CS 510 Final Project Report

Hanna Lu

**Introduction**

Acutely ill patients are discharged to different locations such as SNF (skilled nursing facilities), hospice, long-term care hospitals, rehabilitation facilities, or home according to doctors’ judgement of the patients’ health condition at the end of their stay in the hospital. Those who are sent home are considered to be in a steady and safe condition at the end of their hospital stay. However, a significant percentage of patients who are sent home die within 90 days of discharge from the hospital. In this study, we aim to identify the mortality risk indicators to help doctors reevaluate their patient discharge decisions in hope of reducing mortality rate among patients who are sent home after their stay in the hospital.

The aim of this study is to find significant factors that impact the mortality rate of critically ill patients within 90 days after they are discharged from the hospital. The findings of this study may help doctors to re-evaluate their discharge decisions to reduce the mortality rate.

**Data Source**

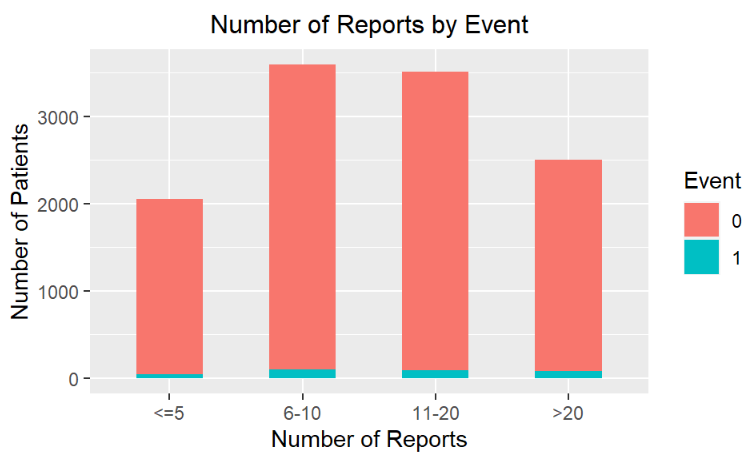
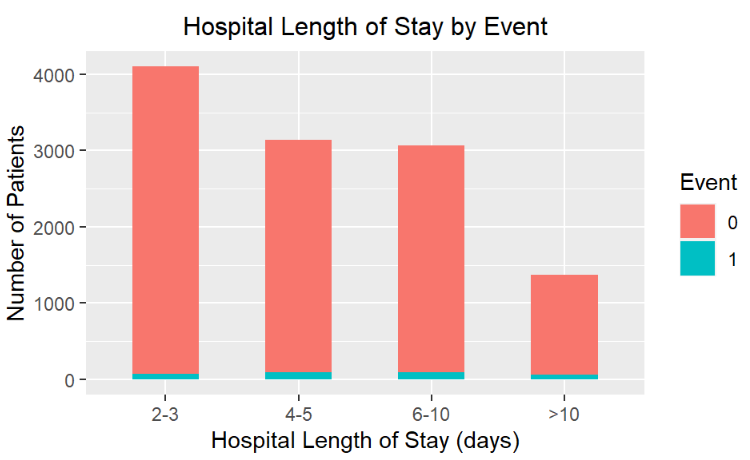
This study investigates variables extracted from 11 data sets from the MIMIC-III (Medical Information Mart for Intensive Care III) database (https://mimic.physionet.org/gettingstarted/access/) where all patients have been admitted to one or more Intensive Care Units (ICU) at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. This database is publicly available, and the patient records are fully deidentified. [1]

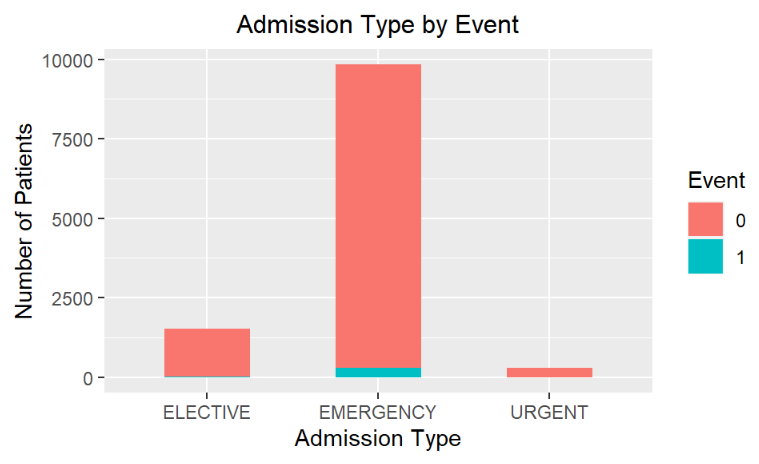
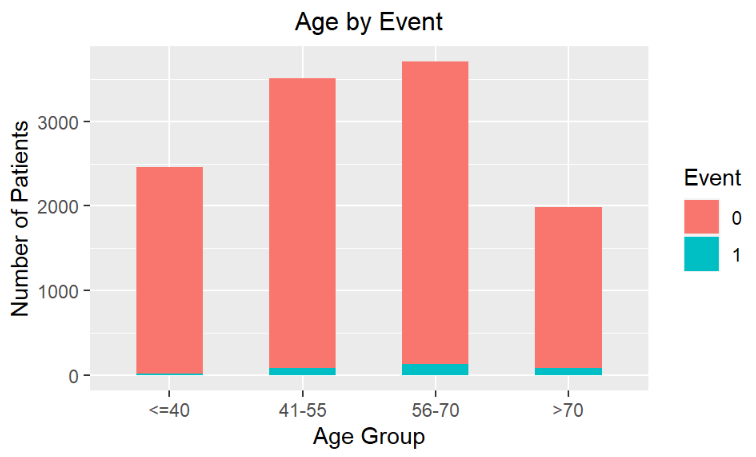
This study focuses on adult patients who are sent home after their hospital stay. Therefore, patients who are below the age of 18 at the time of being discharged from the hospital are excluded from this study. Patients who stayed in the hospital for one day or less, who died in the hospital or within one day of being discharged from the hospital, and those who were discharged to hospice or hospice equivalent facilities are also excluded. After applying these criteria, there are 11,666 patients in the data.

**Data Preprocessing**

In order to reduce the potential bias on the estimated effect size of each variable caused by not including all underlying variables of the true model while keeping the dimensionality low, we investigated variables that are likely to be significant based on our judgement, previous research, and the accessibility of the data. These variables include demographic information such as age, gender, ethnicity, language, religion, and marriage status, general information such as insurance, admission type, the last ICU unit the patient was in before discharge, the last vital sign measurements [3], and other information such as the number of medical notes, and the length of hospital stay.

Most variables have been engineered. Categorical variables are regrouped based on the number of observations in each category and the similarity of different categories to reduce dimensionality and to balance the sample size of each category while keeping the new groups meaningful. Numerical variables are grouped into intervals since they may not have a linear relationship with mortality, especially with the presence of other variables. Figure 1 shows the bar plots of categorical variables by event (event of death).





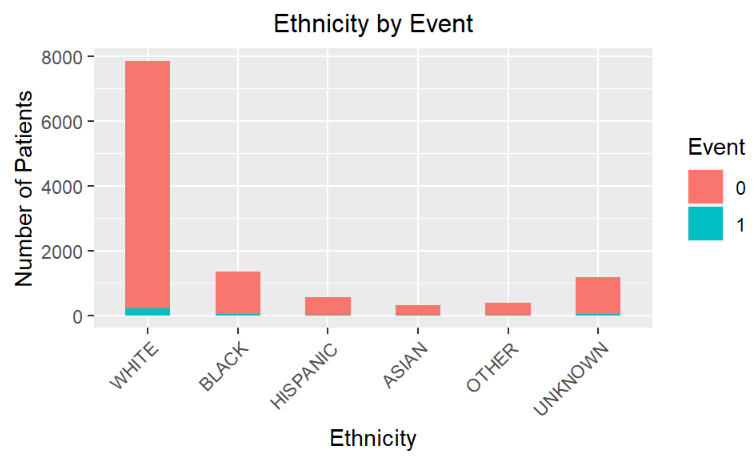
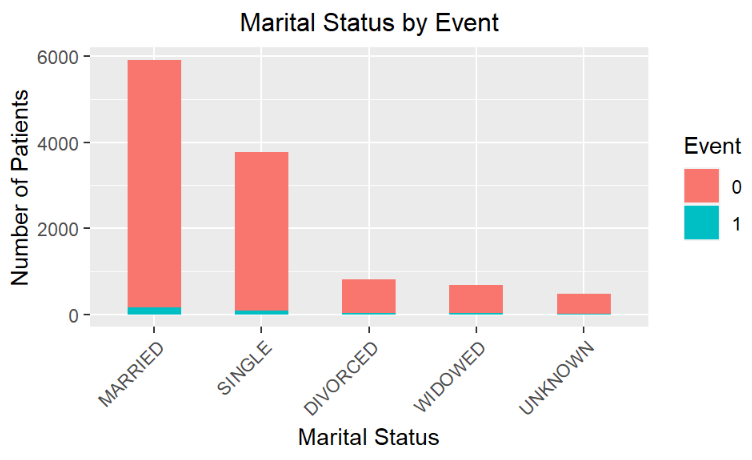


Figure 1. Categorical variables by Event

Since all date variables are masked in the same fashion to deidentify the patients in the database, the age of a patient at time of discharge from the hospital was calculated as the difference between the date of discharge and the date of birth. It was then grouped into 4 intervals.

The data were collected for research purposes and thus it is quite clean. However, there are missing values in all six vital sign variables (Figure 2). The majority of the missing values in the six sign variables are in the same rows (Figure 3), which suggests that the missingness of one vital sign variable is highly correlated with the missingness of each of the other five. This also suggests that the missing values are not Missing Completely At Random (MCAR), nor Missing At Random (MAR). It is reasonable to assume that the missing values are from the patients who were not required to take the measurements. Since they are not missing at random, it is not suitable to impute these values or to delete them because they may carry some information that may influence the mortality risk. Therefore, we will keep the missing values, and code them as “Missing”, and code the non-missing values as “Low”, “Normal”, and “High”, where the “Normal” values are based on domain knowledge.

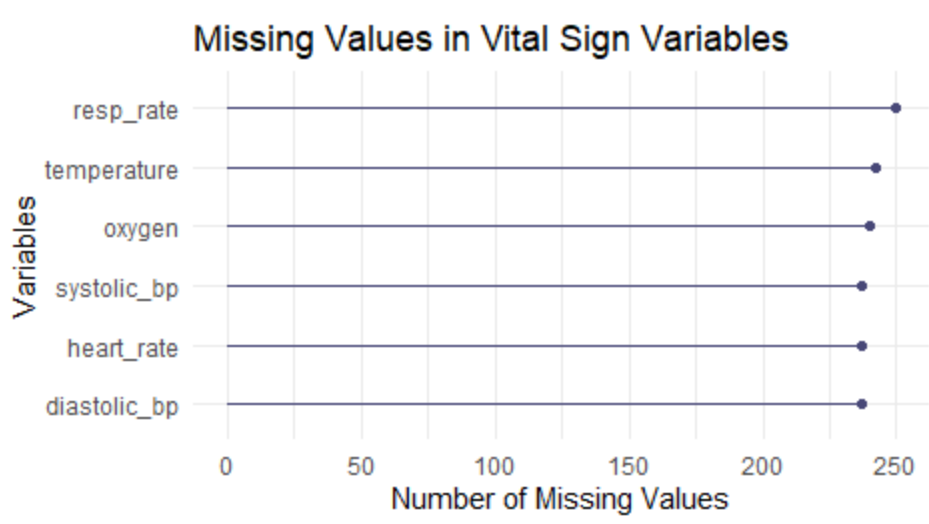


Figure 2. Missing values in vital sign variables

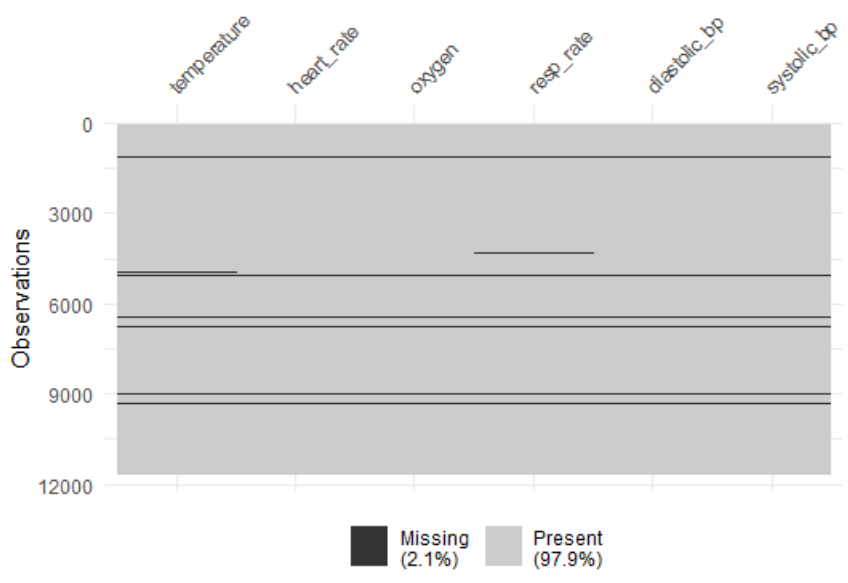


Figure 3. Location of missing values in vital sign variables

There are some outliers in each of the six vital sign variables. They are detected by domain knowledge instead of statistical methods. This is because there is a pre-defined normal range for each of the six vital sign measurements. The values that are below or above the normal range may still be valid values since many patients especially critically ill patients may have an abnormal vital sign measurement. However, those valid values may be identified as outliers with statistical methods. While using domain knowledge, we can precisely identify the values that are erroneous. For example, negative or 0 body temperature, 0 or over 100 oxygen saturation, 0 blood pressure or heart rate would be considered outliers. These erroneous outliers are temporarily treated as missing values and grouped in the Missing category. Other outlier treatment options will be explored in future work.

**Methods**

This is a typical survival analysis problem where the beginning of the study period is the time when the patient is discharged from hospital, and the end of the study period is 90 days after discharge. Patients who died after 90 days of discharge and those who survived the entire study period were censored. The study period was defined to be 90 days because the patients who were admitted to the hospital after 2008 were followed up for a maximum of 90 days after being discharged from the hospital. Although those who were admitted before 2008 were followed up for a maximum of 4 years after discharge, to make it consistent for all patients, we take the shorter follow-up period of the two as the study period. Two key methods that were employed in this study are the Cox Proportional Hazards model for survival analysis, and the comorbidity index scores to enhance the model results.

The Cox Proportional Hazards model was used to perform the survival analysis because this model allows multiple variables to be analyzed for their effects on the hazard (i.e. the probability of dying) [statistics review]. The Cox Proportional Hazards model from the ***survival*** package in R requires an Event variable and a Duration (or Time) variable. The Event variable has binary values with 1 indicating death within 90 days of discharge, and 0 indicating survival at 91st day after discharge. The Duration variable contains the number of days between discharge and death. If a patient survived the follow up period, the duration is coded as 300 days.

Comorbidity index is often used to assess the impact of comorbid diseases on a certain health condition [2], or to analyze mortality of a certain disease. They are derived from the ICD9 or ICD10 (International Statistical Classification of Diseases) diagnosis code data using the ***comorbidity*** package in R. The comorbidity function in this package produces variables relative to each comorbidity domain, comorbidity score, weighted comorbidity score, and the categorization of such scores with one row per individual. There are two score options: the weighted Charlson score (charlson) and the weighted Elixhauser score (elixhauser). The weighted Charlson score is used in this study. The comorbidity categorization data and each of the comorbidity scores were joined with the other variables for the survival analysis separately to investigate their effect independently to avoid correlation.

**Results**

Concordance index is a measure of performance of survival analysis models. The concordance index of the final Cox Proportional Hazards model with all significant variables is 0.829. Table 1 shows the covariates (and their levels if they are categorical), their sample sizes, hazard ratios (and the confidence intervals of the hazard ratios in the parentheses), and the p-values. The hazard ratios are the exponentiated coefficients of the Cox Proportional Hazards model. The black squares and the horizontal lines that run through them are a graph visualization of the hazard ratios and their confidence intervals. The dashed vertical line represents the hazard ratio of 1. Those that have the confidence interval line completely on the left side or completely on the right side of the dashed vertical line (i.e. not crossing the vertical line) are significantly different from their base lines.

Table 1 shows that that age and some of the comorbidities are the most significant contributors to the risk of mortality. The older the patients are, the higher the risk. The risk of mortality for a patient of 70 years or older would be 4.23 (hazard ratio) times higher than a patient of 40 years or younger. The hazard ratio is 3.31 for a patient between the age of 56 and 70 against a patient of 40 years or younger, 2.64 for a patient between the age of 41 and 55. The risk of mortality for a patient with metastatic solid tumor (the metacanc variable) is 6 times higher than a patient without this disease. The hazard ratio is 3.72 for a patient with moderate or severe liver disease (the msld variable), 2.91 for a patient with cancer of any malignancy (the canc variable), 2.69 for a patient with hemiplegia or paraplegia (the hp variable), 1.65 for a patient with renal diseases (the rend variable), 1.81 for a patient with congestive heart failure (the chf varible), 1.48 for a patient with peripheral vascular disease (the pvd variable) with a marginal p-value though (0.056). The risk of mortality for a patient admitted to the hospital through emergency departments is 3.09 times higher than a patient admitted to the hospital electively. There’s no significant difference in the risk of mortality between electively admitted patients and those admitted through urgent care. The risk of mortality for a patient who stayed in the hospital for more than 10 days would be 2.29 times higher than a patient who stayed for 2-3 days. The hazard ratio is 1.38 for patients who stayed in the hospital for 6-10 days against those who stayed for 2-3 days, and 1.44 for patients who stayed for 4-5 days. The risk of mortality for a patient with a low body temperature last measured at the hospital is 1.44 times higher than a patient with a normal last measured body temperature. There’s no significant difference in the risk of mortality between patients with a high last measured body temperature is and those with a normal body temperature, nor between the patients who did not have their body temperature measured and those with a normal body temperature. The risk of mortality of a patient who last stayed in the TSICU (Trauma Surgical Intensive Care Unit) unit would be 0.59 times lower than a patient who last stayed in the CCU (Coronary Care Unit) unit. There’s no significant difference in the risk of mortality between patients who last stayed in the CSRU (Cardiac Surgery Recovery Unit), MICU (Medical Intensive Care Unit), or SICU (Surgical Intensive Care Unit) and those who last stayed in the CCU unit.

**Conclusion**

The concordance index of the model is 0.829, which suggests that the variables explain a large of amount of variation in the patient mortality within 90 days of discharge from the hospital for those who are sent home. In other words, the variables have an impact on the 90-day mortality. The most impactful variables are age and comorbidities. The older the patients are, the higher the risk. Patients with cancer related disease have a significant higher risk of mortality. Gender, ethnicity, number of reports, and the vital sign variables except for the body temperature are not significantly impacting the risk of mortality. This information may help the doctors with their patient discharge decisions to reduce mortality rate among patients who are discharged home.

**Future Work**

In addition to the data investigated in this study, the MIMICIII database has many other health record data including medical notes from doctors and nurses. Previous studies show that information extracted from these notes using machine learning techniques help with predicting patient mortality [4]. Our next step is to analyze the notes using Topic Modeling algorithm and then incorporate the results into the Cox Proportional Hazards model and to improve the model results. The last step is to predict mortality using machine learning techniques. If the accuracy is decent, the model can be implemented and used in hospitals to help doctors to improve their patient discharge decisions.

A picture containing table

Description automatically generated

Table 1. Hazard ratio of each covariate to its baseline for patients discharged home

**Reference:**

1. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035,10.1038/sdata.2016.35 PubMed PMID: 27219127; PMCID: PMC4878278.
2. Rogliani, Paola; Sforza, Maurizia; Calzetta, Luigino. The impact of comorbidities on severe asthma, Current Opinion in Pulmonary Medicine: January 2020 - Volume 26 - Issue 1 - p 47-55.
3. Anthony J. Bleyer, Sri Vidya, Gregory B. Russell, Catherine M. Jones, Leon Sujata, Pirouz Daeihagh, Donald Hire. Longitudinal analysis of one million vital signs in patients in an academic medical center, ELSEVIER, Resuscitation. 2011; 82: 1387-1392, Resuscitation 82 (2011) 1387-1392.
4. McCoy TH, Castro VM, Cagan A, et al. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study. PLoS One. 2015; 10(8): e0136341.