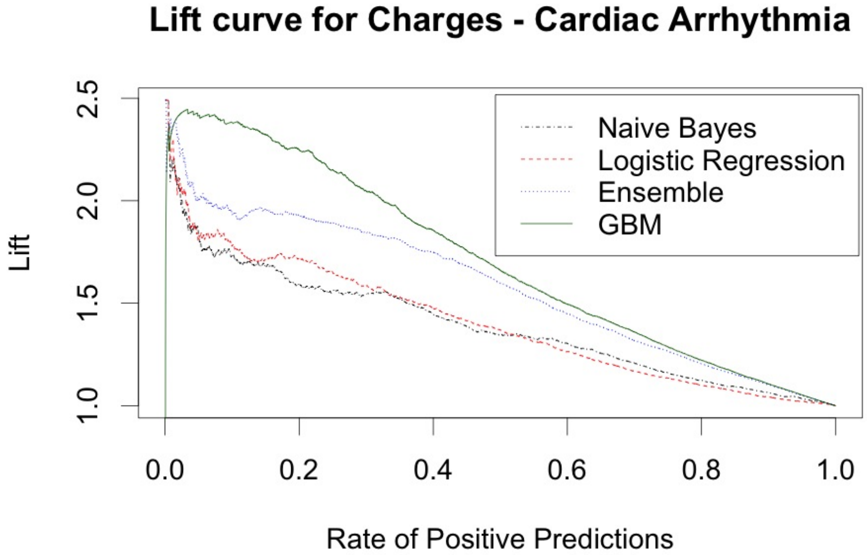
(a)

(b)

Fig. 3: ROC and Lift Curves for Heart Failure Length of Stay Predictions

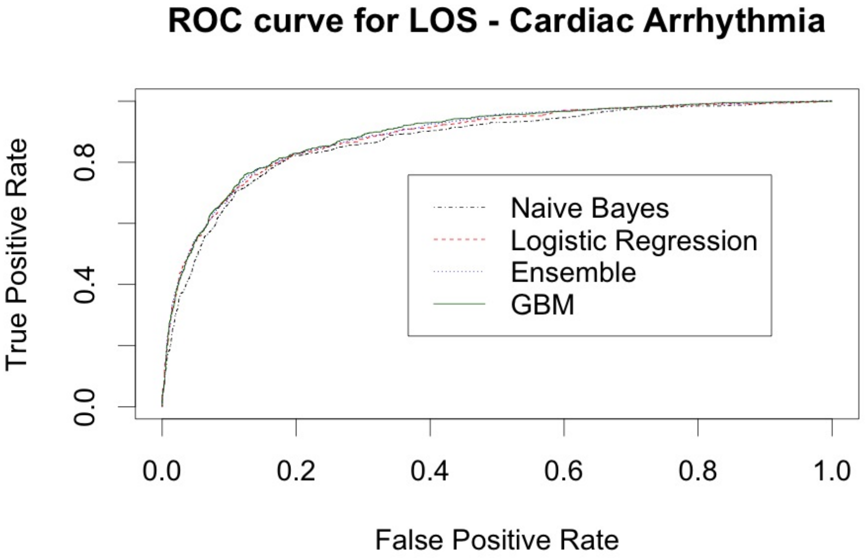


(a)

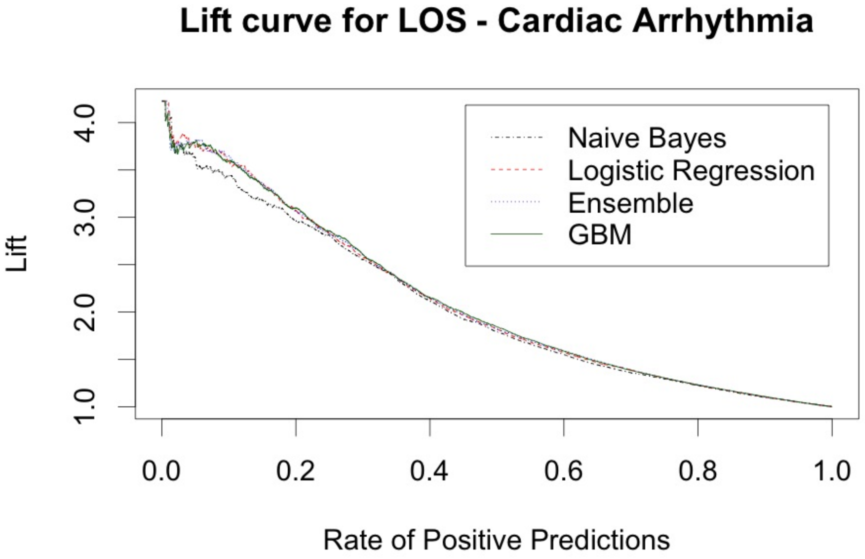


(b)

Fig. 4: ROC and Lift Curves for Cardiac Arrhythmia Charge Predictions



(a)



(b)

Fig. 5: ROC and Lift Curves for Cardiac Arrhythmia Length of Stay Predictions

Model	**% Accuracy**	**AUROC**
Baseline	59.8	-
Naïve Bayes	65.5	0.70
Logistic Regression	67.9	0.70
Gradient Boosting	79.5	0.87
Ensemble	76.0	0.80

TABLE VI: Accuracy and AUROC for Cardiac Arrhythmia Charge Predictions

Model	**% Accuracy**	**AUROC**
Baseline	76.3	-
Naïve Bayes	84.6	0.87
Logistic Regression	85.3	0.88
Gradient Boosting	85.7	0.89
Ensemble	85.3	0.89

TABLE VII: Accuracy and AUROC for Cardiac Arrhythmia Length of Stay Predictions

may not be separable in the current input feature space. Transforming the input feature space is a very challenging task because the demographic features are all categorical. In this section we briefly discuss the impact that different input feature spaces have on our model's performance, focusing specifically on using GBM to predict *LOS* for cardiac arrhythmia patients who underwent procedure 37.34.

A. Combined Age and Race

In the results presented in Section V-B1, we do not include race in the input feature space, because it was among the lowest relative importance of features as indicated by GBM. Additionally, we suspect that including income level as an input feature could improve our performance, but we do not have this information available in our dataset. As a proxy for income level, we create a new input feature that simply combines the age category and race features into one categorical feature. This combination is motivated by the observation that while patients of the same age or race category alone may have an income distribution with large variance, the combined categories may result in income distributions with lower variances. However, including this combined feature in our models does not result in a performance improvement.

	Predicted 0	**Predicted** 1
Actual 0	2497	191
Actual 1	312	521

TABLE VIII: Confusion Matrix for Cardiac Arrhythmia Patients

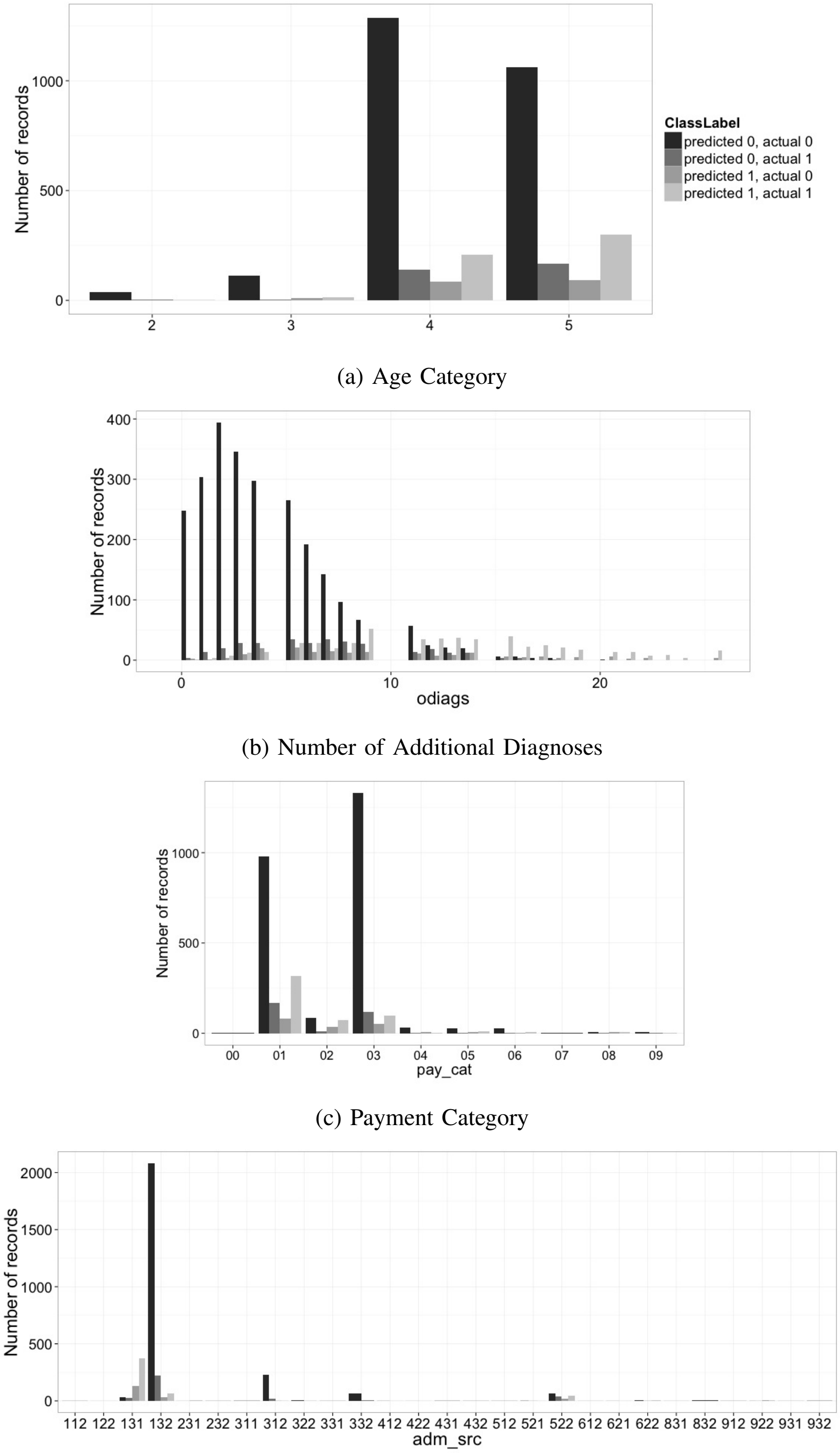
B. Pruned Data

We observe that for all of the models, we can predict records whose true label is $Class0$ (less than or equal to the mean) with high accuracy. The models are not very accurate in predicting records whose true label is $Class1$ (greater than the mean). This disproportion is evident in the confusion matrices produced by our models; one such confusion matrix for predicting *LOS* for cardiac arrhythmia patients using GBM is representative and is provided in Table VIII. Based on this observation, we investigate the distribution of our predictions across different input features. These distributions are shown in Fig. 6, for age category, number of additional diagnoses, payment category, and source of admission.

Based on these distributions, we further refine our dataset by forming the following individual subsets: records for which age category = 5, which corresponds to 65 years or greater; records for which number of additional diagnoses ≤ 10; records for which payment category = 1, which corresponds to Medicare; and records for which source of admission = 131, which corresponds to patients who arrived at the hospital from their homes and upon arrival were admitted through the hospital's Emergency Room. We select these subsets based on the fact that they represent a significantly large proportion of the given dataset and contain a large number of records whose actual label is $Class1$ but are misclassified as $Class0$ by our models. In other words, there is room for improvement in predicting $Class1$ for each of these subsets. However, our accuracies for these subsets do not improve from our overall accuracy as previously reported.

C. One Feature Considered at a Time

We assess the naïve Bayes model's performance using only a single input feature at a time. For nearly all input features, as expected, the performance is similar to the baseline. However, when admission source is used as the only input feature, we achieve an accuracy that matches that of any of our models, about 85%. Thus, it appears that source of admission may be the best predictor of length of stay for cardiac arrhythmia patients, while other demographic factors such as age, gender, race, and type of insurance play a lesser role in classification.



(a) Age Category

(b) Number of Additional Diagnoses

(c) Payment Category

(d) Source of Admission

Fig. 6: Distribution of *LOS* Predictions by Input Feature, for Cardiac Arrhythmia

Based on this understanding, we retry the classification of *CHG* for cardiac arrhythmia patients. We use GBM to assess the performance of isolated input features instead of naïve Bayes. GBM is selected because, in the larger input feature space, GBM provides an accuracy improvement of 14 percentage points over naïve Bayes in predicting CHG. In this case, the accuracy of 79% achieved using only the number of additional procedures and hospital size as input features matches the accuracy achieved when we use the previously used larger feature set with GBM. Once again, it appears that demographic factors are less valuable predictors than the number of additional procedures and hospital size for the charges a patient will incur.

VII. Conclusion

We analyzed the Patient Discharge Data from the state of California to mine for insights on the patterns of occurrences of heart failure, as well as differences in overall payments made across various categories of insurance. Our results from this analysis indicate that there is significant variation in mean payment values in different counties of California.

We applied predictive modeling, in the form of binary classification, to predict whether a patient will incur charges that are lower or higher than the mean charge values for a specific disease. A similar approach was applied for length of stay predictions. Our results for this problem indicate that making accurate predictions for these medical outcomes, without making use of detailed medical information, is a significantly hard problem.

References

[1] "California Patient Discharge Data." [Online]. Available: http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/

[2] A. B. Bindman, K. Grumbach, D. Osmond, M. Komaromy, K. Vranizan, N. Lurie, J. Billings, and A. Stewart, "Preventable hospitalizations and access to health care," *JAMA: the journal of the American Medical Association*, vol. 274, no. 4, pp. 305–311, 1995.

[3] K. Geary, D. Irvine, and A. Croft, "Does military service damage females? an analysis of medical discharge data in the british armed forces," *Occupational Medicine*, vol. 52, no. 2, pp. 85–90, 2002.

[4] J. Green and N. Wintfeld, "How accurate are hospital discharge data for evaluating effectiveness of care?" *Medical care*, vol. 31, no. 8, pp. 719–731, 1993.

[5] S. F. Jencks, D. K. Williams, and T. L. Kay, "Assessing hospital-associated deaths from discharge data," *JAMA: the journal of the American Medical Association*, vol. 260, no. 15, pp. 2240–2246, 1988.

[6] L. K. Demlo and P. M. Campbell, "Improving hospital discharge data: lessons from the national hospital discharge survey," *Medical Care*, vol. 19, no. 10, pp. 1030–1040, 1981.

[7] S. Vijiyarani and S. Sudha, “An efficient classification tree technique for heart disease prediction,” *IJCA Proceedings on International Conference on Research Trends in Computer Technologies 2013*, vol. ICRTCT, no. 3, pp. 6–9, February 2013, published by Foundation of Computer Science, New York, USA.

[8] C. S. Dangare and S. S. Apte, “Improved study of heart disease prediction system using data mining classification techniques,” *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44–48, June 2012, published by Foundation of Computer Science, New York, USA.

[9] J. Soni, U. Ansari, D. Sharma, and S. Soni, “Predictive data mining for medical diagnosis: An overview of heart disease prediction,” *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, March 2011, published by Foundation of Computer Science.

[10] Y. Xing, J. Wang, Z. G. Zhao, and Y. Hong, “Combination data mining methods with new medical data to predicting outcome of coronary heart disease,” in *Proceedings of the 2007 International Conference on Convergence Information Technology*, ser. ICCIT ’07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 868–872.

[11] J. E. Banta, K. P. Joshi, and L. B. H. B. Nguyen, “Patient and hospital characteristics associated with inpatient severe sepsis mortality in california, 2005-2010,” *Critical care medicine*, vol. 14, no. 11, pp. 2960–2966, 2012.

[12] N. Sharma, “Comparing the performance of data mining techniques for oral cancer prediction,” in *Proceedings of the 2011 International Conference on Communication, Computing and Security*, ser. ICCCS ’11. New York, NY, USA: ACM, 2011, pp. 433–438.

[13] C. Yang, N. W. Street, D.-F. Lu, and L. Lanning, “A data mining approach to mpgn type ii renal survival analysis,” in *Proceedings of the 1st ACM International Health Informatics Symposium*, ser. IHI ’10. New York, NY, USA: ACM, 2010, pp. 454–458.

[14] A. Shukla, R. Tiwari, and P. Kaur, “Knowledge based approach for diagnosis of breast cancer,” in *Advance Computing Conference, 2009. IACC 2009. IEEE International*, 2009, pp. 6–12.

[15] P. Liu, L. Lei, J. Yin, W. Zhang, W. Naijun, and S. Belciug, “Healthcare data mining: Prediction inpatient length of stay,” in *Intelligent Systems, 2006 3rd International IEEE Conference on*, 2006, pp. 832–837.

[16] J. L. Breault, C. R. Goodall, and P. J. Fos, “Data mining a diabetic data warehouse,” *Artificial Intelligence in Medicine*, vol. 26, pp. 37–54, 2002.

[17] A. Al Jarullah, “Decision tree discovery for the diagnosis of type ii diabetes,” in *Innovations in Information Technology (IIT), 2011 International Conference on*, 2011, pp. 303–307.

[18] M. Zorman, G. Masuda, P. Kokol, R. Yamamoto, and B. Stiglic, “Mining diabetes database with decision trees and association rules,” in *Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems, 2002. (CBMS 2002).*, 2002, pp. 134–139.

[19] C. Groselj, “Data mining problems in medicine,” in *Computer-Based Medical Systems, 2002. (CBMS 2002). Proceedings of the 15th IEEE Symposium on*, 2002, pp. 377–380.

[20] K. J. Cios and G. W. Moore, “Uniqueness of medical data mining,” *Artificial Intelligence in Medicine*, vol. 26, pp. 1 – 24, 2002.

[21] “Population Estimates - County Totals, U.S. Census Bureau.” [Online]. Available: http://www.census.gov/popest/data/counties/totals/2012/index.html

[22] H. Chauhan, K. McArdle, and R. Sundar, "Analyzing the Effects of Non-Medical Factors on Medical Care Outcomes - Extended Report." [Online]. Available: http://pdsl.ece.utexas.edu/himanshu/papers/dm13Report.pdf

[23] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, p. 841, 2002.

[24] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, Mar. 2002.

[25] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[26] "Online ICD9/ICD9CM Codes." [Online]. Available: http://www.census.gov/popest/data/counties/totals/2012/index.html

VIII. Appdenix

This appendix presents prediction results for length of stay (LOS) and charges (CHG) for several additional diagnoses.

Model	**% Accuracy**	**AUROC**
Baseline	71.6	-
Naïve Bayes	80.9	0.82
Logistic Regression	81.2	0.84
Gradient Boosting	82.8	0.86

TABLE IX: Accuracy and AUROC for Diabetes Charge Predictions

Model	**% Accuracy**	**AUROC**
Baseline	67.9	-
Naïve Bayes	77.0	0.79
Logistic Regression	77.9	0.81
Gradient Boosting	78.7	0.82

TABLE X: Accuracy and AUROC for Diabetes Length of Stay Predictions

Model	**% Accuracy**	**AUROC**
Baseline	73.1	-
Naïve Bayes	80.5	0.82
Logistic Regression	81.0	0.84
Gradient Boosting	83.2	0.87

TABLE XI: Accuracy and AUROC for Hypertension Charge Predictions

Model	**% Accuracy**	**AUROC**
Baseline	77.1	-
Naïve Bayes	80.7	0.80
Logistic Regression	81.4	0.81
Gradient Boosting	82.0	0.82

TABLE XII: Accuracy and AUROC for Hypertension Length of Stay Predictions

Model	% Accuracy	AUROC
Baseline	65.9	-
Naïve Bayes	75.1	0.80
Logistic Regression	75.9	0.82
Gradient Boosting	79.2	0.86

TABLE XIII: Accuracy and AUROC for Lung Cancer Charge Predictions

Model	% Accuracy	AUROC
Baseline	70.1	-
Naïve Bayes	76.5	0.75
Logistic Regression	77.5	0.77
Gradient Boosting	78.2	0.79

TABLE XIV: Accuracy and AUROC for Lung Cancer Length of Stay Predictions

Model	% Accuracy	AUROC
Baseline	71.2	-
Naïve Bayes	71.2	0.64
Logistic Regression	72.9	0.67
Gradient Boosting	75.8	0.78

TABLE XV: Accuracy and AUROC for Schizophrenia Charge Predictions

Model	% Accuracy	AUROC
Baseline	70.7	-
Naïve Bayes	69.0	0.61
Logistic Regression	71.3	0.63
Gradient Boosting	72.1	0.70

TABLE XVI: Accuracy and AUROC for Schizophrenia Length of Stay Predictions