ADS1001 Project: Predicting patient outcomes in ICU
Adam Choong, Yilin Han, Simon Sun, Farrel Wiharso & Yija Xue.

Group Contributions

The project team consisted of 5 individuals, Adam Choong, Simon Sun, Yilin Han, Farrel Wiharso, Yijia Xue. Adam Choong was primarily involved in the coding aspect of this project but also assisted in the drafting of this project's report, particularly the component regarding the data manipulation and modeling process. Simon Sun was primarily responsible for the research component of this project but also assisted in the coding aspect of this project by providing useful insights so that meaningful conclusions could be made about the results produced by the data. Yilin Han was heavily involved in the construction of the presentation for this project. Farrel Wiharso was also responsible for the construction of this project's presentation and also provided insights that were used in the report, specifically the background information. Yijia Xue collaborated with Adam Choong in the coding process and reviewed the data manipulation process to confirm the accuracy of the methods involved.

Background Information

In hospitals, Intensive Care Units (ICUs) are specialist hospital wards which provide treatment to critically ill patients who require serious medical care. Healthcare staff working within the ICU are multi-disciplinary, composed of experienced nurses, doctors, and specialists well-trained in providing critical care for patients who face life-threatening health circumstances. (Bergmeir, 2020). In order to create advancements in the medical field, healthcare providers frequently report on the quality of patient care in ICUs, aiming to evaluate the effectiveness of various types of medications and treatments (Bergmeir, 2020). Throughout recent years, data science has played an increasingly significant role in the healthcare sector. Hospitals and medical institutions are heavily reliant on a data-driven approach in order to evaluate a patient's condition while undergoing treatment (*The Vital Role of Data Science in Advancing Medicine*, 2019). A notable example of this is the PhysioNet/Computing in Cardiology Challenge held in 2012 (Goldberger, . This project utilizes the data collected from PhysioNet, where over 12,000 unique ICU stays were recorded, with the data analyzing the patient's first 48 hours in the ICU. In this experiment, 42 variables are considered, further highlighting the intensive nature of care given in an ICU. The variables are all quantitative, where 6 descriptors are utilized to identify each patient, such as their physical attributes, while the other variables are recorded in intervals throughout 48 hours.

RecordID	Age	Gender	Height	ICUType2	ICUType3	ICUType4	Mean_Weight.x	Mean_GCS.x	Mean_HR.x
132543	68	1	180.3	0	1	0	84.6	14.88888889	72.97142857
132545	88	0	169.787227	0	1	0	83.05413594	15	79.52
132547	64	1	180.3	0	0	0	114	8.333333333	81.31818182
132551	78	0	162.6	0	1	0	48.4	13.25	78.125
132554	64	0	169.787227	0	1	0	60.7	15	129.3636364
132573	77	1	162.6	0	0	0	92.89259259	15	74
132577	65	1	169.787227	0	1	0	66.57428571	10.6	84.73913043
132582	84	1	182.9	0	1	0	82.5	14.71428571	94.63636364
132584	78	0	169.787227	0	1	0	72.8	8.833333333	84.5
132588	48	0	154.9	0	1	0	42.3	15	103.85
132590	58	1	188	1	0	0	105.8461538	9.5	98.13888889
132591	81	1	169.787227	0	1	0	63.7	15	73.18518519
132597	66	0	137.2	0	1	0	82	15	68.23333333
132598	80	0	169.787227	0	0	1	60	7	72.41463415
132602	80	1	180.3	0	1	0	70	15	77.69230769
132605	90	0	169.787227	0	1	0	55	7.5	67.53125
132610	72	1	172.9	0	1	0	72.26	15	74.34615385
132614	77	1	162.6	0	0	0	59	15	70.86363636
132615	46	0	152.4	0	1	0	88.16551724	6.833333333	74.20930233
132617	77	1	170.2	0	0	0	75.0962963	15	76.21052632
132622	71	0	160	0	1	0	79	15	73.03571429

The image displayed above represents the preprocessed version of the ICU dataset. As demonstrated, the data is sorted based on the patient's recorded ID number, and alongside are the patient's descriptors, such as their age, gender, and height. Additionally, it classifies each patient into the ICU section they have been allocated in, where the values 1 to 4 indicate the Coronary Care Unit, Cardiac Surgery Recovery Unit, Medical ICU, and Surgical ICU respectively.

Mean_HR.x	Min_HR.x	Max_HR.x	Mean_HR.y	Min_HR.y	Max_HR.y
72.97142857	57	88	68.2	57	79
79.52	67	94	70	65	79
81.31818182	71	91	94.88	90	101
78.125	62	111	65.34146341	55	79
129.3636364	120	137	125.2916667	115	133

In.hospital_death
0
0
0
1
0
0
0

Furthermore, the 36 variables analyzed during the patients treatment are evaluated through their mean, minimum, and maximum values. In the dataset, these values are recorded chronologically, where the x and y classifiers attached to each variable indicate data collected during the first and second day respectively. The final column of the dataset is the in.hospital_death variable, which indicates the patient's status after 48 hours in the ICU. A value of 0 indicates the patient's survival, while 1 indicates an in-hospital death. It is important that this is presented into a dummy variable, as the data can be easily processed for analysis in further parts of the investigation.

Considering the background information and context of the dataset is crucial in order to explore a viable research topic. After careful consideration, the investigation has been centered on analyzing arterial related exposures among ICU patients. This leads to the construction of the research question, where it aims to explore the question: "Which arterial related variables are most correlated to death in an ICU?", with a focus on chemical-related variables. To further elaborate on the research question, arterial variables refer to the condition of the patient's arteries, which are the blood vessels which deliver blood from the heart to the body's tissues.' This research question has been selected as the exploration is neither too general or narrow, and can be answered accurately with the given data set. Arterial variables refer to the condition of the patient's arteries, which are the blood vessels which deliver blood from the heart to the body's tissues.' In the dataset, these variables are denoted with ABP, and are measured in the amount of millimeters of mercury present within the patient's body.

Preprocessing Data

The data was preprocessed into a csv file before it was provided for use in Python. However, it is important to note that initially, when the Physionet challenge commenced, the data was raw and took the form of several html files that were not compressed into a singular spreadsheet. The data was preprocessed in such a way that the variables that fluctuated, such as arterial blood pressure, were split and labeled with an 'x' or 'y', to indicate which day the readings were taken, since the dataset was a compilation of the maximum, minimum and mean readings of each patient, taken independently in day one and day two. Furthermore, when the data was compiled into a csv file, several suspected pieces of empty data were detected and subsequently disguised. For example, several patients had the exact same height, which could easily be labeled as unusual since the accuracy of the height recorded was higher than other patients that actually had recorded heights. The specific case in question can be seen when several patients had a height of exactly 169.787227 centimeters, despite their differing ages and genders. This may have implied that their heights were not recorded during their stay at their respective ICU wards because the median value of the height was 169.787227 centimeters.
As a result of substituting median values for missing data, all variables could be analysed, albeit not as accurately as one might have desired to. However, the presence of these median values, which could have been interpreted as empty values, aided in the analysis of the data because it assisted in narrowing some of the variables that should be selected for exploration. Furthermore, the choices that were made, regarding the selection of variables to analyse, was also influenced by the preprocessing of data, to some extent, as some variables were unsuited for analysis, not just based on their property of having disguised missing values, but also because of their abundance of outliers, which were later found in Python.

Processing and Manipulation of Data in Python

Using the pandas module, the data was filtered to show only the data of the patients who were admitted to the coronary ICU since this specific type of unit specialised in treating arterial related issues. Furthermore, the variables analysed were selected based on the theme of being chemically based. For example, the project disregarded the readings of systolic arterial blood pressure in favour of other variables such as glucose levels because although systolic arterial blood pressure would have aided in the investigation, it would have further complicated the analysis because it did not prescribe to the theme of being chemically related due to the fact that it would have been considered a physical factor. A physical factor was

defined in this project as being a reading that did not yield some kind of quantity of a chemical's presence in the body of a patient. Furthermore, in principle, adding more variables to study would have led to more complications when trying to reach a conclusion.

Further steps that were taken in the manipulation of data included changing the filtered data set so that it only displayed the mean values of each chemical reading over the two days. The minimum and maximum values were disregarded because there were still some missing values that were disguised, using the method mentioned above, which would have made these values irrelevant.

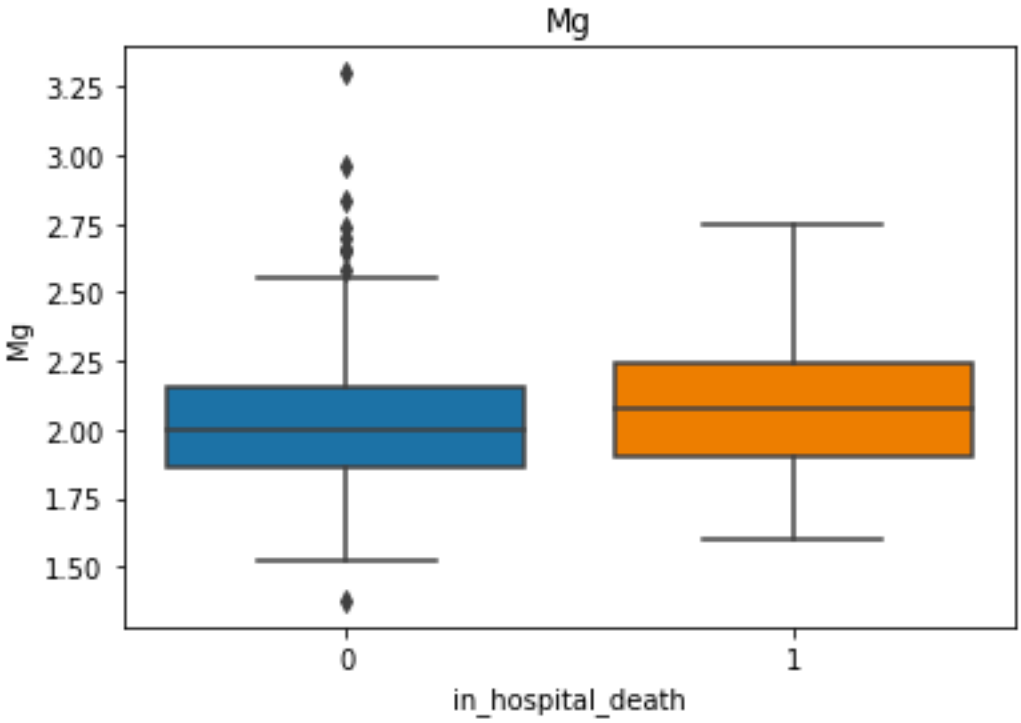
Before further manipulation occurred, a rough model was generated, in the form of a box plot, because a visual communication of the outliers and properties of each variable was required for the purposes of comprehending the data set more easily before 3 main variables were selected for in-depth analysis. In the process of generating two boxplots for each variable, one showing the interquartile range and median values for the patients who survived and the other conversely showing the same statistical properties for the patients who passed in the same ward, variables were discounted on the premise that if the difference between the median values and interquartile ranges shown in each of their plots was subjectively small, then they would not be considered. External research factors also affected the choices to discount certain variables. Conversely, variables whose median values and interquartile ranges had a sufficiently significant difference were considered for analysis. In addition, it was decided that the median values of ideal candidate variables would have to lie in the interquartile range because it would be easier to show the distinctions between the two boxplots generated for each variable.

When the variables were ultimately selected for analysis, their outliers were not discounted because of the relatively high quantity of outliers present which would have heavily affected the accuracy of the conclusions made. Furthermore, it was suggested that the difference between the mean values of each variable be used to quantify the suitability of each variable for analysis.

Data Modeling

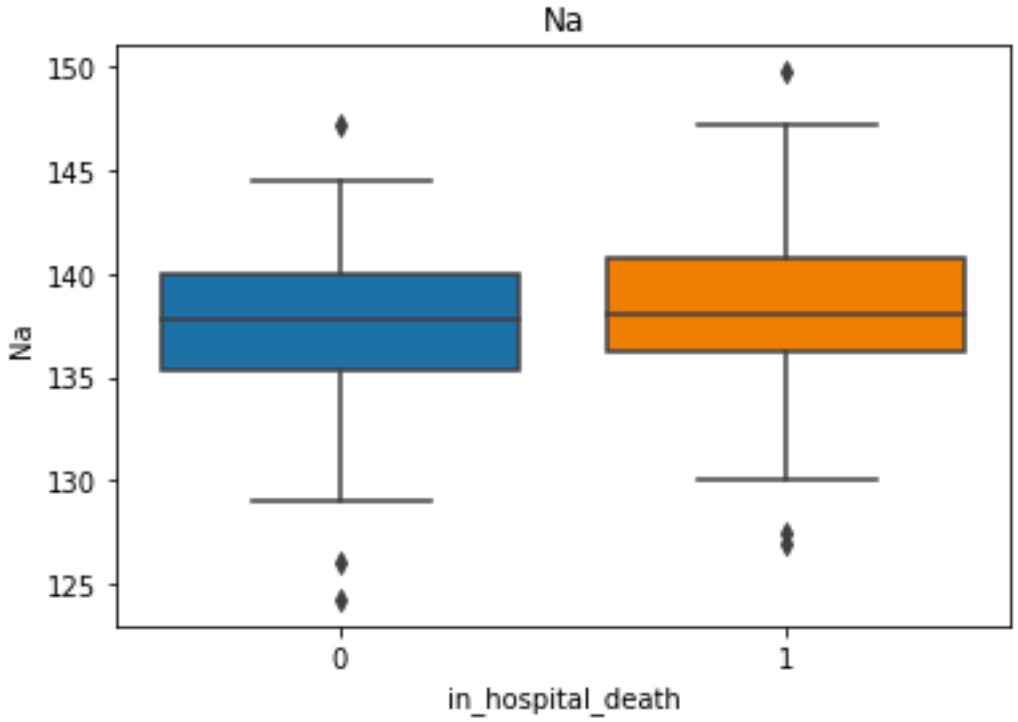
The main techniques used were the generation of boxplots and heatmaps from the seaborn module as well as the use of data frame manipulation techniques under the pandas module. In addition, logistic regression techniques were explored. Boxplots were used to detect outliers easily whilst heatmaps quantified correlation and logistic regression models assisted in generating a predictive model.

As mentioned, the choice to include outliers in the models was made because of the significant impact they had on the statistical properties of each variable. This could be credited to the relatively high abundance of outliers present. For example, in one of the models shown, there is a sufficiently significant presence of outliers, which made discounting them impractical.



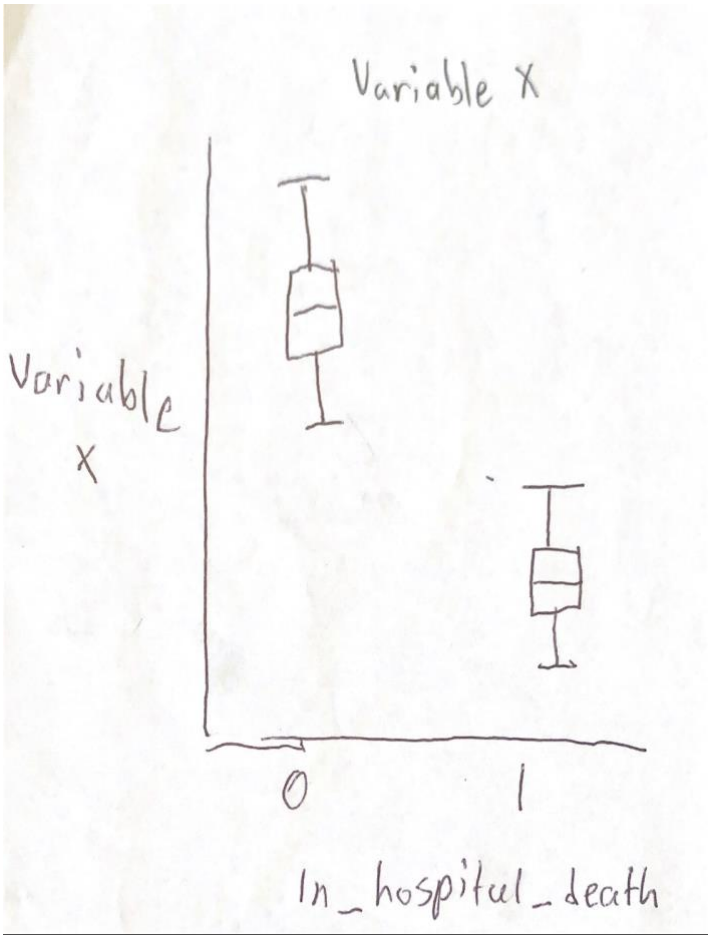
It can be seen that on the left boxplot, there is a significant presence of outliers which would lessen the accuracy of the results. Although outliers were included in the analysis of results, they were considered with discretion. In the case concerning Lactate's correlation with in-hospital death, which is mentioned later, for example, a sufficiently significant proportion of results consisted of outliers, thus the variable was discounted. The boxplot method of modeling has some shortcomings. For example, by using boxplots, one would have to select the appropriate variables to analyse, and correlation between variables and in hospital death could not be quantified objectively. Instead, the use of boxplots to determine correlation between the chosen variable and in hospital death would be mostly subjective, leading to inaccuracies. Furthermore, the decision to include outliers in all results was ade so consistency would be maintained.

Another important comment to be made on the box plot shown is that there is a noticeable, relative difference between the medians of the box plots, which is favourable because it would allow for the conclusion that higher levels of Mg would lead to in hospital death, to be validated. The use of a box plot assisted in determining the relative differences between the medians of each variable as it took into account the different measurement systems of each variable and plotted them in such a way that irregardless of the interval of the measurement system being used for each variable, visual distinctions could easily be made with regards to the impact of each variable of in hospital death. For context, in another box plot shown below, it can be seen that unlike the Mg box plot, it lacks a significant impact on the outcome of patients in ICU despite the difference between medians and means being higher due to the measurement scale.



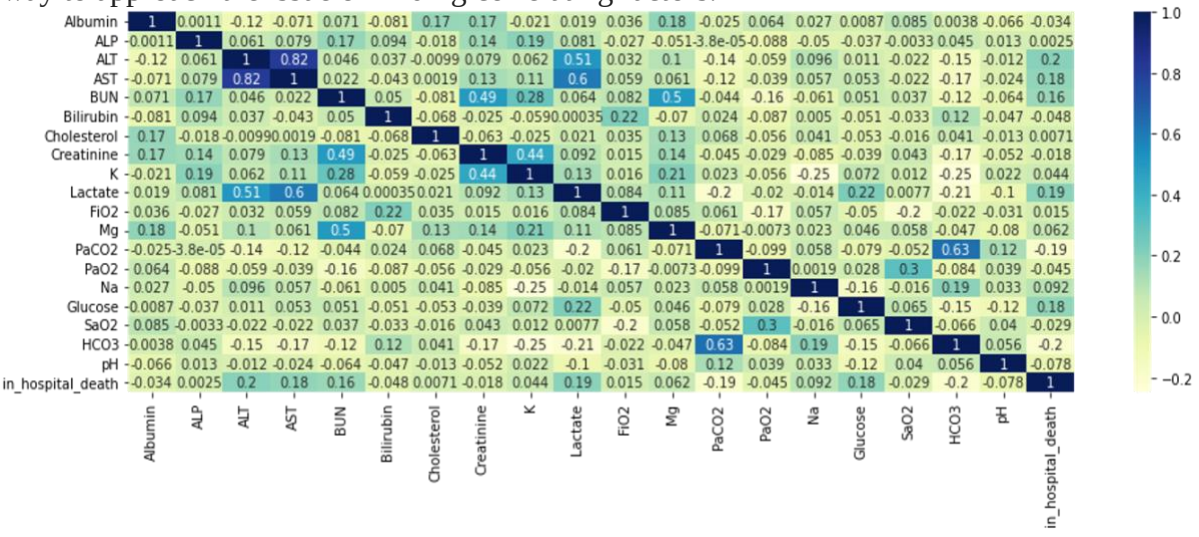
However, there are other methods that can be used to account for this factor, using normalisation but box plots were chosen because they were easy to analyse and clearer to comprehend. Yet, it is important to note that by using boxplots as the main method of modeling, accuracy was reduced. If the usage of normalisation was considered, there would have been far more accurate results.

In this dataset, no ideal box plots could be identified. The properties of an ideal pair of boxplot were that both boxplots had completely different median values, in the respect that both boxplots' medians were contained within completely different interquartile ranges. Another property that was identified in box plots that were judged as ideal, was the presence of a pair of boxplots with interquartile ranges that did not have any coinciding values. A hypothetical example of the ideal pair of box plots in the succeeding figure.



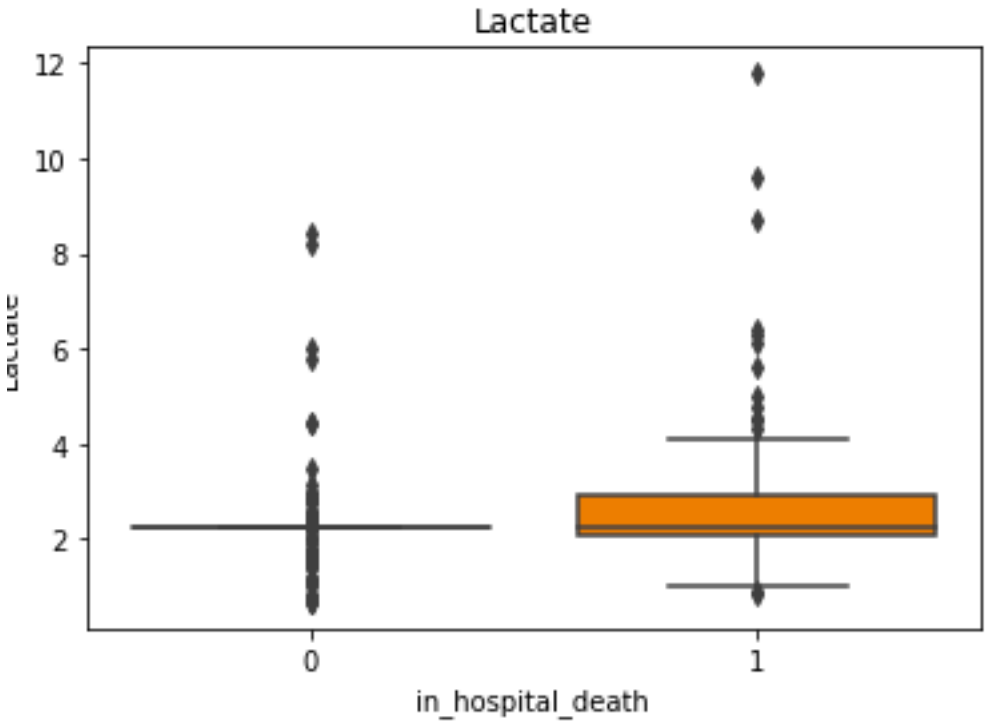
Note that the interquartile ranges and means are visibly distinguishable. This property would have allowed for a more definitive comparison between variables to be made.

The use of a heatmap to quantify correlation between variables was a decent method to propose because the original premise of using a heatmap was to find a comprehensive, objective method to present strength of correlation which, in principle, was an appropriate way to approach the issue of finding correlating factors.

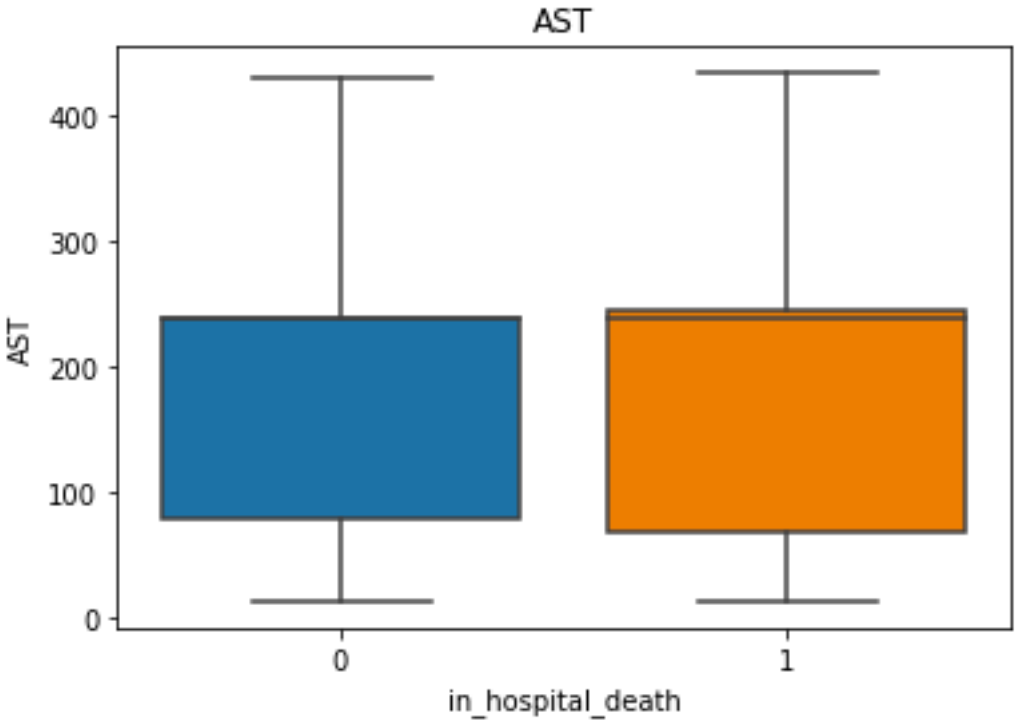


It can be seen that ALT, AST, BUN, Lactate and Glucose have a strong correlation to in hospital death. However, Mg was also chosen as a variable that had strong correlation

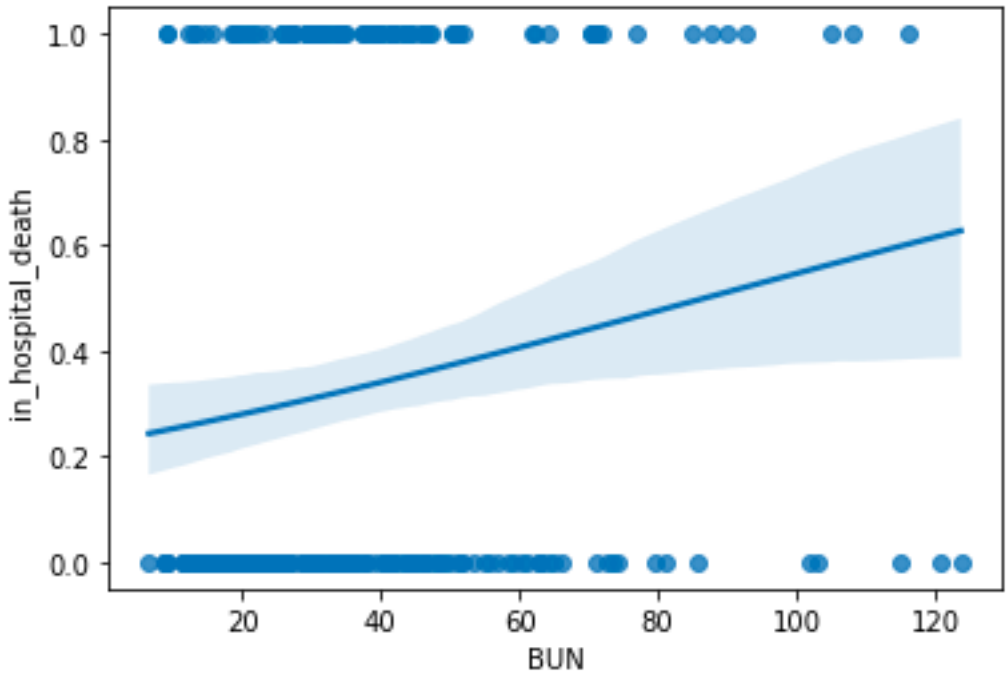
because of its boxplot properties. In addition, Lactate could not be used for analysis because of the proportion of outliers present.



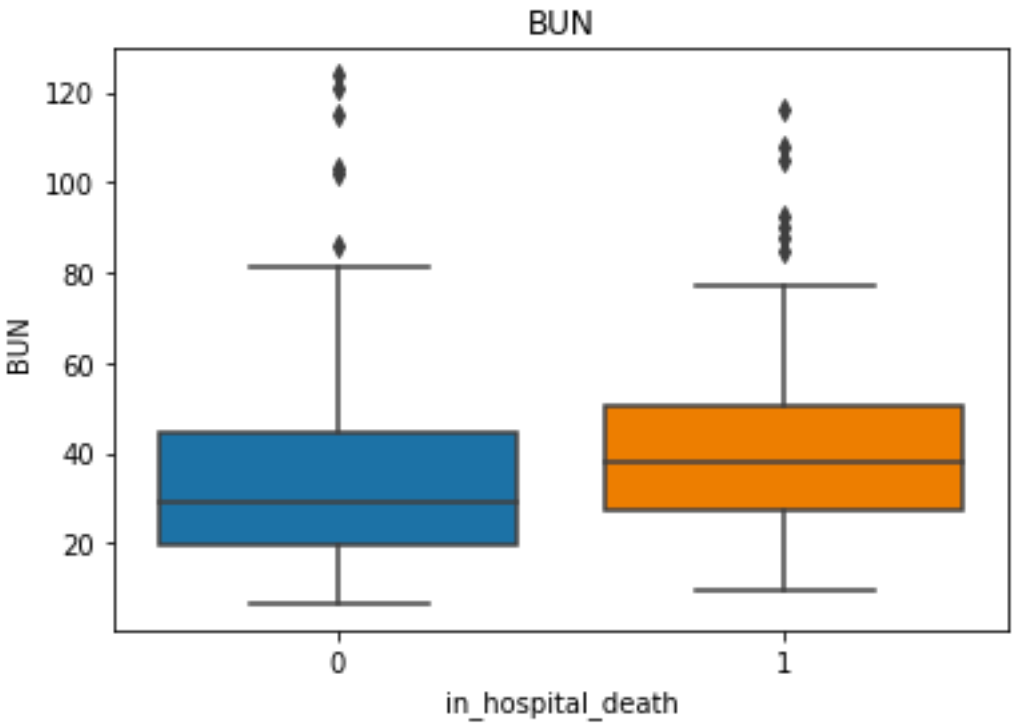
In the 0 column, just under a third of patients who survive are considered outliers which made this variable not ideal to analyse. For the AST variable, the interquartile ranges of both boxplots were too similar when compared to variables such as Mg.



Although logistic regression was not covered in the presentation, attempts were made to construct a regression model. These models affirmed the conclusions that were deduced in the boxplots. Here is an example of a regression plot that affirmed such conclusions.



It can be seen that high levels of BUN correlated to decreased survivability which is supported in the boxplot shown.



The interquartile range and median value of patients who survived is lower than that of those who passed, supporting the argument that was proposed by the interpretation of the regression line.

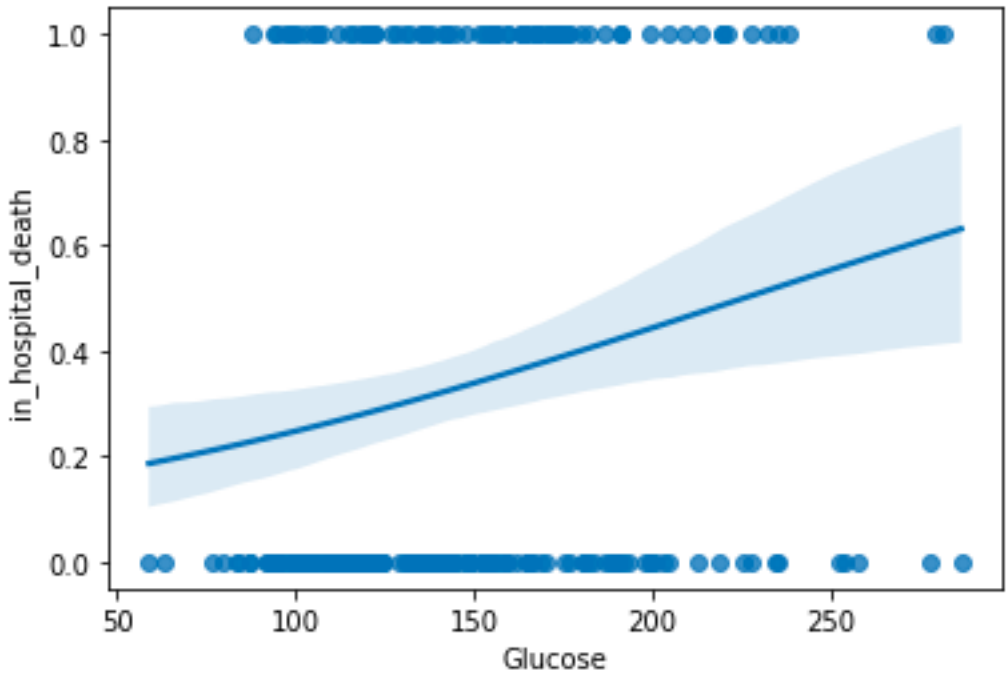
To quantify the correlation between each variable and patients' outcome, coefficients were calculated for BUN, Mg and Glucose. A parameter of the model used to calculate the results was that a training set size of 80% and a testing set size of 20% was used. It was found that Mg had the strongest correlation with a coefficient of 0.16 out of the 3 main variables analysed. Furthermore, the accuracy of the regression models were tested with the specific example of magnesium yielding an accuracy score of 0.679 for the training set and 0.612 for the testing set, indicating a reasonable level of accuracy. The other significant aspect of the method used to quantify accuracy of the models was the utilisation of a classification system which showed the hypothetical outcomes of each patient in the testing set based on the fitted training set. These hypothetical outcomes were then compared to the true outcomes of each patient to determine the model's accuracy. Any patient with a probability of survival less than 0.5 was assigned to be hypothetically dead and this value of 0.5 was determined through experimentation of other probabilities around that similar range. It was eventually decided that the most accurate of these analysed probabilities was 0.5.

Exploratory Data Analysis

From the models generated, relationships can be determined between variables, which can be explained by external research conducted about the context of the project.

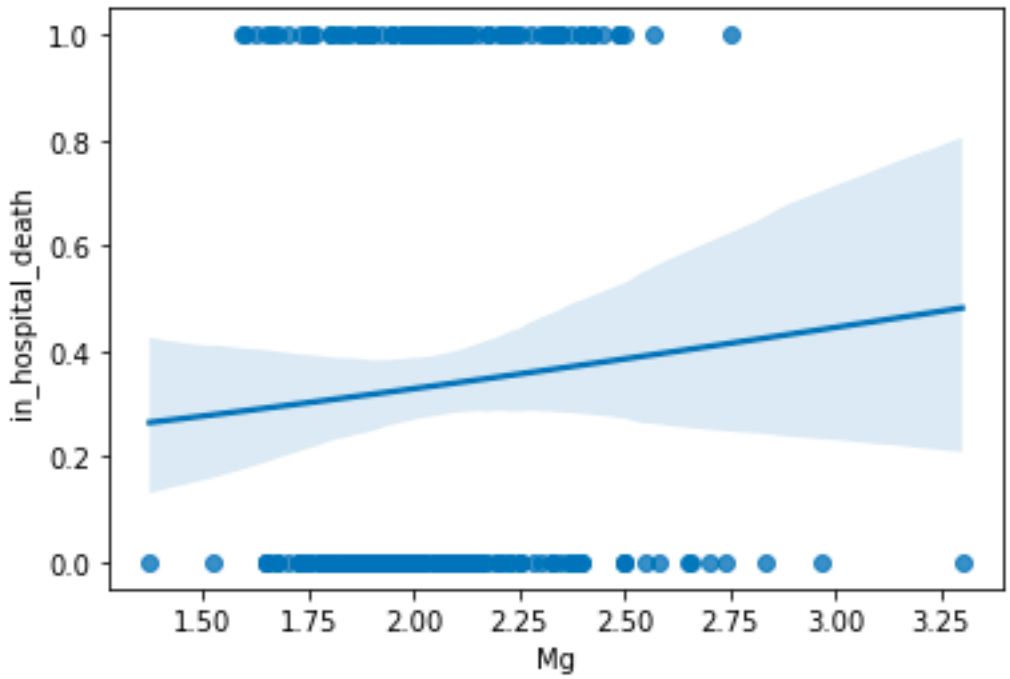
Blood urea content (BUN) is the concentration of urea in the bloodstream. Weakening of arterial walls can lead to overall low blood pressure and thus kidney function is hindered as not enough blood is being filtered such that excess urea stays within the blood. Similarly, high blood pressure caused by arterial obstruction can damage the small blood vessels that circulate around the kidneys, further limiting urea-blood to be expelled. Furthermore, elevated levels of BUN are indicators of larger issues such as severe dehydration, kidney failure and heart failure (Mayoclinic, 2021). As a result of BUN's large indicative nature of complications, it thus supports the model's positive relationship between BUN and ICU death as shown in the data modeling section of this report.

Serum glucose is a blood test to measure glucose concentration in the blood, usually to screen for diabetes. Excess levels of glucose (hyperglycaemia) are a strong indicator of coronary artery disease, stroke, and peripheral arterial disease (Aronson & Rayfield, 2002). This result is due to excess glucose damages the arterial walls, decreasing their elasticity and causing them to narrow and restrict blood flow. Furthermore, hyperglycaemia seems to support the growth of other processes such as atherosclerosis, which is the buildup of plaque in the arteries which also restrict blood flow (Aronson & Rayfield, 2002). As a result of glucose's damaging nature to artery structure, it thus supports our model's increased rate of ICU death, with higher levels of blood glucose. Evidence of the model supporting the aforementioned claims is shown below.



It can be seen that in this logistic regression model, there is indeed a correlation of positive regression between glucose and in-hospital death which translates to higher levels of glucose increasing the likelihood of death in coronary ICU. Further variations of evidence supporting the hypothesis that increased levels of glucose in the bloodstream leads to increased probability of death can be found in the appendices provided.

Magnesium (Mg) functions as an electrolyte and serves to support muscle and nerve function. Magnesium is stored in the bones, and little circulates in the blood thus excess magnesium in the blood (hypermagnesemia) is uncommon (Lewis III, 2021). Hypermagnesemia can cause poor muscle control (muscle spasms and weakness), low blood pressure and can lead to the heart completely stopping. Thus, the data highlights that generally higher levels of magnesium generates higher death outcomes in ICUs however the heatmap determined magnesium not to be as correlated as the logistic regression model and box plots suggested. A possible reason for this is, hypermagnesemia often occurs when patients are supplemented with magnesium salts or drugs to combat kidney failure, such that patients' magnesium levels are inflated within the original data set. Thus a possible explanation is hypermagnesemia generally does not lead to death, however it is an indication that patients are undergoing treatment for kidney failure, which could potentially increase their vulnerability. Thus, this explains the positive regression between magnesium and ICU death where the relationship should not be as pronounced. Examples that affirm the implications and arguments mentioned above are shown below.



It can be seen that there is positive regression occurring which affirms the result that increased levels of magnesium may lead to reduced probability of survival. Another important observation is that the positive regression is perhaps not as pronounced as glucose or BUN despite the coefficient of the model being the highest, which also supports the claim the magnesium is not directly correlated but is rather a more subtle indicator of increased vulnerability in ICU. For example, it was hypothesised that patients with high levels of magnesium were likely undergoing treatment for kidney failure in the form of taking medication containing elevated levels of Mg content, citing the indirect nature of the relationship between patient outcome and magnesium levels. This hypothesis is affirmed by the heatmap showing a reduced correlation value compared to the other observed variables.

Conclusion

After analysis of the PhysioNet dataset, 3 distinct arterial related variables were linked with ICU death. Those variables were BUN, Mg and Glucose as the most relevant to ICU death. All of these three variables are positively correlated to the ICU death, which means the increasing of these three variables may lead to an increasing chance of ICU death. However, an important concept to remember and may be as applicable to health-related events as any in other fields is correlation does not always equal causation. Before data modeling and analysis, a qualitative pre-screening of all the variables was conducted, to make predictions which variables would likely correlate to coronary health outcomes from already established clinical data. From
this pre-screening, it was assumed that almost all the variables had a negative effect on the arteries; such as artery obstruction, blood viscosity and deviating blood ph. However, after modeling, the data showed no significant ICU death rate difference in the majority of these variables. One variable in particular was troponin. Troponin is a protein that is released from your heart when it is damaged. A higher concentration of troponin in your blood would indicate a more damaged heart and thus should produce higher death outcomes; however, the model showed almost identical box plots; with the same median, interquartile, upper and lower quartiles.

It would appear that either the accuracy of the preliminary research or the model/data was off; however, linking back to correlation and causation, applying to all the variables even if the data showed no correlation; particularly when it comes to health, all these variables in isolation are indicators of a larger problem. Any negative health outcome is almost never due to a single causing factor. A single variable can have a huge cascading effect and trigger other events. But ultimately, what causes someone to die is due to a compounding effect of multiple variables and can all interact with each other and accelerate their effects.

This is not to say that the produced model can be improved. Further usage of more advanced machine learning techniques than logistic regression is recommended to better tackle the aim of the Physionet challenge which was to predict with higher accuracy, the probability of survival. Alternate approaches could have been taken, such as utilising the minimum and maximum values in the analysis or individually analysing each data point on a two day basis instead of using merging techniques. Most importantly though, ABP variables could have been used instead as they are also highly relevant to the coronary ICU.

References

American Heart Association, (2022, March 4). How High Blood Pressure Can Lead to Kidney Damage or Failure, American Heart Association. https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-kidney-damage-or-failure

Bergmeir, C. (2020). Predictions of Mortality in ICU patients project, Monash University.

Doron, A. & Rayfield, E. J. (2002, April). *How hyperglycemia promotes atherosclerosis: molecular mechanisms*.Cardiovascular Diabology https://cardiab.biomedcentral.com/articles/10.1186/1475-2840-1-1

Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.C., Mietus, J.E, Moody, G.B., Peng, C., Stanley, E. (2000).*PhysioBank, PhysioToolkit, and PhysioNet*: *Components of a new research resource for complex physiologic signals.* American Heart Association. 101 (23), pp. e215–e220." https://www.researchgate.net/profile/Shlomo-Havlin/publication/243775682_Components_of_a_new_research_resource_for_complex_physiologic_signals/links/5656e6a208ae4988a7b50f03/Components-of-a-new-research-resource-for-complex-physiologic-signals.pdf?_sg%5B0%5D=TCHDkKLEJX8n5ynADh_h7yBUMIsVnv2ULczN393izV_ful8TG3dK_7ezWueDJ4a-kYs5T8gcgj-avns4TKqZzw.nT99_aaXvkTATeTJkUSPhqW2iBVHvP3m_8Q_bVXIxvEYLgsZXK2d4GpBPzjM4kGEB40dGf0s9bpUOyYWXq1naA&_sg%5B1%5D=aU4m7qc_1z0RPgUnmK8q7tf8jYqUrKdg1HX2OUqv-EQp-SwG7vvKwoBr5QP8kyyORrd07diroGiPoCO7oCXw4xlvloa366R_4Q65LVD57qtK.nT99_aaXvkTATeTJkUSPhqW2iBVHvP3m_8Q_bVXIxvEYLgsZXK2d4GpBPzjM4kGEB40dGf0s9bpUOyYWXq1naA&_iepl=

James Cook University. (2022). *The vital role of data science in advancing medicine*. <https://online.jcu.edu.au/blog/data-science-medicine#:~:text=The%20Use%20of%20Data%20Science,medical%20research%20more%20data%2Ddriven.> [Accessed 29 May 2022].

Lewis III, J. L. (2021 October). Hypermagnesemia (High Level of Magnesium in the Blood). MSD Manuals. https://www.msdmanuals.com/home/hormonal-and-metabolic-disorders/electrolyte-balance/hypermagnesemia-high-level-of-magnesium-in-the-blood

Mayo Clinic, n.d. Blood Urea Nitrogen (BUN) test, Mayo Clinic. https://www.mayoclinic.org/tests-procedures/blood-urea-nitrogen/about/pac-20384821#:~:text=Generally%2C%20a%20high%20BUN%20level,Urinary%20tract%20obstruction