DSC 424 – Advanced Data Analysis Final Project

Patient Survival Prediction Analysis

Shanele Youngquist, Mary Doerries, Cristian Casado, Can Dai, Michael Goss
DePaul University

Executive Summary

Getting an accurate understanding of a patient’s overall health has been particularly important during the COVID-19 pandemic as healthcare workers around the world have been flooded with patients that are diagnosed to be in critical condition. Oftentimes, Intensive Care Units (ICUs) do not have access to verified medical history for incoming patients and transferring medical records from outside facilities could take days. In addition, patients that are in distress, confused or unresponsive may be unable to provide their past history on chronic illnesses. By providing healthcare workers with knowledge on chronic conditions, it will better equip them to make clinical decisions and ultimately improve the odds of a patient’s survival.

Studies utilizing machine learning techniques to create ICU patient mortality prediction models found there are factors that are more helpful in predicting patient mortality than others depending on their disease group (Safaei, N. et al, 2022). A similar study found that Acute Physiology and Chronic Health Evaluation Score III (APACHE III) and the Logistic Organ Dysfunction Score (LODS) could be used to create ICU patient mortality prediction models. The goal of these models is to assist medical personnel judge the outcome of critically ill patients (Pang, K. et al, 2022). The National Report Card on the State of Emergency Medicine (Epstein, S. et al, 2009) evaluated the emergency care in the United States, giving them a C- in 2009.

To better understand these conditions, our research team focused on exploring six questions: (1) Are there certain attributes of hospital/ICU patients that have a higher risk of death? (2) Which sets of attributes capture the most information about the ICU patients? (3) What are the underlying trends in the attributes of hospital patients? (4) What is the relationship between vitals and demographic info (including patient mortality)? (5) What types of disease groups are more common to Hispanic patients? and (6) Are there certain groups of hospital patients that can be diagnosed in common disease groups?

Linear regression was used to verify if there were any hospital characteristics or patient attributes that were linked to a higher risk of death. Principal component analysis was used to identify the sets of attributes that capture the most information about the hospital patients. Factor analysis was applied to look for underlying trends in the attributes of the hospital patients. Canonical correlation analysis was utilized to explore the relationship between vitals and demographic information. Correspondence analysis was used to help understand which types of disease groups are more common to Hispanic patients. Lastly, clustering was employed to help identify if there were groups of hospital patients that were being diagnosed with common disease groups.

After completing some analysis, we found that particularly old age, a higher heart rate, low peripheral oxygen saturation levels, low blood pressure and/or a low core body temperature were correlated with patient deaths. However it is important to note that their inverse were not necessarily correlated or good predictors for patient survival, e.g. higher blood pressure did not have the same degree of correlation or significance in predicting patient survival. Three high level types of patients were identified: anemic patients, healthy patients and overweight patients.

Another analysis found that ethnicity and the diagnosis group are related and that the three most likely reasons Hispanic patients are admitted to the ICU are musculoskeletal, neurological, and respiratory issues. In contrast, trauma is the least likely of the reasons they are admitted within the given diagnosis group. This analysis also found that Native-American patients are more likely to be admitted to a hospital due to metabolic disorders than by any other reason.

Throughout this project, there were several limitations that potentially could have impacted our results. First, the research team identified a lot of missing information in the dataset. To combat this issue, any columns that were missing more than 50% of its data were removed. Second, the research team did not have the benefit of working with healthcare subject matter experts to provide insights on whether we should pay additional attention to specific aspects of the data. Furthermore, the amount of time to prepare the data, build working examples and interpret the results was very limited. If we had a lengthier timeframe, there would have been a greater opportunity to explore different theories related to the dataset and present additional examples. Lastly, there is a possibility that the Patient Survival Prediction dataset contained a lot of inconsistencies due to the wide time frame the data was originally collected – for example, it is unclear whether the data was collected within the first hour of a patient’s visit or 24 hours into their visit. On the other hand, some of our findings indicate that future work could be focused on specific patient attributes such as age, weight, BMI and blood pressure. This could be demonstrated using either numerical or categorical variables only and by building models that support a larger dataset and greater computing resources.

The research conducted also demonstrates that age and blood pressure are important predictors of survivability. It is good advice to attend routine medical checks as we grow older to proactively identify any situations. If possible, individuals should also place a high importance on improving their physical and mental health to reduce the risk of harmful blood pressure levels. In addition, more exposure to nature seems to be related to a healthier life. Surprisingly, we discovered that Native Americans appear to be considerably further apart from known disease groups such as cardiovascular issues. Further education should be done to understand the Native American lifestyle to see if any of their habits has an impact on their overall health. Although this research provides insight on how to prevent hospital deaths – in reality, the collective mortality rate at hospitals across the United States, Australia and New Zealand are relatively low. However, we recognize that any loss of life can be considered a tragedy.

Technical Paper

Abstract

As visits to emergency departments in hospitals increase and emergency services and medical staff continue to experience issues related to capacity and performance, it is important that we are able to understand patient profiles quickly to be able to potentially reduce the number of fatal outcomes for patients requiring advanced medical care. In this study, we leveraged several multivariate data analysis techniques including multiple linear regression, canonical correlation analysis, principal component analysis, factor analysis, cluster analysis, and correspondence analysis to understand several aspects related to patient survival. Through the implementation of these techniques we found evidence that age, heart rate, oxygen saturation, blood pressure and body temperature are variables bearing some of the most impact on predicting patient mortality. There is also evidence that there is differentiation on the way different diagnosis/disease groups affect different patient profiles. Our results compare positively with related research which has identified these same variables as significant in predicting patient mortality, reinforcing their importance as predictive variables. The mortality rate for patients admitted to hospitals collectively across the United States, Australia and New Zealand is low, at 9%. We recognize that even a 1% rate for loss of life can have devastating effects on individuals and/or families. Research results offer insight into actionable recommendations to focus proactive medical care in the monitoring of age and blood pressure as well as highlighting the benefits of living in cleaner, less polluted spaces as found in less populated or rural areas where proximity to nature is a given.

Keywords: *patient survival, multivariate analysis, advanced data analysis*

Introduction

According to the 2009 National Report Card on the State of Emergency Medicine, the health care system in the United States continues to be under heavy pressure (Epstein, S. et al, 2009). The recent COVID-19 pandemic has been a huge stressor, in addition to long term situations related to hospital capacity, cost of technology, medicine costs, insurance coverage and efficiency of emergency services. As visits to emergency departments in hospitals across the country increase, which is proven by the National Hospital Ambulatory Survey that was conducted in 2006, it displayed that hospital visits increased with an average of 32% more visits from 1996 to 2006. As more data is being collected each year, it is important that we are able to understand patient profiles and generate models that allow for volume predictions and potentially reduce the number of fatal outcomes. In this study, we are leveraging several multivariate data analysis techniques to understand several aspects related to patient survival in the United States, Australia and New Zealand.

Literature Review

Recent studies exploring patient health related data have utilized machine learning for predicting ICU mortality status upon patient discharge, using information available during the first 24 hours of admission. Different prediction models have been generated for patients affected by different disease groups. Based on their findings – age, heart rate, respiratory rate, blood urine nitrogen and creatinine levels are amongst the most critical features in mortality predictions (Safaei, N. et al, 2022). Other studies have proposed machine learning models to predict mortality risk of ICU patients based on characteristics of Acute Physiology and Chronic Health Evaluation Score III (APACHE III) and the Logistic Organ Dysfunction Score (LODS). Researchers have discovered that mortality risk of ICU patients can be better predicted using both of these tools and would be helpful in assisting medical personnel in judging the outcome of critically ill patients, especially those with a survival outcome that appears uncertain (Pang, K. et al, 2022).

There continues to be pressure on the emergency care system in the United States. The National Report Card on the State of Emergency Medicine (Epstein, S. et al, 2009) evaluated the emergency care environment state by state, providing a snapshot of the emergency care system in the United States. The report provides rankings in five different categories and provides a grade for the country overall and individually for each state. The categories assessed are access to emergency care, quality and patient safety environment, medical liability environment, public health and injury prevention and disaster preparedness. In 2009, the overall grade for the United States was a C-.

Methods

This study used patient survival prediction data available for download on the data science online community website Kaggle.com. The original source for the selected data is the eICU collaborative research datasets (Raffa, J. et al, 2019). The data was collected from more than 380,000 patients admitted to the ICU with a stay at least four hours long between 2014-2015 in 366 hospitals across the United States, Australia and New Zealand. Out of the data that was collected, the usable number of patient information was reduced to 91,713 samples and compared hospital and patient characteristics. Our research team's pre-processing techniques reduced this number again to 34,776 samples after removing columns that had 50% or more missing data.

Multivariate analysis was performed to solve six research questions. Linear regression was applied to verify if there were certain attributes of hospital/ICU patients that had a higher risk of death. Principal component analysis was used to determine which sets of attributes captured the most information about the ICU patients. Factor analysis was applied to uncover underlying trends in the attributes of hospital patients. Canonical correlation analysis was employed to explore the potential relationship between vitals and demographic info and correspondence analysis was carried out to understand which types of disease groups were more common to Hispanic patients and clustering was run to identify if there were groups of patients being diagnosed with common disease groups. Canonical correlation analysis was appropriate for the evaluation of relationships between patient characteristics. The dimensionality of patient characteristics was evaluated by exploring the results of principal component and factor analysis via a factor solution including item-factor correlations and levels of variance explained by the extracted factors. Clustering served as a method that could potentially surface previously unknown associations or data qualities for multiple patients' profiles. For exploring the relationship between specific pairs of patients' characteristics, a correspondence analysis was carried out. Regression analysis was appropriate for the goal of predicting a deadly outcome for a given patient.

Discussions and Results

Model building and multiple linear regression was used to explore the first research question: Are there certain attributes of hospital/ICU patients that have a higher risk of death? In addition to data cleaning shown in *Appendix I - R Code to clean Patient Survival Dataset*, backwards selection was used to remove variables and attempt to reduce multicollinearity. In addition, 5-fold cross-validation was also completed in the beginning to give an idea on the output of the test. The mean square error of our 5-fold test was 0.35. The model predictions are not as close to the observations, therefore the mean square error was larger than anticipated. Initially, a standard linear model was built using the original data without any stepwise regression so a comparison could be made after backwards selection was completed. Multicollinearity was very high at this point so additional models were created to reduce this. Once multicollinearity was removed completely, variables that had high p-values in their t-tests were removed to see if the model could be improved to strengthen its reliability.

By focusing on variables with p-values in their t-tests that are the closest to zero, we discovered five patient variables and two hospital characteristics that have significant impact on the results. Patients within the first 24 hours displaying a high heart rate, low oxygen levels, low blood pressure, and low body temperature in addition to being older in age, have an increased risk of dying in the hospital. The two APACHE predictions are also helpful because it scores a probabilistic prediction of mortality for the patient by using APACHE III scores and other covariates including diagnosis. The higher the percentage levels predicted, the higher the chance of patient death as well.

In addition, Ridge regression and LASSO regression were performed and compared against the final iteration of the standard linear model. In the Ridge regression model, none of the variables were removed and the variability in hospital death is roughly 21.17% explained by the model. Furthermore, our lambda value is 0.01056302, which explains a little bit of bias. Because all the variables were used in this model, there are some variables with p-values in the t-test that show significance, which may not have appeared in the prior model. However, if we look at the variety of plots in the model, there are clear patterns of the values being too fitted and some variables should be removed.

Lastly, LASSO regression was performed on the data*,* which removed 18 variables. The lambda value is 0.0002497616 which appears to be a lot more optimal than Ridge regression due to its smaller value. (*Figure 1)* Although some variables were removed in this model, roughly 21.2% of the variability in hospital death is explained, which is similar to Ridge regression. Because it is a reduced model, the p-values for some of the t-tests have improved and are closer to zero which deem these variables as important in predicting patient death. Plus the significance codes have increased slightly as well which improve the overall strength of the model. However, if we look at the variety of plots in the model, there are clear patterns of the values being too fitted and some variables should still be removed.

Similar to the simple linear regression model, both Ridge and LASSO regression confirm that any patients that are older and have symptoms of high heart rate, low oxygen, low blood pressure and/or have a very low core body temperature have an increased risk of dying in the hospital, however there are other variables that display significance as well.

Principal Component Analysis (PCA) was the multivariate method that was used to explore the second research question: Which sets of attributes capture the most information about the ICU patients? Using the KMO test, Bartlett's Test of Sphericity, and Cronbach's reliability test, we can see that the data is suitable for PCA since the variables are closely related. The MSE from the KMO test was 0.82, the p-value from Bartlett's Test of Sphericity is close to zero, and Cronbach's alpha is 0.86. The method used to determine the number of components to use for PCA was the knee method with a scree plot. Using this method, the optimal number of components to use is somewhere between three and six. After experimenting using these different values to make sure that there were no cross-loadings or variables that did not contribute enough to the components, we decided to use four components for the analysis. In order to get the component loadings and discover which variables contributed the most to the four components, the varimax rotation method was utilized to rotate the four components.

The component loadings, which can be found in *Appendix VI - PCA Component Loadings*, show that the variables contributing to component 1 are related to the blood pressure readings after one day or one hour in the ICU. This component can be called "Blood Pressure", and the total variance explained by this component is 33.35%. The variables contributing to component 2 are related to the blood urea nitrogen (BUN) concentration and creatinine concentration in the patient's serum or plasma after one day in the ICU. This component can be called "BUN and Creatinine", and the total variance explained by this component is 14.26%. The variables contributing the most to component 3 are related to the patient's heart rate during the first hour or the first day in the ICU. This component can be called "Heart Rate", and the total variance explained by this component is 10.14%. The variables contributing to component 4 are related to the patient's red blood cell count during the first day in the ICU. This component can be called "Red Blood Cells", and the total variance explained by this component is 9.19%. After uncovering the sets of variables that contribute the most to the principal components that explain the highest percentage of variance in the data, we are capable of answering the research question. The sets of attributes that capture the most information about the ICU patients are Blood Pressure, BUN and Creatinine, Heart Rate, and Red Blood Cells.

Factor Analysis (FA) method was chosen to explore the third research question: What are the underlying trends in the attributes of hospital patients? Factor Analysis is commonly used to reduce the number of variables to a smaller number of factors, and to concisely describe the relationship among variables. Therefore, it was a suitable methodology to understand the underlying trends in the attributes of the data. Before conducting the FA, three distinct sets of tests were conducted in order to measure the reliability and factorability of the clean dataset (see *Appendix I - R Code to clean Patient Survival Dataset*). The results show a KMO value of 0.81, the Bartlett's Test of Sphericity with a p-value less than 2.22e-16 and a Cronbach's Alpha test with a raw alpha value of 0.8199. These values indicate that the data is reliable and stable.

To determine the number of factors to use for the FA analysis, it was decided to use a scree plot and utilize the knee method for the best number of factors. As a result, it was decided to use five factors since there is an elbow between five and six factors, and the use of five factors allows us to grasp a bigger variance of the data. The resulting Factor Analysis loadings for each factor can be viewed in *Appendix IV – Factor Loadings Figure*. The results indicate that the cumulative variance of using five factors is 0.424, and the root mean square of the residuals (RMSA) is 0.08.

Variable loadings for each respecting factor which are bigger than 0.4 can be viewed in *Appendix IV – Factor Loadings Figure*. Using factor loadings plots, we can describe the factors accordingly: Factor 1 loadings indicate that this factor is about a patient's highest blood pressure during their unit stay (during their first hour or during their first 24 hours). Factor 1 can be called "maximum blood pressure". Factor 2 loadings show that this factor is about a patient's blood urea nitrogen, creatinine, potassium concentrations. Studies suggest that urea and creatinine are nitrogenous end products of metabolism (Hosten, et al, 1990). Therefore, this factor can be called "Metabolism end products". Factor 3 loadings indicate that this factor is about a patient's heart rate, respiration and temperature during their first hour or first 24 hours of the stay – therefore, this factor can be called "vitals". Factor 4 loadings indicate that this factor is about a patient's red blood cells proportion and hemoglobin levels during their first 24 hours of stay. Hemoglobin levels measure the red blood cell production in the blood, therefore this factor can be called "hemoglobin levels". Factor 5 loadings are almost the opposite of Factor 1 and measure a patient's lowest blood pressure during their unit stay (during their first hour or during their first 24 hours). Using this information, we can call Factor 5: "minimum blood pressure".

Coming back to the research of question of what the underlying trends in the variables of the Patient Survival Prediction dataset, we can conclude that maximum blood pressure (Factor 1), metabolism end products (Factor 2), vitals (Factor 3), hemoglobin levels (Factor 4), and minimum blood pressure (Factor 5) are the underlying factors that play a crucial role to diagnose a patient in the hospital.

Looking at the results of applying PCA and FA to the data, we can compare the underlying factors discovered from using both methods. The first underlying factor from using FA was Maximum Blood Pressure, while the first underlying factor from using PCA was Blood Pressure. While both underlying factors are related to blood pressure, the underlying factor from using FA has the maximum blood pressure variables contributing to it. The second underlying factor from using FA was Metabolism End Products, while the second underlying factor from using PCA was BUN and Creatinine. Both underlying factors have blood urea nitrogen and creatinine concentration variables contributing to them, but only Metabolism End Products has a potassium concentration contributing to it. The third underlying factor from using FA was vitals, while the third underlying factor from using PCA was Heart Rate. Both underlying factors have the heart rate variables contributing to them the most. The fourth underlying factor from using FA was Hemoglobin Levels, while the fourth underlying factor from using PCA was Red Blood Cells. These underlying factors have the exact same variables contributing to them. Since only four components were used for PCA, the fifth underlying FA factor cannot be compared to anything.

Canonical correlation analysis was used to investigate the fourth research question: "What is the relationship between vitals and demographic info?" Some variables used in this analysis were not normal so the Bartlett's Chi-Squared Test was used instead of the Wilks Lambda test because it is a little more robust and less sensitive to outliers and skewed data. Bartlett's Chi-Squared test identified the first five variables as significant. This means that there is nearly 99% of overlapping variance (Figure 2) between vitals and the demographic variables and that vitals are a very good indicator of the numeric demographic variables in the data set.

The strongest positive correlations from vitals are low blood pressure measured invasively and non-invasively. This supports our research since three of the demographic variables are age, weight, and BMI which are all known contributors to blood pressure. The strongest negative correlation from the demographic variables is age. The redundancy coefficients show us that 5% of the demographic variables can be explained by the vitals variables and 3% of the vitals variables can be explained by the demographic variables.

For correspondence analysis, the process focused on the following two categorical variables: ethnicity and diagnosis group. The key research question for this analysis was to find out which types of diagnosis/diseases are more common in Hispanic patients. The Chi-Squared statistic for testing the independence of ethnicity and the diagnosis group was 451.18 with 50 degrees of freedom, with an associated p-value < 0.00000000000000022. This result suggests that ethnicity and the diagnosis group are related.

When reviewing the model's summary for the patients, we see that most of the eigenvalue variance is explained by the first two dimensions. Of the total inertia, the first component accounts for 59.23% of the inertia (Proportion = 0.5923) and the second component accounts for 24.74% of the inertia (Proportion = 0.2474). Together, these two components account for 84% of the total inertia (Cumulative = 0.8397). This suggests using two components for the correspondence analysis should be sufficient.

Exploration of the graphical representation of ethnicity versus the diagnosis group suggest that at hospital admission time, Hispanic patients are more associated with the following diagnosis groups: respiratory, neurological, musculoskeletal, genitourinary, hematological, and gynecological. Also, at hospital admission time, Hispanic patients are less likely to be admitted because of trauma, with the top 3 diagnosis/disease groups for Hispanics being: musculoskeletal, neurological, and respiratory (Figure 3). Finally, Native-American patients are more likely to be admitted to a hospital due to metabolic disorders than by any other reason.

The key research question for clustering analysis was to find out if there were different patient groups exhibiting common medical characteristics. A K-means clustering algorithm was applied to the patients' sample, looking for a 3 cluster solution (figure 4). The clusters were labeled as "Anemic", "Overweight", and "Healthy" patient groups (Table 1). "Anemic" patient cluster: represents 32% of the patients, and was named by a negative strong association with blood pressure and vitals. "Overweight" patient cluster: as the largest cluster it represents 39% of the patients, and was named due to a strong association with higher blood pressure and hemoglobin levels. "Healthy" patient cluster: represents 29% of the patients, and was named by a strong association with lower blood pressure and stronger vitals.

Limitations & Future Work

A large amount of missing data was identified in the Patient Survival Prediction data. To solve this problem, listwise deletion was executed and potentially relevant independent variables had to be removed from the data. The study could have benefited from a subject matter expert meaning that current study design choices were probably not optimal. In this case, we didn't have the opportunity to engage with a domain expert as we worked through the analysis.

Most of the techniques leveraged by the team in the realization of the study required the use of numerical variables. Because of this, categorical variables had to be removed from the data and it is possible that potentially representative interactions between continuous and categorical variables could have been missed. In addition, available time for overall preparation, execution and interpretation of the study was in short supply due to a shortened class schedule. With more time, we could have extended our analysis in different directions. Finally, there is a possibility that the data contained potentially inconsistent data due to variance in terms of the original data collection timing. Each event had a reference to collection at the first hour and then the first day, but there is no clarity as to what the exact moment was when each measurement was taken.

Some of our findings highlight interesting areas of focus for future research that could be explored in subsequent studies; one of the most critical ones is the recommendation to prioritize further studies which target age, weight, BMI and blood pressure as critical predictor variables for overall health and life span. In terms of the data types, our work targeted either numerical or categorical variables. There is pending work required to incorporate some interactions between different types of variables into analysis based on similar patient health data related datasets. A necessary future exploration to extend in the multivariate analysis presented in this study, would be to explore additional types of clustering techniques due to the large dataset size requiring more computing resources to run different amounts of trials and tests.

Conclusion

Age and blood pressure are important predictors of survivability. It is good advice to attend routine medical checks as we grow older to proactively identify any potentially adverse situations. In addition, more exposure to nature seems to be related to a healthier life. Native Americans are less likely to be diagnosed with known disease groups such as cardiovascular issues. The mortality rate of hospitals studied within the United States, Australia and New Zealand are relatively low at 9%, when looking at overall percentages. We recognize even a 1% rate for loss of life could be a dramatic event depending on the scale.

References

Anzum, S. (2022). Patient Survival Prediction Dataset. https://www.kaggle.com/datasets/sadiaanzum/patient-survival-prediction-dataset?select=Dataset.csv

Benediktsson, S., Hansen, C., Frigyesi, A., & Kander, T. (2020). Coagulation tests on admission correlate with mortality and morbidity in general ICU patients: An observational study. Acta Anaesthesiologica Scandinavica, 64(5), 628–634. https://doi-org.ezproxy.depaul.edu/10.1111/aas.13545

Boyle, J., Jessup, M., Crilly, J., Green, D., Lind, J., Wallis, M., Miller, P., & Fitzgerald, G. (2012). Predicting emergency department admissions. Emergency medicine journal : EMJ, 29(5), 358–365. https://doi.org/10.1136/emj.2010.103531

Epstein, S. K., Burstein, J. L., Case, R. B., Gardner, A. F., Herman, S. H., Hirshon, J. M., Jermyn, J. W., McKay, M. P., Mitchiner, J. C., Sullivan, W. P., Wagner, M. J., Beer, S., Tiberi, L., Price, C., Cunningham, R., Wilkerson, D., Bromley, M., Geist, M., Gore, L., Singh, C. A., … Schwalberg, R. H. (2009). The National Report Card on the State of Emergency Medicine: evaluating the emergency care environment state by state 2009 edition. *Annals of emergency medicine*, *53*(1), 4–148. https://doi.org/10.1016/j.annemergmed.2008.10.028

Hosten AO. BUN and Creatinine. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. Boston: Butterworths; 1990. Chapter 193. Available from: https://www.ncbi.nlm.nih.gov/books/NBK305/

Pang, K., Li, L., Ouyang, W., Liu, X., & Tang, Y. (2022). Establishment of ICU Mortality Risk Prediction Models with Machine Learning Algorithm Using MIMIC-IV Database. Diagnostics (2075-4418), 12(5), N.PAG. https://doi-org.ezproxy.depaul.edu/10.3390/diagnostics12051068

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raffa, Jesse1; Johnson, Alistair1; Celi, Leo Anthony2,,1; Pollard, Tom1; Pilcher, David3; Badawi, Omar4 33: THE GLOBAL OPEN SOURCE SEVERITY OF ILLNESS SCORE (GOSSIS), Critical Care Medicine: January 2019 - Volume 47 - Issue 1 - p 17. doi: 10.1097/01.ccm.0000550825.30295.dd

Safaei, N., Safaei, B., Seyedekrami, S., Talafidaryani, M., Masoud, A., Wang, S., Li, Q., & Moqri, M. (2022). E-CatBoost: An efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database. PLoS ONE, 17(5), 1–33. https://doi-org.ezproxy.depaul.edu/10.1371/journal.pone.0262895

Figures

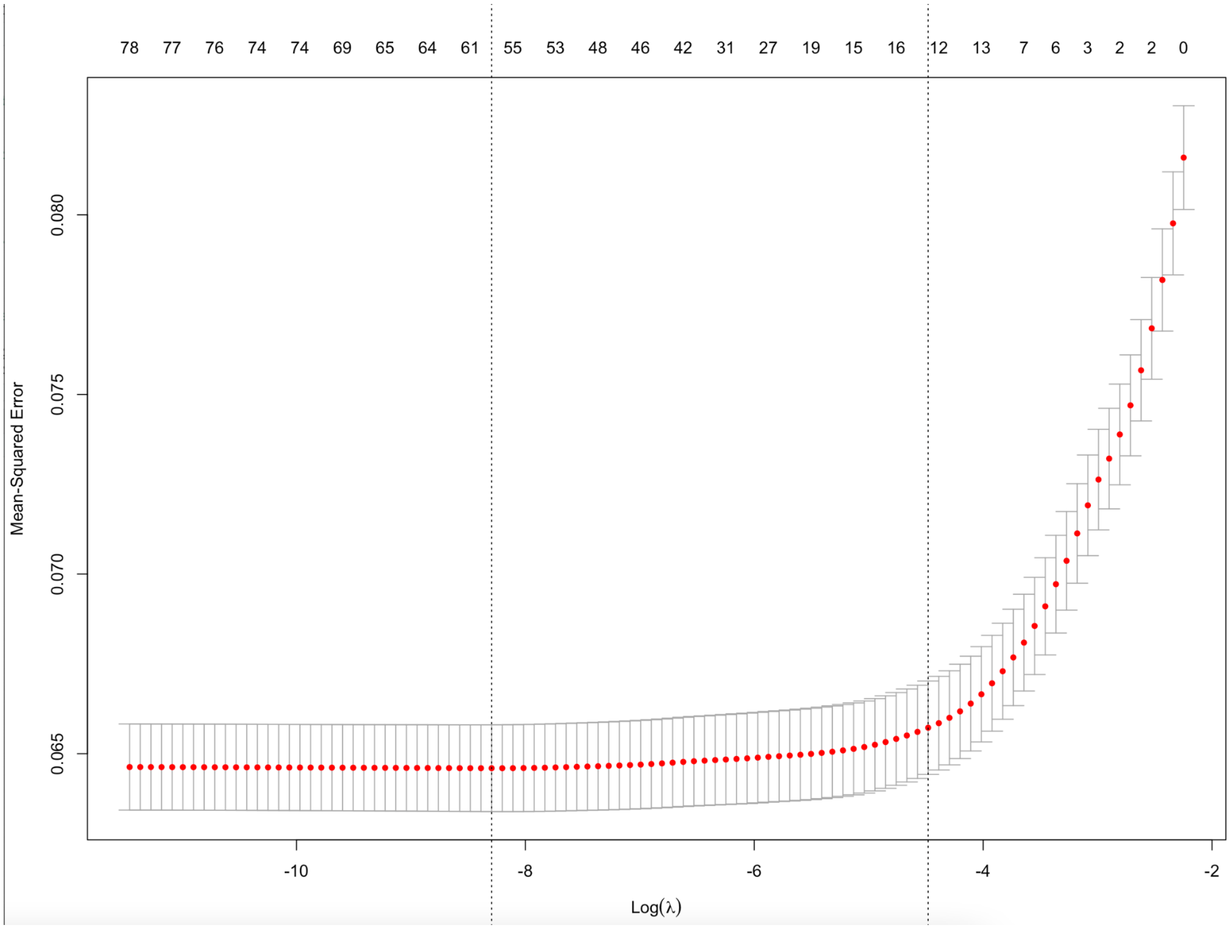

Figure 1. LASSO Regression Plot.

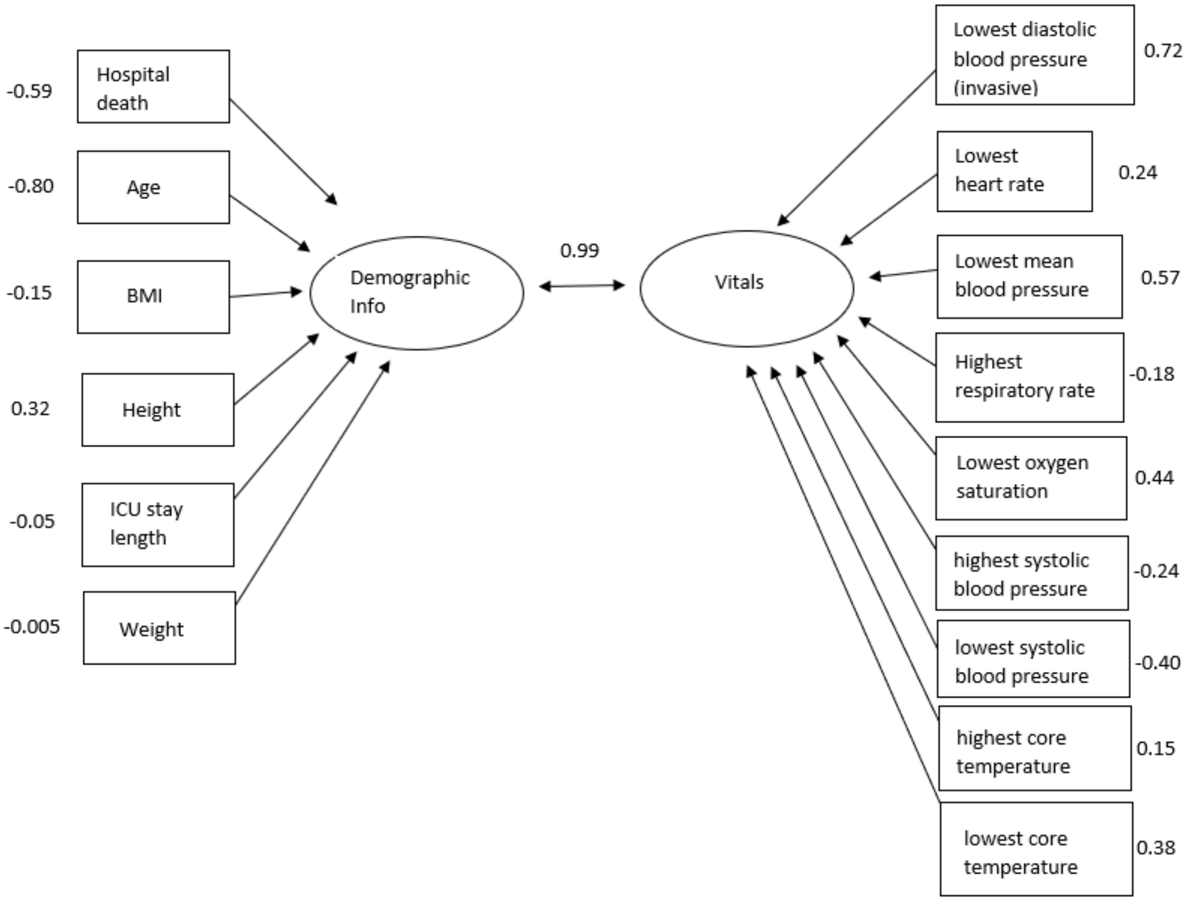

Figure 2. Canonical Correlation Analysis visualization

* note that not all of the vitals variables are present for space spacing saving purposes.

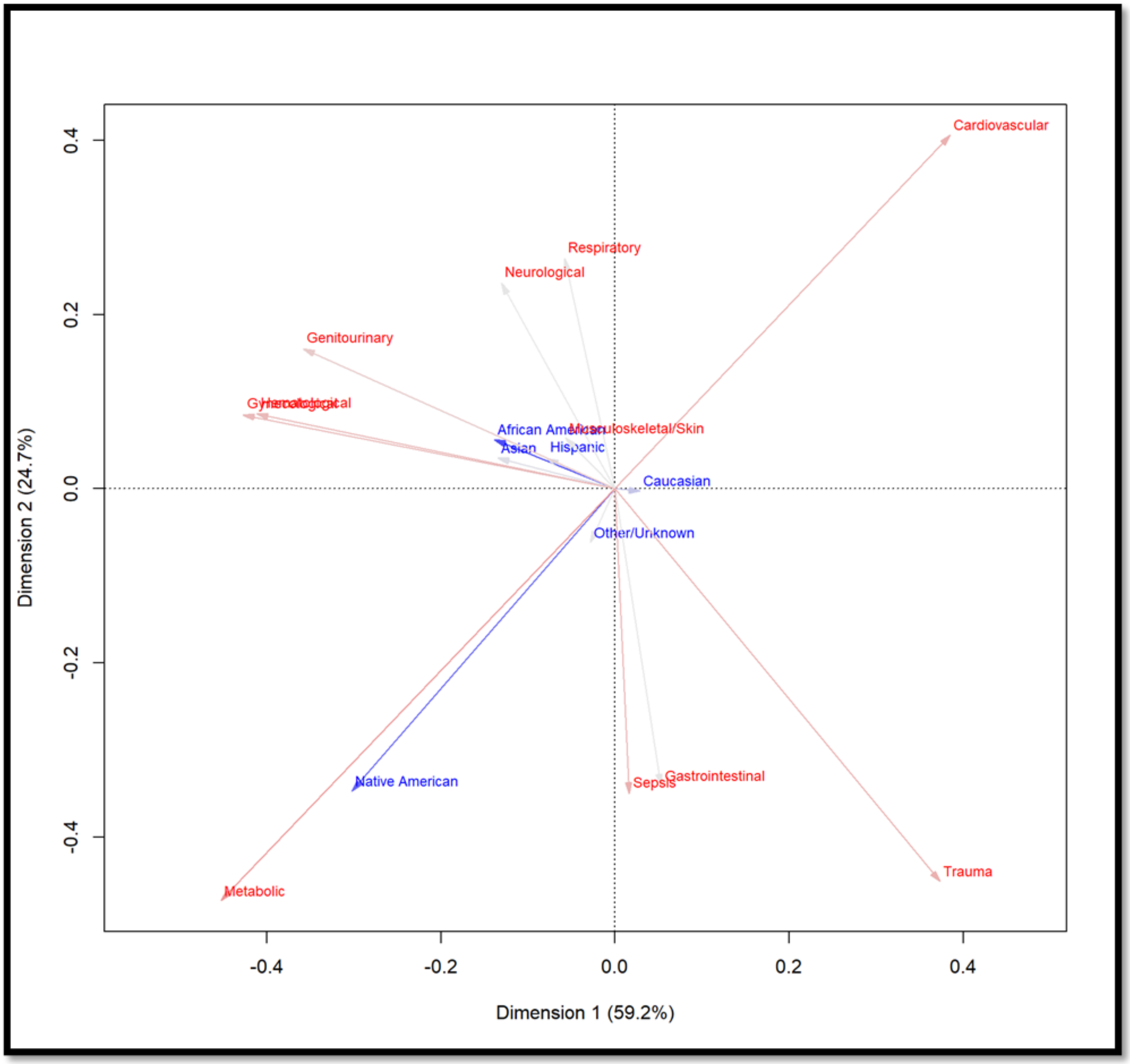

Figure 3. Correspondence analysis solution for Hispanic patients (with arrows)

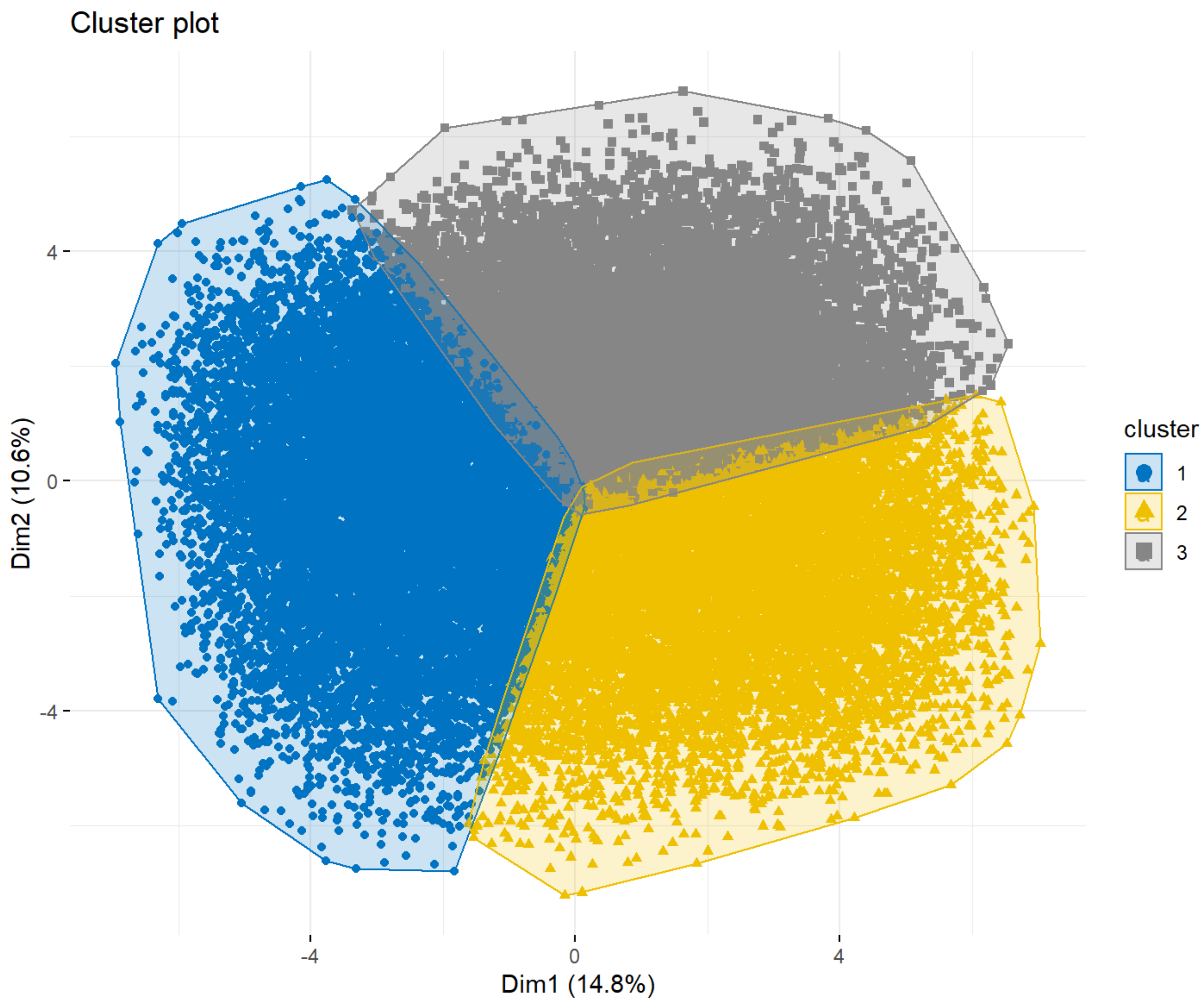

Figure 4. K-means Clustering model.

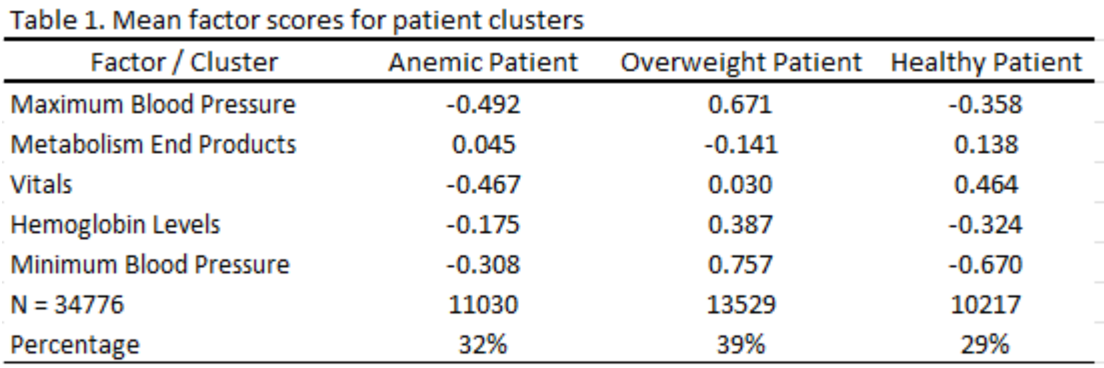

Table 1. Mean factor scores for patient clusters

Factor / Cluster	Anemic Patient	Overweight Patient	Healthy Patient
Maximum Blood Pressure	-0.492	0.671	-0.358
Metabolism End Products	0.045	-0.141	0.138
Vitals	-0.467	0.030	0.464
Hemoglobin Levels	-0.175	0.387	-0.324
Minimum Blood Pressure	-0.308	0.757	-0.670
N = 34776	11030	13529	10217
Percentage	32%	39%	29%