HEALTH CARE DATA MINING: REPORT 2

GROUP 12

ISYS 650

10/26/2015

PRADEEP ALAGIRI pradeep_alagiri@tamu.edu
SHARATH R SUBRAMANIYA sharath_rs@tamu.edu
BHARADWAJ VENKATESWARAN bharadwaj.v@tamu.edu

Table of Contents

INTRODUCTION

Survey of healthcare analytics area

The healthcare industry has modernized its operations, and is increasingly adopting electronic health records (EHRs). This has led to the deployment of new health information technology systems that constantly create, collect, and manage their information. Therefore, the amount of data available to clinicians and hospital administrators in the healthcare domain is growing at an exponential rate. However, despite such advances in technology and proven success in making data driven decisions in other domains such as retail, marketing etc., healthcare providers often report only minor improvements in decision making capability through the use of existing data.[1]

Costs and risk are not spread evenly across a population in many healthcare systems. Therefore, relatively small number of patients who are classified as high-risk patients tend to consume or utilize more medical resources than their peers. Also, studies are showing that deficits in managing care for these patients could lead to higher expenses. These findings necessitate and warrant systematic efforts that focus on *identifying high risk patients* to ensure that they receive the most efficient and effective care possible. [2]

McKinsey has recently stated that big data analytics has the potential to enable savings of more than $300 billion per year in U.S. healthcare. Also, McKinsey believes big data could help minimize inefficiencies in the following three areas [3]:

1. *Clinical operations*
2. *Research & development*
3. *Public health*

Healthcare delivery problem

Our world has a population of more than 7 billion people today. This growing trend of global population has triggered some of the biggest healthcare challenges that the healthcare industry is facing. One of the major business problems in healthcare domain is the assessment and management of clinical problems in the hospital. Governments and patients evaluate a hospital's quality of care by looking at performance data. In many countries, the data used to compare and evaluate outcomes is frequently based on Diagnosis Related Groups (DRGs) [5]. This involves classifying a patient's severity of illness or his mortality rate, or reducing the length of stay (LoS) of patients in the hospital based on the initial diagnosis.

For example, by classifying the patients based on the severity of the illness, the doctors can plan a treatment schedule well ahead of time. The hospital can also plan the logistics for the proposed treatment which can help in the efficient treatment of the patient. Similarly, by predicting the length of stay of a patient, hospital can solve the problem of manpower allocation. This can help the hospital in effective scheduling for admission of elective patients. [3]

Data mining problem: Classify risk of mortality

Risk of mortality is defined as the likelihood of dying. The four risk of mortality subclasses are numbered sequentially from 1 to 4 indicating respectively, minor, moderate, major, and extreme severity of illness. Based on our preliminary research we found that the variables listed below impact the risk of mortality: [4]

1. Principal Diagnosis coded in ICD-9-CM (PRINC_DIAG_CODE)
2. Secondary Diagnoses coded in ICD-9-CM (OTH_DIAG_CODE)
3. Procedures Coded in ICD-9-CM (PRINC_ICD9_CODE,OTH_ICD9_CODE)
4. Age (PAT_AGE)

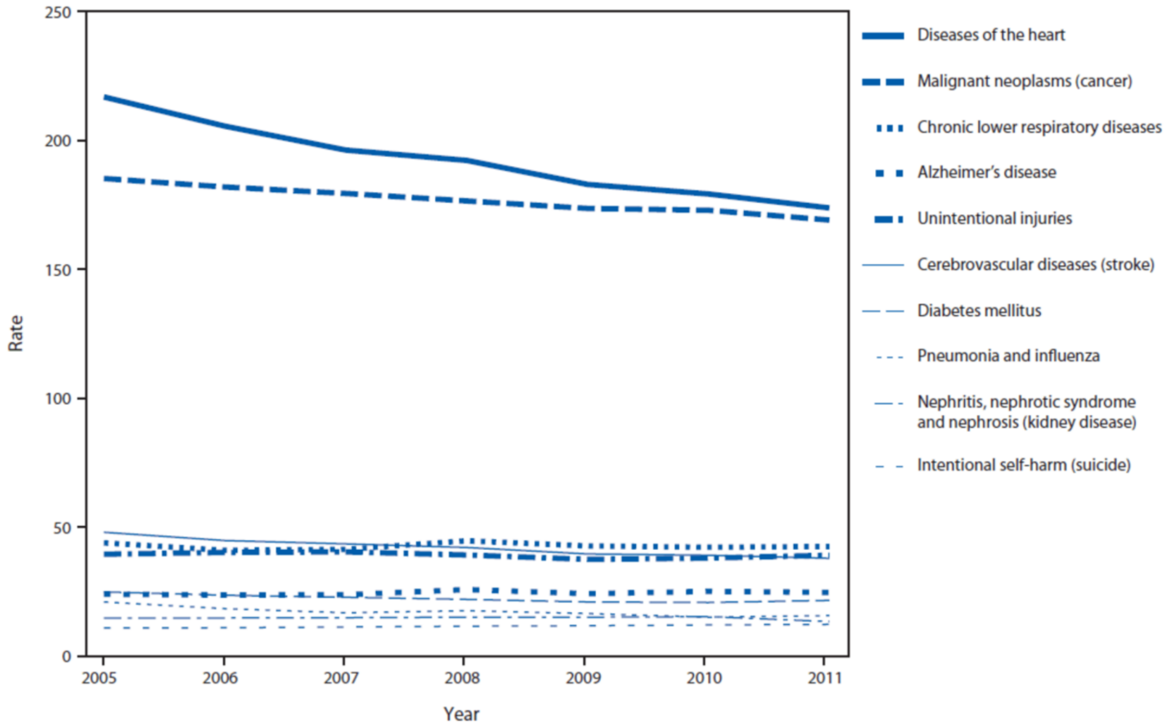
5. Sex (SEX_CODE)
6. Discharge Disposition (PAT_STATUS)

MOTIVATION AND PROBLEM FORMULATION

Motivation

Risk mortality refers to the in-hospital deaths and is one of the most important factors that define the quality of service at hospitals. In most of the cases the risk mortality of the patient is estimated during admission using history of the patient either through intuition or experience of the hospital staff. This helps them provide better services and improve intensive care to higher need patients, but with increasing complexity of diseases and increasing influential factors, it is imperative to have a more effective and reliable scoring system to provide accurate information on the risk mortality of the inpatients. Hospitals are profiled based on risk-standardized rates with increasing emphasis on improving patient management.

In recent years, there has been a significant increase in the number of patients admitted for heart diseases due to degrading food lifestyle and less health awareness. With increasing incidents, there has also been an increase in the risk mortality of inpatients with heart issues. According to a recent survey by CDC, diseases of the heart is one of the leading cause of deaths among patients in the United States.



Sources: CDC. National Vital Statistics System. Mortality public-use data files, 2011. Hyattsville, MD: US Department of Health and Human Services, CDC, National Center for Health Statistics; 2011. Available at http://www.cdc.gov/nchs/data_access/vitalstatsonline.htm.

Figure 1: Rate of deaths per 100,000 population, by leading cause of death — United States, 2005–2011

Every year more than 750,000 Americans have a heart attack and more than 610,000 people die of heart disease. There following are some of the common heart diseases among the population:

- Rheumatic heart disease
- Coronary artery disease
- Aortic aneurysm
- Pulmonary heart disease

Coronary heart disease alone kills more than 370,000 people every year in the United States of America [14]. As seen in Fig 1. The deaths due to the diseases of heart has seen a slight decline, but there has not been a significant drop in the death rates given the technological advancement in the treatment of heart diseases in the last decade. This makes it imperative to find the leading factors that increase the mortality risk for patients of heart disease. This will enable the hospital management to prioritize services for patients with high pre-determined risk of mortality.

Listed below are some of the important factors that affect the risk mortality among patients with diseases of the heart [14]:

1. Age
2. Sex
3. Race
4. Non-cardiac history - diabetes etc.
5. Cardiac history - Previous attacks or heart diseases
6. Vital measures - measure of vital elements in blood
7. Dependencies - Alcohol or drug dependencies
8. Family history or genetic disorder

Problem formulation: *The above factors have motivated us to concentrate our data mining problem towards a more specific health concern. Therefore, we propose to classify risk mortality for patients diagnosed with disease of the circulatory system or heart.*

Pre-Processing

Since our main focus is on risk mortality associated with heart diseases we have included principal diagnosis codes starting from 390 to 459 (inclusive). Based on our exploration of this data we found that there are around 96,000 records in Q1 and around 89000 records in Q2.

1. Principal Diagnosis coded in ICD-9-CM (PRINC_DIAG_CODE)

Reason:

The Principal Diagnosis is the condition established after ascertaining the condition that is chiefly responsible for the admission of the patient to the hospital for care.

Based on our research, Principal Diagnosis Code is one of the data elements that is used to determine the risk of mortality [4]. For the intents and purposes of this project, we are considering the ICD-9-CM codes that correspond to diseases related to the circulatory system.

Data Exploration:

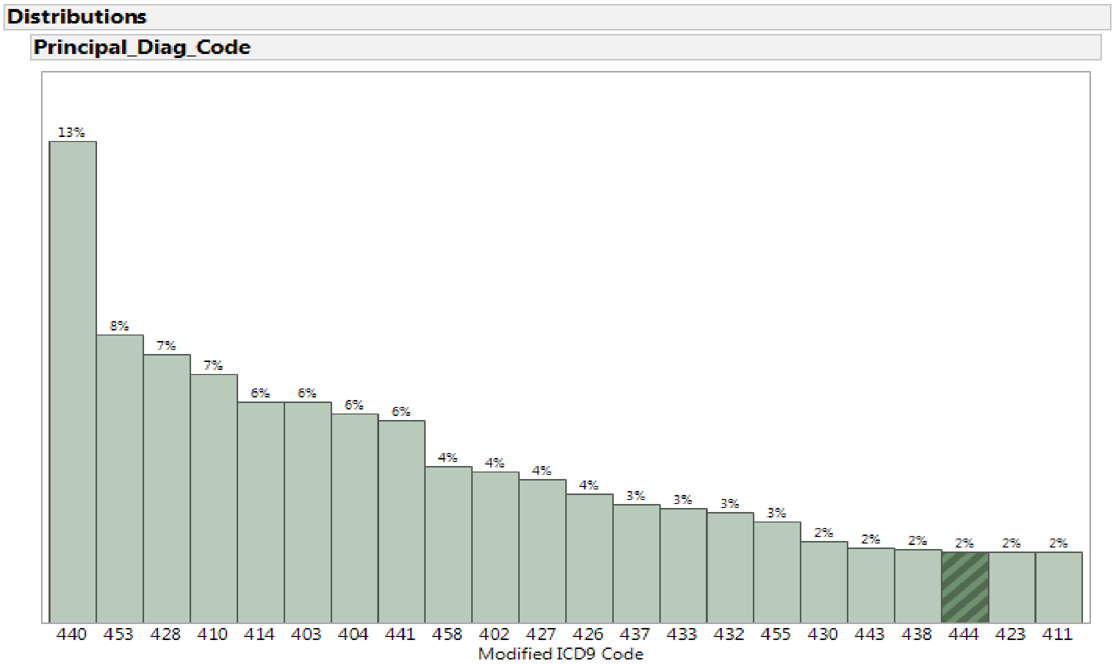


Figure 2: Distribution of records based on principal diagnosis code (first 3 digits)

Inference: In order to reduce the number of categories we have considered only the first 3 digits of the ICD 9 code for the principal diagnosis code column in the dataset. Figure 2 depicts the number of records in each of these *modified ICD 9 codes (first 3 digits)*. We can see that 440 (arterial diseases) constitute almost 13% of the 96000 records.

Data cleansing approach:

Since the records were filtered based on the ICD 9 codes column to get records pertinent to circulatory diseases, the PRINC_DIAG_CODE field does not contain any null or empty values that needs to be cleansed.

2. Secondary Diagnosis coded in ICD-9-CM (OTH_DIAG_CODE)

Reason:

Other Diagnoses include all conditions that coexist at the time of inpatient admission or ambulatory surgical service, or develop subsequently, which affect the treatment received and/or length of stay. Based on our research, we believe that pre-existing medical condition could play a major role in determining the risk mortality of the patient.

Data exploration:

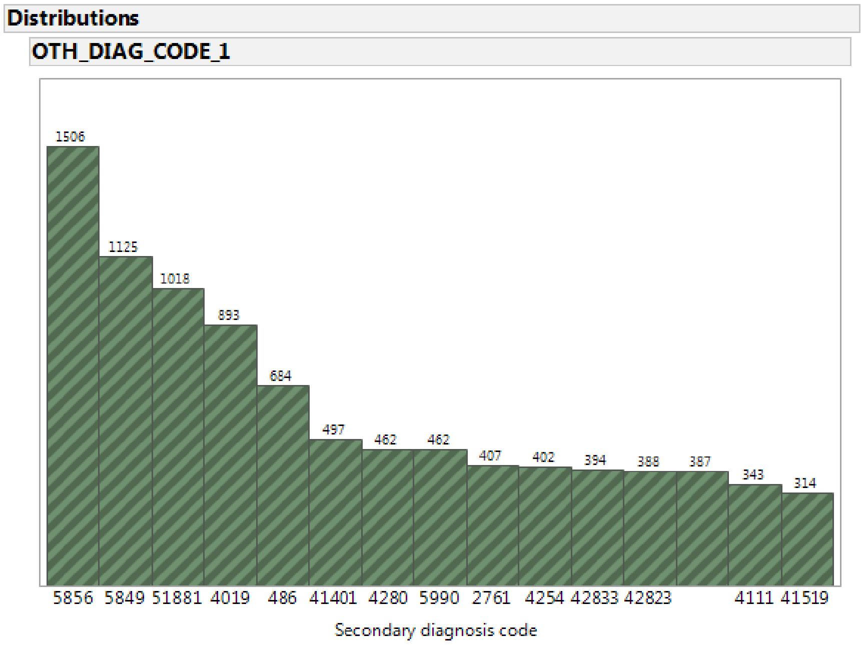


Figure 3: Subset of distribution of number of records based on other diagnosis code (ICD 9 codes)

Inference: As seen in figure 3, the distribution of records for secondary diagnosis is spread across multiple ICD 9 codes. Since this is just a supporting variable in cases where a principal diagnosis code is not present, we do not intend to clean this field.

3. Age (PAT_AGE)

Reason:

Based on our research, patient age is one of the data elements that is used to determine the risk of mortality [4]. For example, the risk mortality of a patient who is 70 years old and suffers from an ischemic heart disease is higher than a patient whose age is 40 years and suffers from the same heart condition.

Data Exploration:

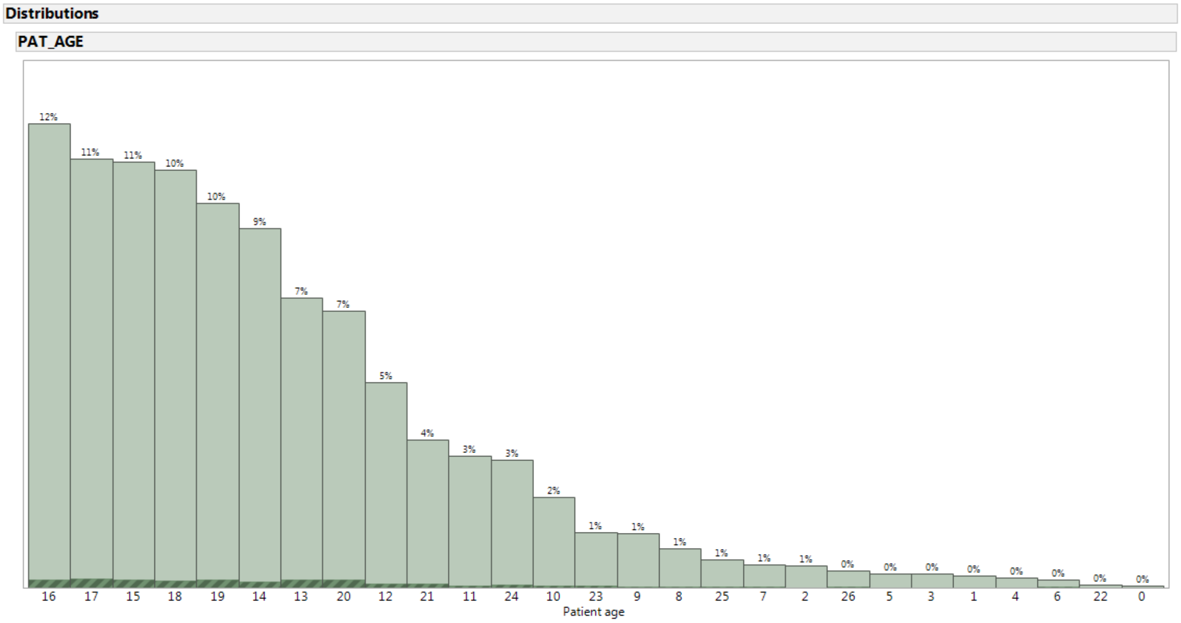


Figure 4. Distribution (percentage) of total records based on patient age

Inference: From figure 4, we can understand that majority of the patients fall into the age group categories 14-19 which signify age groups 55-84. This is because we are considering the case of circulatory diseases which are more prevalent in people of this age group.

Data cleansing approach:

The field PAT_AGE doesn't contain any invalid or erroneous data values.

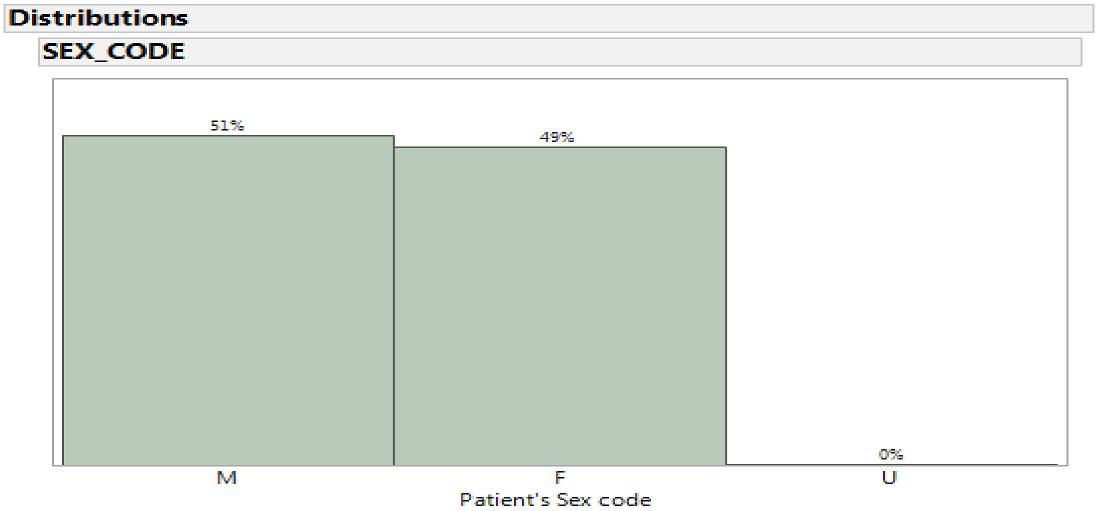
4. Sex (SEX_CODE)

Reason:

Based on our research, patient age is one of the data elements that is used to determine the risk of mortality [4]. For example, Heart attack symptoms in women may be different from those experienced by men. Many women who have a heart attack do not know it. Women tend to feel a burning sensation in their upper abdomen and may experience lightheadedness, an upset stomach, and sweating. Because they may not feel the typical pain in the left half of their chest, many women may ignore symptoms that indicate they are having a heart attack [10]. Accordingly the diagnosis for men and women may be different.

Data exploration:

Figure 5. Distribution (percentage of total records) based on patient sex code



Inference: As seen in figure 5, the distribution of data is pretty even among male and female genders. This will help in classifying the risk mortality for both men and women effectively.

Data cleansing approach:

Upon investigation, we found that only a negligible percentage (4%) of the total number of records had empty values. These records can be removed while preprocessing.

5. Discharge Disposition (PAT_STATUS)

Reason:

A patient discharge status code is a two-digit code that identifies where the patient is at the conclusion of a healthcare facility encounter. Based on our research, patient age is one of the data elements that is used to determine the risk of mortality [4]. This parameter can provide us the details on how a patient discharge status can affect the risk of mortality. For example, if the risk mortality for acute myocardial infarction with a discharge code *07* (Left against Medical Advice or Discontinued Care) is high, the hospital can allocate adequate resource to take care of the patient once he is discharged.

Data exploration:

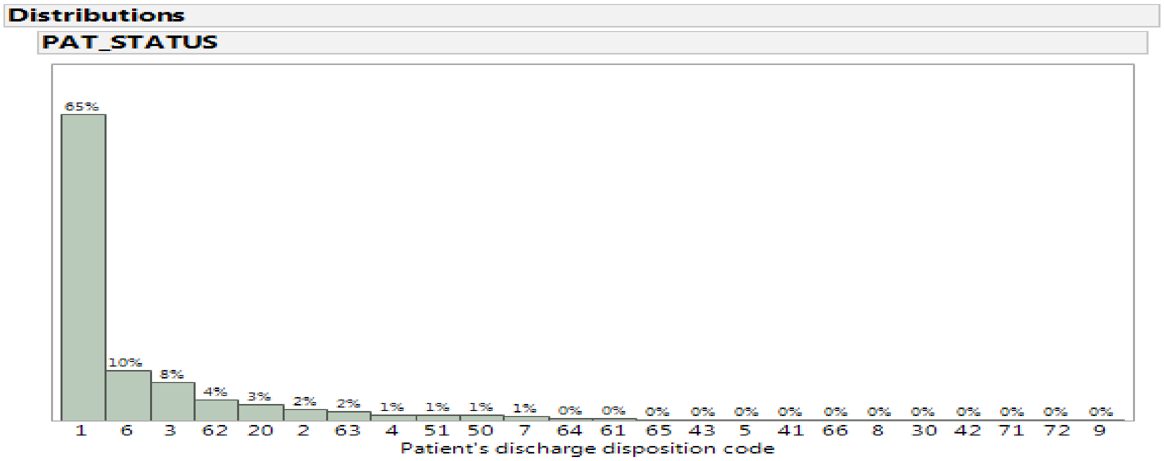


Figure 6. Distribution (percentage of total records) based on patient's discharge disposition code

Inference: As seen in figure 6, almost 65% of the patients had a *routine discharge*. This was expected.

Data cleansing approach:

Upon investigation, we found that only a negligible number of records (247) of the total number of records (96000) had empty values. These records can be *removed* while preprocessing.

6. Race (RACE)

Reason:

Based on our research, patient age is one of the data elements that is used to determine the risk of mortality [4]. For example, most of the studies found hypertension to be significantly higher in Blacks than Whites [15].

Data exploration:

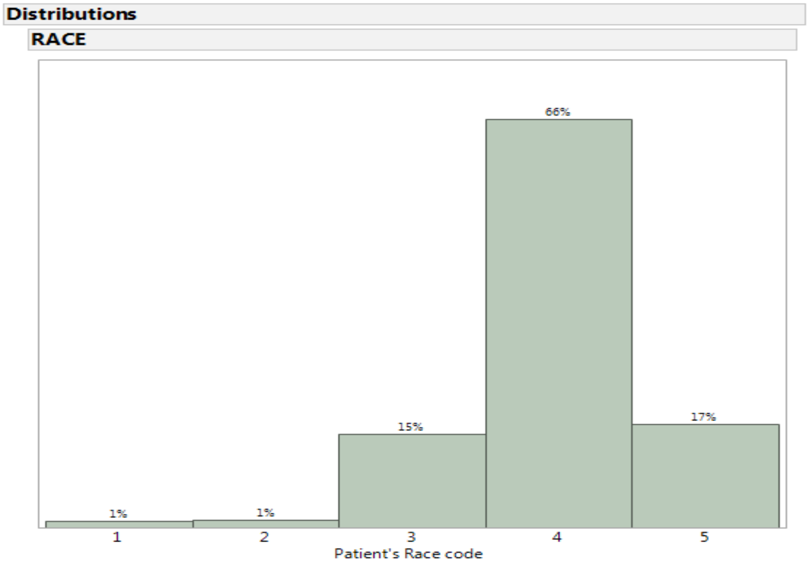


Figure 7: Distribution (percentage of total records) based on patient's race code

Inference: From figure 7, we can clearly see that most of the patients are 'white' (category 4). Also, 17% of the total number of patients are listed in the category 5 which belongs to 'other'.

Data cleansing approach:

Upon investigation, we found that a negligible number of records (411) of the total number of records (around 96000) had empty values. These records can be *removed* while preprocessing.

7. Ethnicity (ETHNICITY)

Reason:

Changes in circulatory diseases differ by race as well as ethnicity. To better differentiate various trends between African-American, Hispanic, and White ethnic groups in circulatory disease we have included the ethnicity as one of the data elements that is used to determine the risk of mortality.

Data exploration:

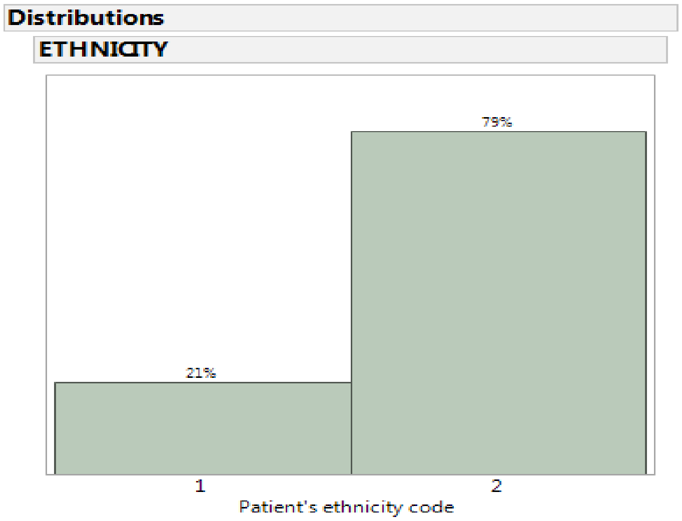


Figure 7. Distribution (percentage of total records) based on patient's ethnicity code

Inference: From figure 7, we can observe that most of the patients are not of Hispanic origin (category 2). Also, 21% of the total number of patients are listed in the category 1 which belongs to 'Hispanic origin' as expected since we are dealing with data from Texas which has a significant Hispanic population.

Data cleansing approach:

Upon investigation, we found that a negligible number of records (187) of the total number of records (around 96000) had empty values. These records can be *removed* while preprocessing.

8. Type of admission (TYPE_OF_ADMISSION)

Reason:

This code indicates the manner in which the patient was admitted to the health care facility. The type of admission (emergency, trauma, urgent etc.) can determine the severity of the patient's condition, which can in turn have an impact on the patient's risk of mortality.

Data exploration:

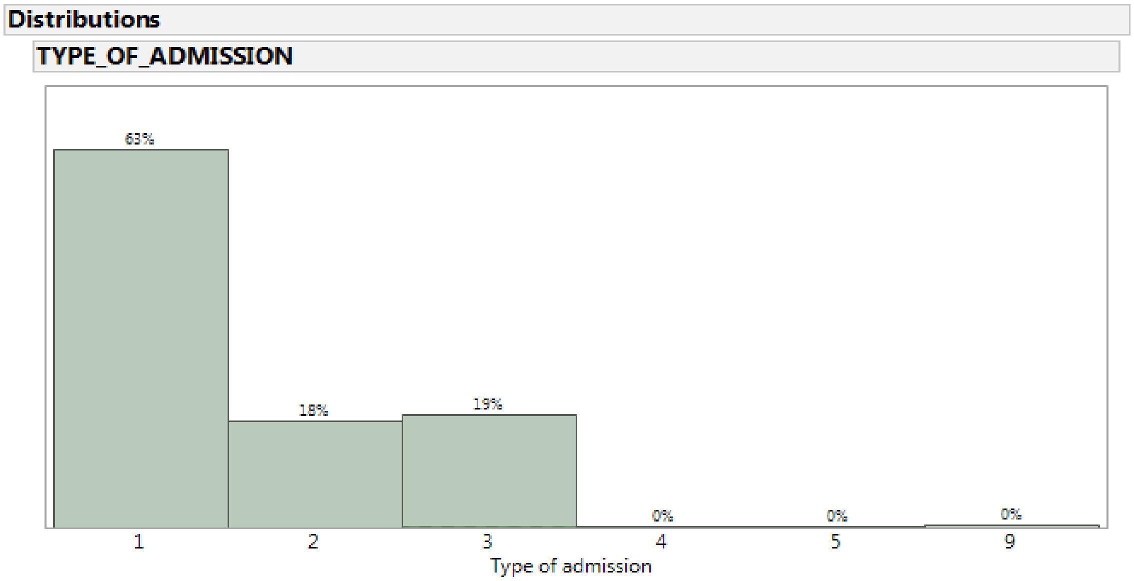


Figure 9: Distribution (percentage of total records) based on type of admission

Inference: From figure 9, we can clearly see that most of the admissions are emergency admissions (category 1). Also, 18% of the total number of admissions were listed in the category 2 which belongs to 'urgent admissions'.

Data cleansing approach:

Upon investigation, we found that though there are no empty values in the field but a negligible number of records (247) of the total number of records (around 96000) belong to category 9 (no information available). These records can be *removed* while preprocessing.

9. Admitting diagnosis (ADMITTING_DIAGNOSIS)

Reason:

This code indicates the admitting diagnosis of the patient who was admitted to the health care facility. The type of admission (ICD-9-CM Codes) can accurately determine the type of condition that patient suffered from when he was initially admitted. Certain admission codes can indicate potential seriousness of the patient condition, which in turn can impact the risk of mortality.

Data exploration:

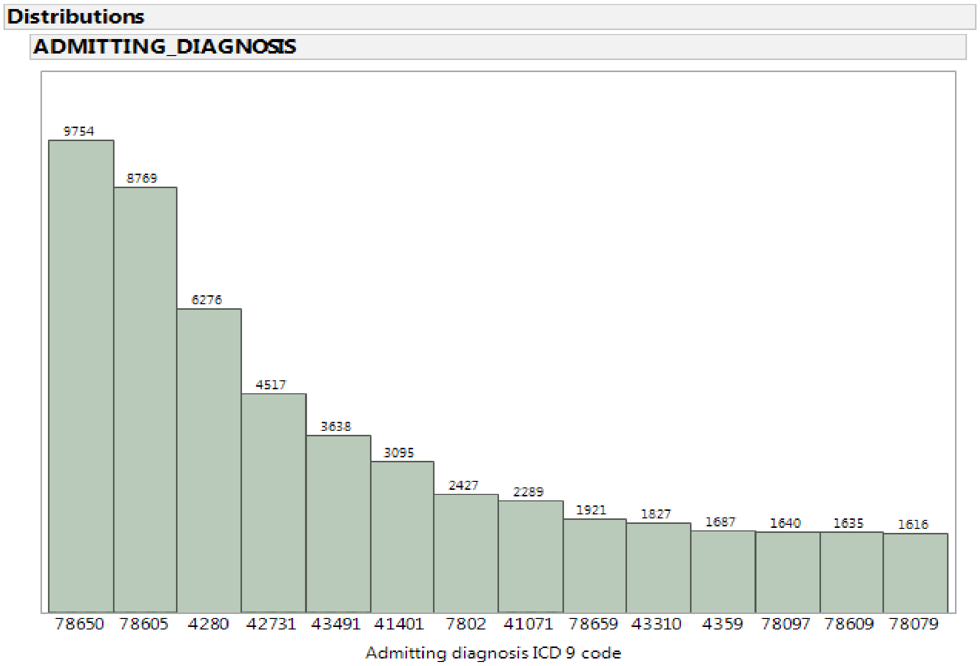


Figure 10. Distribution (percentage of total records) based on admitting diagnosis ICD9 code

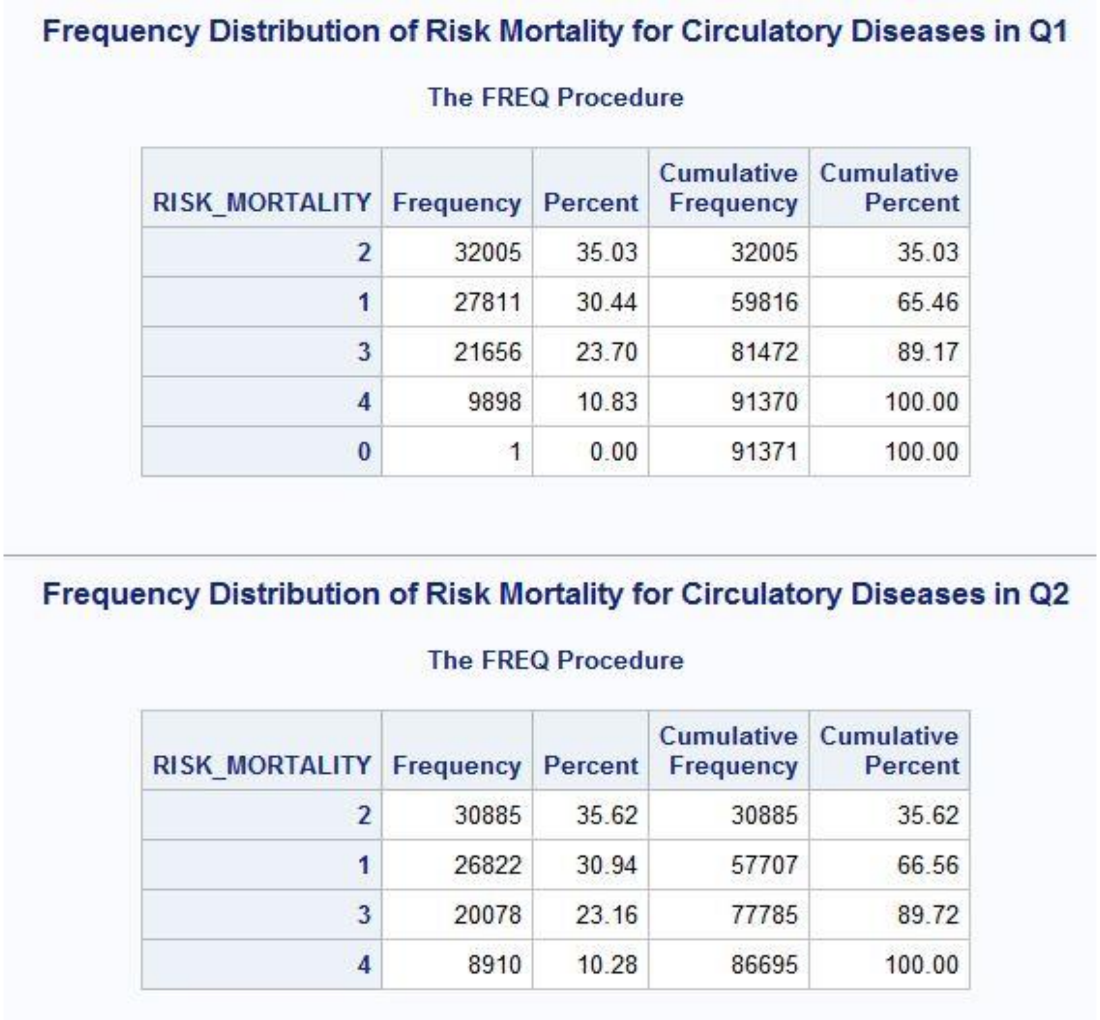
Inference: From figure 10, we can clearly see that majority of the patients are admitted with an ICD 9 code *786.50 (unspecified chest pain) followed by 786.05 (shortness of breath)* which are symptoms of circulatory diseases.

Data cleansing approach:

Upon investigation, we found that a negligible number of records (94) of the total number of records (around 96000) had empty values. These records can be *removed* while preprocessing.

10. Risk Mortality

This variable is the output variable which we are going to classify in our data mining project.



Frequency Distribution of Risk Mortality for Circulatory Diseases in Q1

The FREQ Procedure

RISK_MORTALITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	32005	35.03	32005	35.03
1	27811	30.44	59816	65.46
3	21656	23.70	81472	89.17
4	9898	10.83	91370	100.00
0	1	0.00	91371	100.00

Frequency Distribution of Risk Mortality for Circulatory Diseases in Q2

The FREQ Procedure

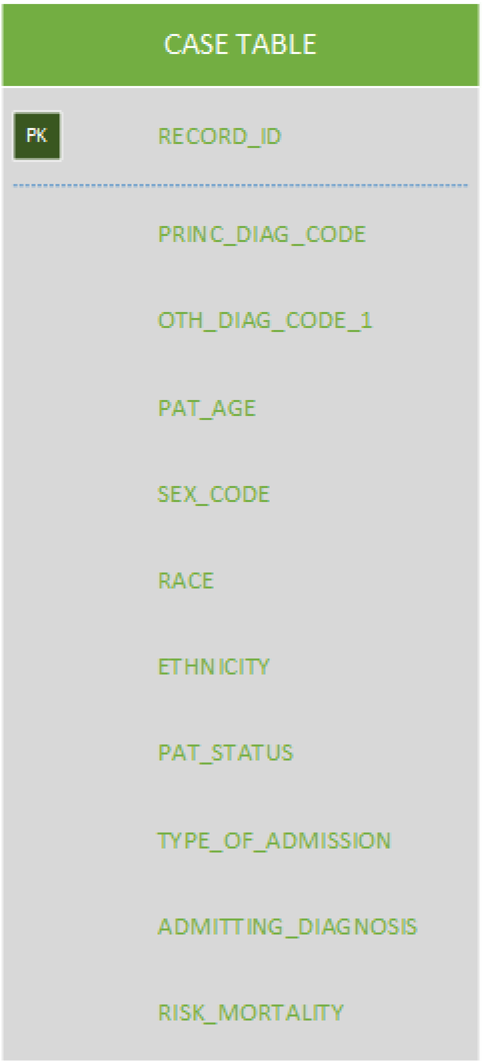
RISK_MORTALITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	30885	35.62	30885	35.62
1	26822	30.94	57707	66.56
3	20078	23.16	77785	89.72
4	8910	10.28	86695	100.00

Figure 11: Distribution (number of records) for every class of risk mortality

In figure 11, we have indicated the number of records under each class of risk mortality for both the quarters Q1 and Q2 separately. In our classification problem, we consider output risk mortality classes 3 (major) and 4 (extreme) important since these demand special medical attention. Thus, from figure 11 we can understand that there might not be a need for oversampling as there around 33% of the records belong to risk mortality classes 3 or 4 in both the quarters.

SOLUTION

Case Table

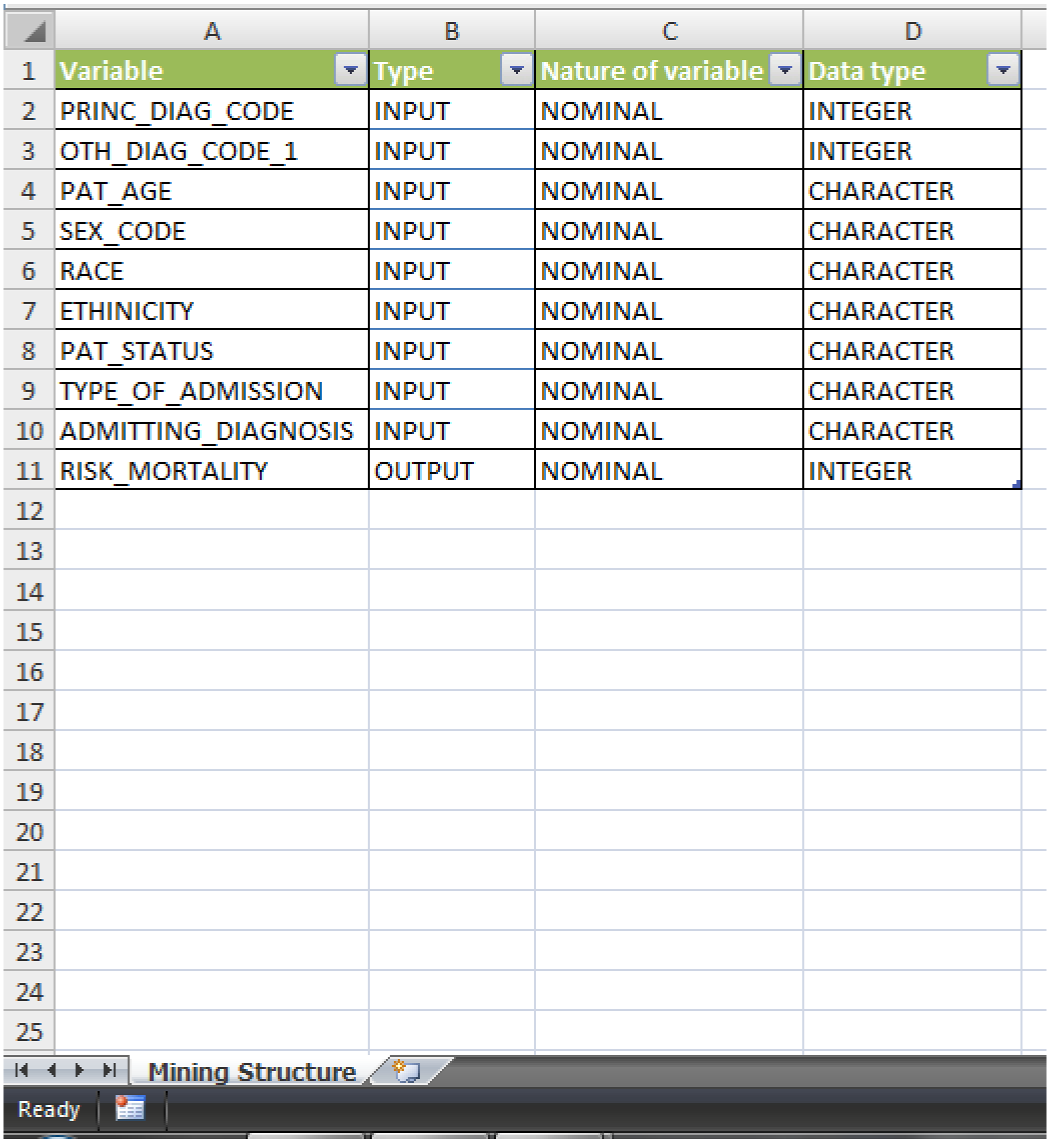


	CASE TABLE
PK	RECORD_ID
	PRINC_DIAG_CODE
	OTH_DIAG_CODE_1
	PAT_AGE
	SEX_CODE
	RACE
	ETHNICITY
	PAT_STATUS
	TYPE_OF_ADMISSION
	ADMITTING_DIAGNOSIS
	RISK_MORTALITY

Figure 12: Case Table

The schema of the case table is shown in figure 12. For our data mining problem, since we are dealing with classification of risk mortality we do not require either dataset 2 or dataset 3 because they deal with insurance and billing information. *Thus, for our solution we do not need a nested case table.*

Mining Structure



Variable	Type	Nature of variable	Data type
PRINC_DIAG_CODE	INPUT	NOMINAL	INTEGER
OTH_DIAG_CODE_1	INPUT	NOMINAL	INTEGER
PAT_AGE	INPUT	NOMINAL	CHARACTER
SEX_CODE	INPUT	NOMINAL	CHARACTER
RACE	INPUT	NOMINAL	CHARACTER
ETHINICITY	INPUT	NOMINAL	CHARACTER
PAT_STATUS	INPUT	NOMINAL	CHARACTER
TYPE_OF_ADMISSION	INPUT	NOMINAL	CHARACTER
ADMITTING_DIAGNOSIS	INPUT	NOMINAL	CHARACTER
RISK_MORTALITY	OUTPUT	NOMINAL	INTEGER

Figure 13: Mining structure

As indicated in figure 13, our mining problem entails dealing with mostly nominal (categorical) variables. Also, a subset of these variables might be used in the future for various mining models to classify the output variable risk mortality.

CONCLUSION

As seen in the above sections, we are working on a more focused problem of classifying risk mortality for patients with circulatory diseases. We have identified the potential input variables for classifying risk mortality and developed a case table and a mining structure based on the identified variables.

REFERENCES

1. Jesus J Caban and David Gotz.Visual analytics in healthcare – opportunities and research challenges. *Journal of the American Medical Informatics Association.*
2. David Gotz, Harry Stavropoulos and Jimeng Sun. ICDA: A Platform for Intelligent Care Delivery Analytics. *US National Library of Medicine National Institutes of Health*
3. Gordon H. Robinson, Louis E. Davis, and Richard P. Leifer. Prediction of Hospital Length of Stay. *Health Services Research*
4. 3M Health Information Systems. APR^{TM}-DRG Classification Software -Overview
5. P. Fontaine. Using Severity Adjustment Classification for Hospital Internal and External Benchmarking. *2004 IFHRO Congress & AHIMA Convention Proceedings, October 2004*
6. Length of Stay (LoS). Wikipedia. *https://en.wikipedia.org/wiki/Length_of_stay*
7. Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. Health Information Science and Systems.
8. Laursen TM, Wahlbeck K, Hällgren J, Westman J, Ösby U, Alinaghizadeh H, et al. (2013) Life Expectancy and Death by Diseases of the Circulatory System in Patients with Bipolar Disorder or Schizophrenia in the Nordic Countries. PLoS ONE 8(6): e67133. doi:10.1371/journal.pone.0067133
9. Statistics about data elements. *https://www.health.ny.gov/statistics/sparcs/sysdoc/elements_837/table_x12.htm*
10. Texas Heart Institute. *http://www.texasheart.org/HIC/Topics/HSmart/women.cfm*
11. Margaret Jean Hall, Ph.D.; Shaleah Levant, M.P.H.; and Carol J. DeFrances, Ph.D.(2013) “Trends in Inpatient Hospital Deaths: National Hospital Discharge Survey, 2000–2010” (Issue No. 118 March 2013)

12. Drye, Elizabeth et al. “Comparison of Hospital Risk-Standardized Mortality Rates Using In-Hospital and 30-Day Models: Implications for Hospital Profiling.”*Annals of Internal Medicine* 156.1 Pt 1 (2012): 19–26. *PMC*. Web. 26 Oct. 2015.
13. Krumholz HM, Wang Y, Mattera JA, Wang Y, Han LF, Ingber MJ, et al. An Administrative Claims Model Suitable for Profiling Hospital Performance Based on 30-Day Mortality Rates Among Patients With Heart Failure. Circulation. 2006; 113(13):1693–701. [PubMed: 16549636]
14. http://www.cdc.gov/heartdisease/
15. Kurian AK, Cardarelli KM. Racial and ethnic differences in cardiovascular disease risk factors: a systematic review.