Introduction

The Problem Statement

I am working on this problem to help identify which patients are at high risk of dying in the hospital to help physicians adjust their treatment plans to the wishes of the patient and family. The criteria for success are to develop a model that has high recall and precision at identifying which patients will die during the current admission. The scope of the solution will be to focus on identifying patients on admission who are at high risk for dying. Constraints on the scope will involve being able to get enough data on expired and non-expired patients. Getting data from the general population and not just diabetic patients. Key stakeholders will involve the hospitalist physicians in the Mid Atlantic Regional Health Center. The algorithm will be integrated and used in the EMR system. The key data source will be the UCI Machine Learning Repository. We will use the discharge disposition ID to predict which pts died. The modeling response will be 1 for expired patients and 0 for non-expired patients. The models used in this project will be Logistic Regression and Random Forest. The deliverables of the project will involve the Jupyter Notebooks for Data Wrangling, EDA, Feature Engineering and Modeling. Also, a presentation slide deck, report, and metric report.

Background

The ability to predict death accurately is crucial for adjusting goals of care to the patients; for making sound medical decisions for management, treatment, and prevention; and for having realistic expectations. Evidence suggests that physicians perform poorly in predicting when patients will die.

Allocating proper resources for high-risk patients will decrease the number of cardiac arrests and rapid responses which are a considerate amount of stress for hospital staff.

Goals

This project aims to provide physicians with an identification of which patients are high risk for mortality. This will allow for the physician to allocate the proper level of care for high-risk patients and start a discussion with patients and families concerning goals of care.

Datasets

The Dataset was downloaded from UCI Machine Learning Repository. The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient healthcare data.
The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.
Please see the Appendix for a list of features, descriptions, type and percent missing data.

Data set included two csv files. One containing records of the inpatient data. The other file contained the lookup tables for description id, admission id and admission source id. The features used in the final data model are the following:

Independent Features:

age
admission_type_id
admission_source_id
time_in_hospital
num_lab_procedures
num_procedures
num_medications
number_outpatient
number_emergency
number_inpatient
diag_1
diag_2
diag_3
number_diagnoses
change
diabetesMed

Target feature:
discharge_disposition_id

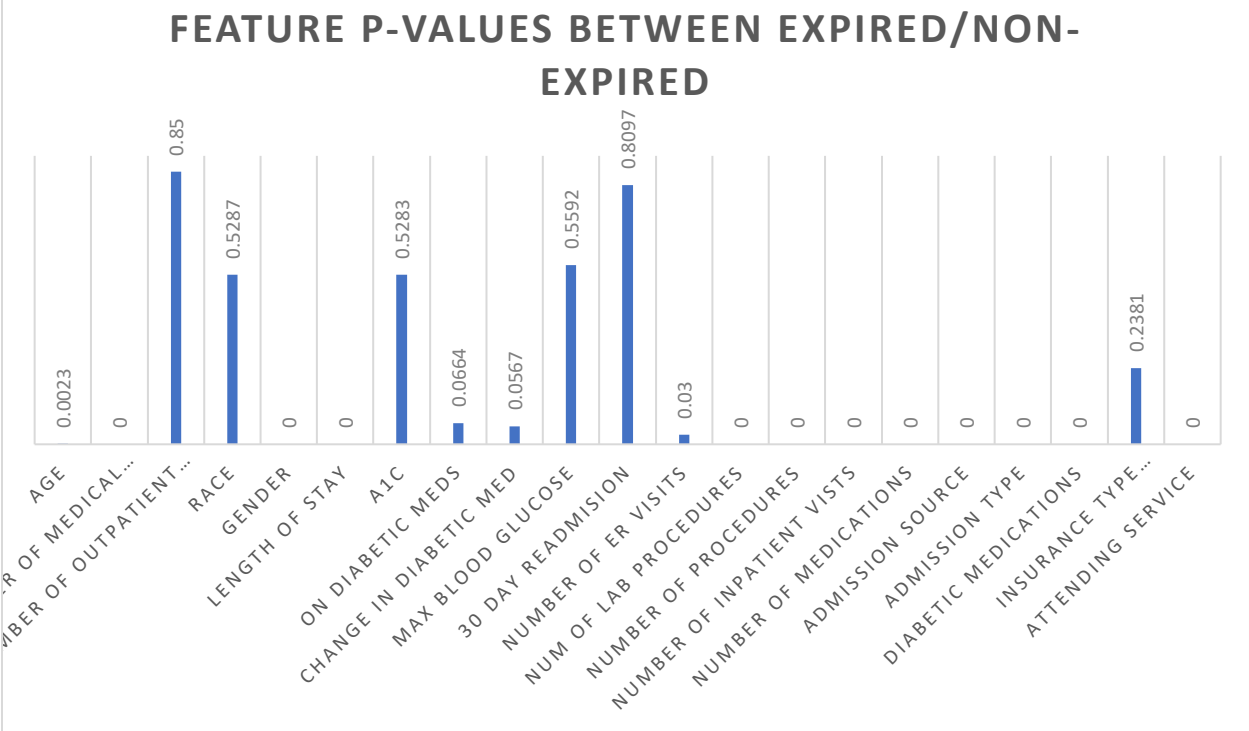
Data Wrangling

First, I loaded two csv files. One was the diabetic data set. The second one was the id_mapping file. Then I had to merge the id descriptions with the actually id in the diabetes data set for the three id features in the dataset. This involved changing some of the mapping id types to create one ID subtype that was used for all unknown values. Then merging all these unknown subtypes into one value for all unknown values. There were 3 Look up tables. (Discharge Disposition, Admission Source, Admission Type) included in the id_mapping.csv file.

Then I went through all unknown value types in the data set and changed them to one universal unknown datatype in the dataset. Then checked if there were any null values and duplicate rows. I had a total of 101,768 patient records. From this total, I was able to identify 2,422 expired patient records and 99,346 non expired patient records. I used SMOTE algorithm to fix the unbalanced data set (99,346 non-expired vs. 2,422 expired). I then checked specific column/feature values to see if they were unknown or missing. All rows of data were unique.

Exploratory Data Analysis

For this part of the project, I went through feature by feature and did t-test if it was a numerical comparison or a chi squared if it was a categorical comparison. I was looking for statistical significance between the two populations. I used a p-value of <.05 to determine if there was statistical significance.



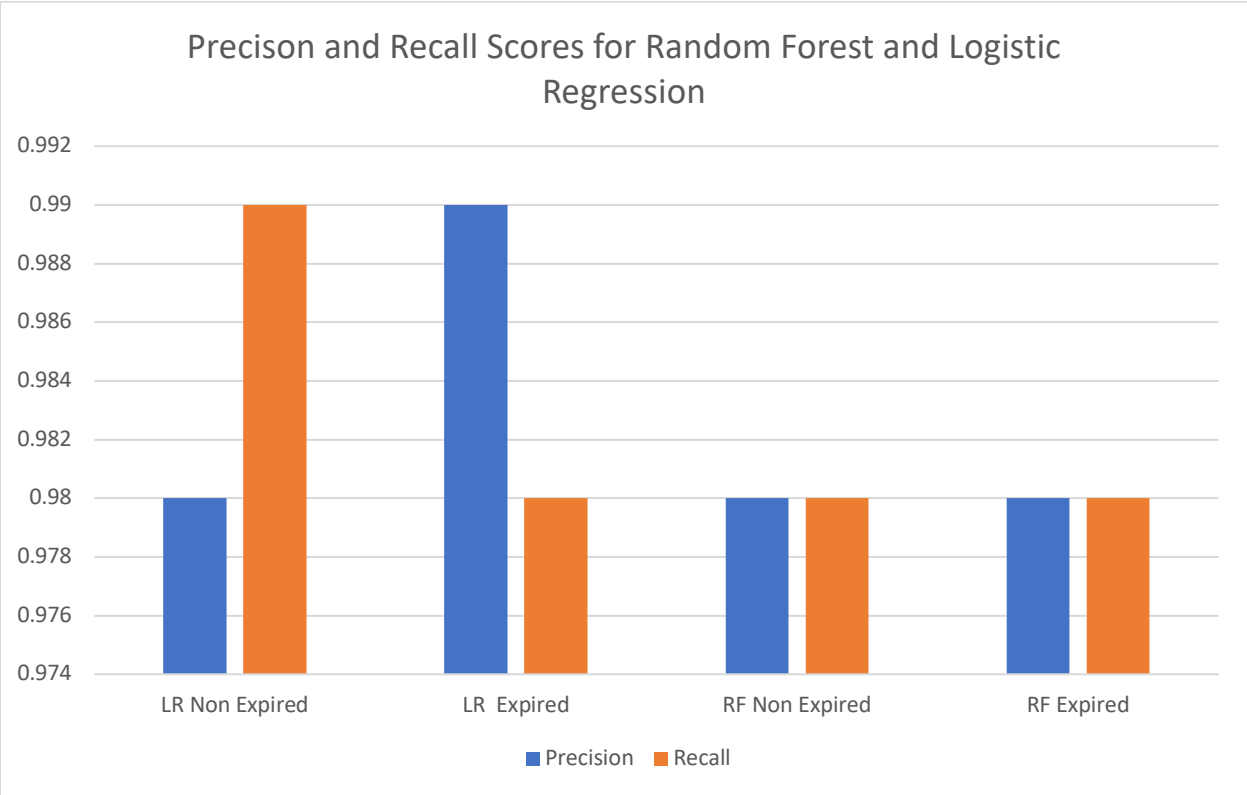
Feature Engineering

In this part I tried to encode all the values for the diabetic medication which were [No,Up,Down,Steady] to either a 1 or 0 if they were on the medication or not. I had to encode Age column to the following values: [0-10), 0, [10-20),1, [20-30), 2, [30-40),3, [40-50),4, [50-60), 5, [60-70), 6, [70-80), 7[80-90), 8 [90-100), 9. I then dropped columns "race", "gender",'weight','payer_code','medical_specialty','max_glu_serum','A1Cresult','readmitted.' These columns did not show a statistical significance in the two population data sets or were not clinically significant. I also dropped the diabetes medications features to reduce the number of features and this information was captured in the DiabetesMed feature as well.

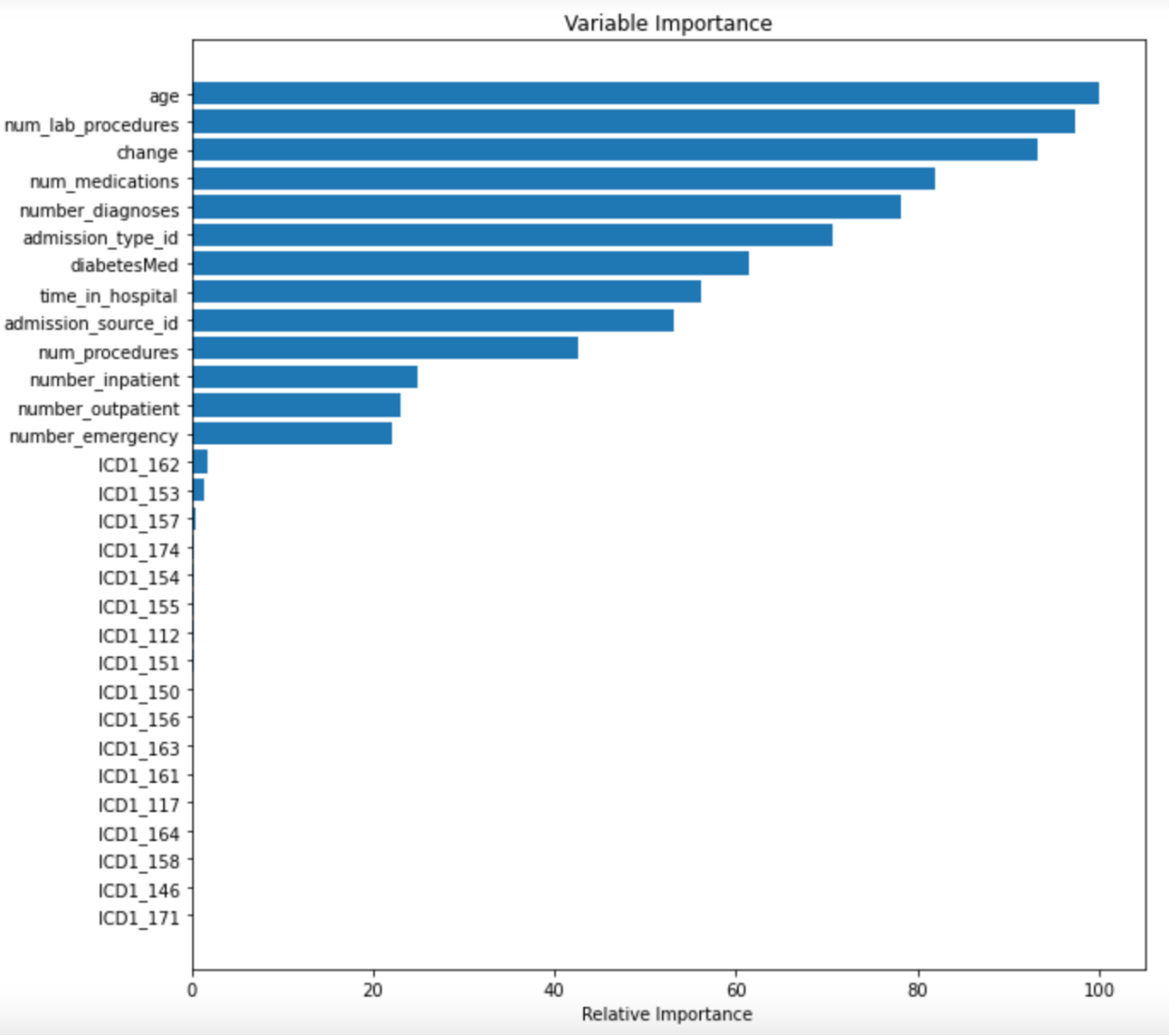
Model Description

I trained two models. I used Logistic Regression and Random Forrest. For Logistic Regression I did hyperparameter tuning on max iterations. I found the value of 100 to be optimal(train score = 0.498). For Random Forest I did n_estimators and found that the optimal value was 100(train score = 1.0).

Model Performance



Model Findings:



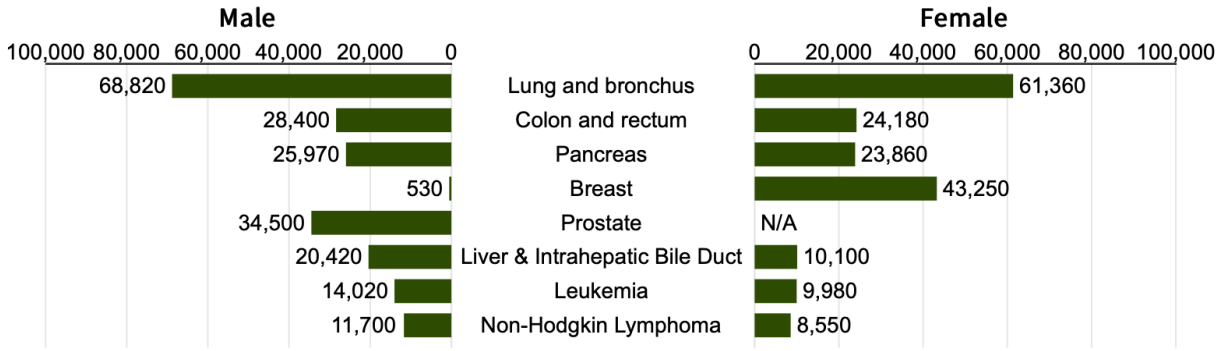
What is impressive is that the model was able to identify the codes that do have a high mortality rate.
ICD-9 162-> Lung Ca
ICD-9 153 -> Colon Ca
ICD-9 157 -> Pancreatic Ca
ICD-9 174 -> Breast Ca
ICD-9 154 -> Rectal Ca
ICD-9 155 -> Liver Ca
ICD-9 112 -> Candida Infection. (Common comorbid disease associated with cancer)
ICD-9 151 -> Stomach Ca
ICD-9 150 -> Esophageal Ca
ICD-9 156 -> Gallbladder Ca

ICD-9 163 -> Cancer of the pleura. (mesothelioma)
ICD-9 161 -> Laryngeal Ca
ICD-9 117 -> Fungal Infection. (Common comorbid disease associated with cancer)
ICD-9 164 -> Thymus Cancer
ICD-9 158 -> Retroperitoneum Ca
ICD-9 146 -> Oropharynx Ca
ICD-9 171 -> Soft Tissue Ca

ICD codes associated with neoplasms (ICD-9 codes 140-239) accounted for **3.6**% of the total of the total encounters.

Neoplasm ICD-9 Codes Found In Expired Patients	Number of Expired Patients with Diagnosis	Description	RF Model Rank of Importance
146	1	Oropharynx Cancer	16
150	1	Esophageal Cancer	9
151	6	Stomach Cancer	8
153	15	Colon Cancer	2
154	6	Rectal Cancer	5
155	12	Liver Cancer	6
156	2	Gall Bladder Cancer	10
157	20	Pancreatic Cancer	3
158	1	Retroperitneum Cancer	15
161	2	Laryngeal Cancer	12
162	66	Lung Cancer	1
163	1	Mesothelioma	11
164	1	Thymus Cancer	14
171	1	Soft Tissue Cancer	17
174	4	Breast Cancer	4
182	2		
183	4		
185	4		
188	2		
189	2		
191	5		
195	2		
196	2		
197	73	Secondary malignant neoplasm of respiratory and digestive systems	Not Ranked
198	34		
199	2		
200	2		
201	3		
202	6		
203	5		
204	3		
205	4		
208	2		
235	1		
238	4		
239	1		
Total	302		
Percentage of Expired Patients with Neoplasm Diagnosis	**12%**		

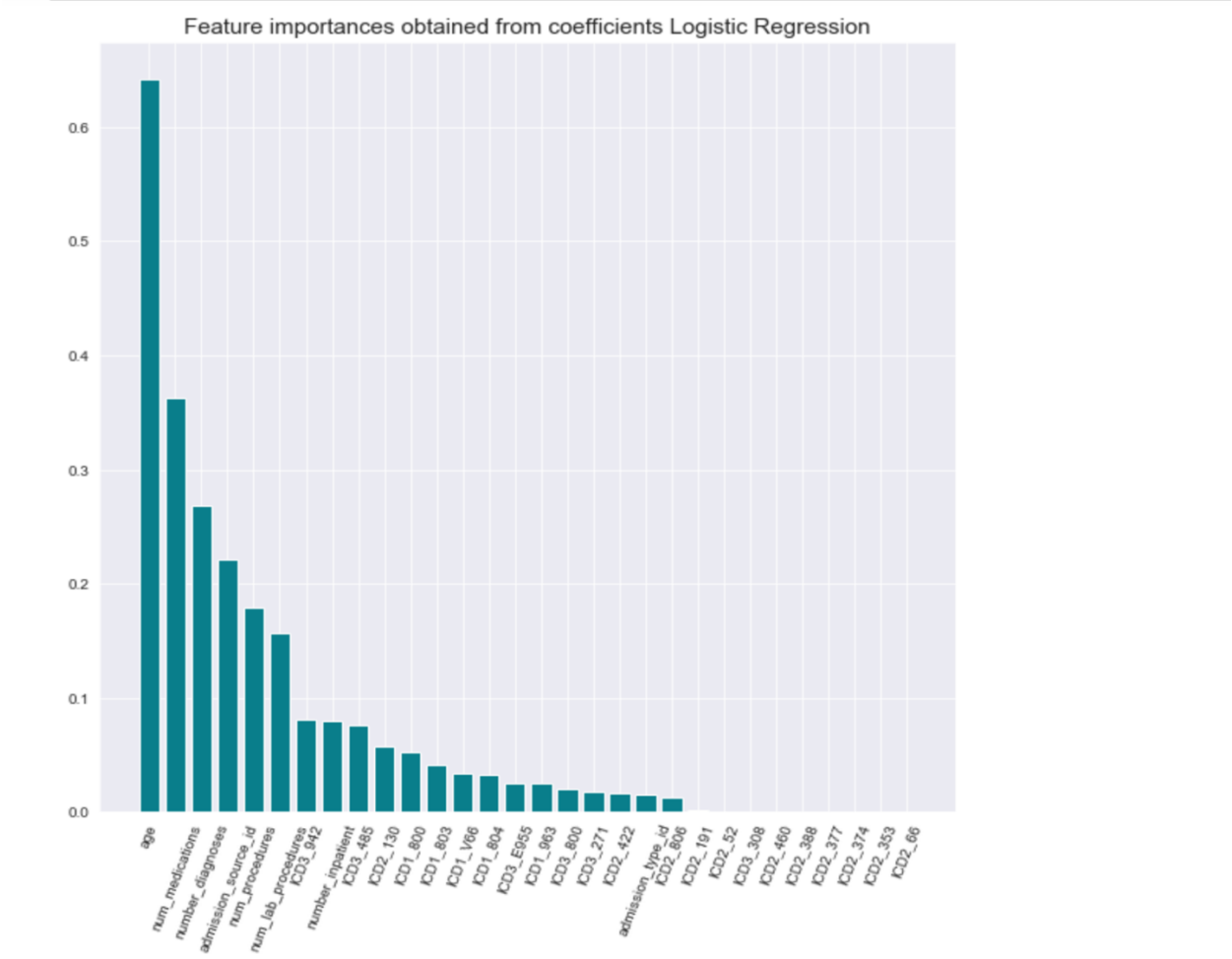
Leading Causes of Malignancy Deaths in the United States



Source: Cancer Facts & Figures 2022, American Cancer Society (ACS), Atlanta, Georgia, 2022.

It is remarkable how the top 5 leading causes of malignancy related deaths in the United States almost matches the ranking of importance of based on the random forest model! This shows that the model was able to draw an accurate conclusion of what features were associated with death. Consider that only 12% of the patients in the expired patient population had a malignancy diagnosis for Diagnosis 1 feature, yet the model was able to determine that these diagnoses when associated with other features such as Age, Number of Diagnoses, Number of Medications, etc. were a high predictor of mortality. Its ranking of diagnoses accurately reflects real world findings based on the CDC data that shown above.

Next, we will look at feature importance of the logistic regression model.



As we can see here, the first 6 features ranked in importance are clinically relevant and would correspond to how a clinician would judge overall morbidity in a patient. Here is the description of the ICD-9 Codes in order of importance from the model:

942 – Trunk Burn
485 – Pneumonia
130 – Toxoplasmosis
800 – Skull Fracture
803 – Closed Skull fracture
V66 – Encounter for Palliative Care
804 – Multiple fractures of the skull
E955 – Suicide and self-inflicted injury by handgun
963 – Poisoning by antiallergic and antiemetic drugs
271 - Disorders of carbohydrate transport and metabolism
422 – Acute Myocarditis
806 - Fracture of vertebral column with spinal cord injury
191 - Malignant neoplasm of brain

Overall, the diagnoses are in line with what would be considered grave conditions that would be considered for end-of-life care. However, the diagnoses are not as compelling as the diagnoses given in the random forest model.

I created a fictious patient with the following features: 99 years old, admitted from the ER, has been in the hospital for 15 days. Has had 10 lab draws done, has had 6 procedures done, is on 8 medications and has had 3 previous hospitalizations, 3 previous ER visits, 3 outpatient vists. Patient has a past medical history of 10 different diseases which include Lung Cancer, Pancreatic Cancer, AND Colon Cancer. This very sick patient has 0% survival from the logistic regression model and a 78% survival from the Random Forrest. I would have expected that the chance of survival would be low for both models. The model needs to be improved to improve its precision and recall scores to make it more reliable in the clinical setting.

NEXT STEPS

For further research I would like to get a dataset from inpatient hospitalization that contains a population that is general and not just diabetic population. I would like to be able to examine more features that would be clinically relevant when judging overall morbidity and mortality in a patient. I would also like to get more records from inpatient deceased patients to get a more overall balanced dataset.

APPENDIX

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Numeric	Grouped in 10-year intervals: 0, 10), 10, 20), ..., 90, 100)	0%

Feature name	Type	Description and values	% missing
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%

Feature name	Type	Description and values	% missing
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 0% 8%, “normal” if the result was less than 7%, and “none” if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%

Feature name	Type	Description and values	% missing
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” 0% if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and 0% “No” for no record of readmission.	0%