1 Introduction

The study aims to identify predictors of in-hospital mortality and length of stay (LOS) using data from 978 randomly selected patients admitted to a large Virginia medical centre with over 500 beds between January and September 2014. The dataset includes patient characteristics such as age, gender, BMI, illness severity, oxygen saturation (spO2), blood pressure, and temperature,key determinants of outcomes like mortality and LOS (Greenwood et al., 2021).

This analysis evaluates the influence of these factors' on a patient outcomes to improve hospital management, optimise resource utilisation, and enhance patient care (Fenton et al., 2020). By uncovering these predictors, healthcare providers can better assess patient prognosis and deliver efficient, high-quality care.

2. Model Selection

2.1 Testing for Multi-Collinearity

Multicollinearity refers to a situation where predictor variables in a regression model are highly correlated, which can inflate standard errors and make it difficult to assess the individual effect of predictors. With mortality as the regressor, I carried out a Variance Inflation Factor (VIF) test to ensure the integrity of the model, for our independent variables x_i , as shown in Table 1. I excluded avpu, risk and severity from this analysis due to their categorical nature.

All VIF values are below the threshold of 10, with an average (VIF) ~ 1, indicating no significant multi-collinearity among the variables. This suggests that each variable contributes insights into the factors affecting mortality affirming the validity of the upcoming analysis.

Variable	VIF
age	1.1705
gender	1.0594
bmi	1.0741
sp02	1.0260
sbp	1.6270
dbp	1.7038
pulse	1.4213
respiratory	1.2067
temperature	1.1330

Table 1: Variance Inflation Factor Results

2.2 Model of Choice: Logistics Regression

A logistic regression model is appropriate for understanding the relationship between mortality (binary: died or survived) and variables such as age, sex, BMI, severity, risk, and physiological measures (BP, SpO2, pulse rate). Coefficients are exponentiated to interpret log-odds ratios, showing the impact of a one-unit predictor increase on mortality.

Logistic regression accommodates continuous and categorical variables, making it suitable for this dataset. Commonly used in medical research, it identifies factors by analysing patient characteristics and outcomes (Plum, 2020).

Logistic Lasso regression refines this approach by retaining only predictors with significant impacts, shrinking irrelevant variables to zero, enhancing interpretability and accuracy.

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n$$

where:

- $P(Y=1)$ is the probability of patient dying,
- β_0 is the intercept,
- $\beta_1, \beta_2, \beta_3$ are the coefficients of predictors X_1, X_2, X_3,

Figure 1: Logistic Regression Model Equation

2.2.1 Assumptions

The logistic lasso regression model is built with several assumptions in mind:

1. **Binary outcome:** The dependent variable can be binary, allowing effective handling of dichotomous outcomes by modeling the probability
2. **Homoscedasticity:** The variance of the residuals remains constant across all levels of the independent variables.
3. **Normality of Residuals:** The residuals should be normally distributed.
4. **No Multicollinearity:** Independent variables should not be highly correlated with each other, as confirmed by our VIF analysis.

2.3 Model Selection: GAM Model

The Generalised Additive Model (GAM) effectively predicts length of stay (LOS) as it captures both linear and nonlinear relationships through smooth terms for continuous variables. This approach reduces residual error without requiring log transformation. GAM assumes data follows a normal distribution (Gaussian family), suitable for LOS as a continuous variable. The approximate degrees of freedom (edf) quantify smooth term complexity, with high edf values indicating potential overfitting and edf near one reflecting adequate smoothing.

2.3.1 Assumptions

The GAM model is built with several assumptions in mind:

1. **Linear & Non-Linear:** The relationship between the dependent variable and independent variables can be linear or non-linear
2. **Independent observations:** The probability of a patient staying in hospitals for longer is independent of other patients
3. **Homoscedasticity:** The variance of the residuals remains constant across all levels of the independent variables.
4. **Normality of Residuals:** The residuals should be normally distributed.
5. **No Multicollinearity:** Independent variables should not be highly correlated with each other

Ensuring these assumptions hold is necessary for the validity of the regression results and subsequent interpretations.

3 Results and Interpretation.

3.1 Logistic Regression Model for Mortality

The logistic regression model was developed to predict the probability of mortality, a binary outcome: "died" or "survived". The odds ratios, obtained by exponentiation of coefficients, represent the nature and magnitude of change in the odds of mortality per change in each predictor variables.

Variable	**Odds Ratio**	**Coefficient**
intercept	$5,780,751$	15.570
age	1.029	0.029
gender (male)	2.209	0.793
bmi	0.965	-0.036
severity 2 - Moderate	0.692	-0.368
severity 3 - Major	2.498	0.915
severity 4 - Extreme	6.757	1.911
risk 2 - Moderate	0.539	-0.619
risk 3 - Major	1.376	0.319
risk 4 - Extreme	6.418	1.859
spO2	0.896	-0.110
sbp	1.002	0.002
dbp	0.992	-0.008
pulse	1.002	0.002
respiratory	1.054	0.053
avpu (pain)	3.004	1.100
avpu (unresponsive)	2.504	0.918
avpu (voice)	3.761×10^{-8}	-17.096
temperature	0.885	-0.122

Table 2: Logistic Regression Results: Odds Ratios and Coefficients

The logistic lasso regression model identifies key predictors of mortality by penalising less important coefficients as seen in Table 3. **Severity 4** (β=0.511) has the strongest association, significantly increasing the log-odds of mortality for the most severe patients, emphasising its role as a critical determinant of survival. **Risk 4** β=0.261) also shows a positive association, reflecting elevated mortality odds for high-risk patients, highlighting the compounded vulnerability of critically ill individuals.

A negative relationship between mortality and oxygen saturation (β=−0.038) indicates reduced mortality risk with higher spO2 levels, aligning with the clinical importance of oxygenation in maintaining organ function. Among continuous variables, **age** (β=0.032) shows a modest positive relationship, with older patients facing slightly higher mortality risk, consistent with established evidence linking age

to poorer health outcomes. This model underscores the central roles of severity, risk, and oxygenation in predicting patient survival.

Variable	Coefficient
intercept	−2.523
age	0.032
bmi	−0.011
severity 4 - Extreme	0.511
risk 4 - Extreme	0.261
spO2	−0.038
sbp	−0.009
dbp	−0.005
pulse	0.017
respiratory	0.046

Table 3: LASSO Logistic Regression Coefficients

.

3.2 GAM Model for Length of Stay (LOS)

The fitted Generalised Additive Model (GAM) assesses predictors' impact on hospital length of stay (LOS), summarised in Tables 5 and 6. Parametric coefficients represent linear relationships with categorical predictors, while smooth terms capture relationships with continuous predictors. The formula is provided below:

$$\text{los} \sim s(\text{age}) + s(\text{bmi}) + \text{severity} + \text{risk} + s(\text{sp02}) + s(\text{sbp}) + s(\text{dbp}) + s(\text{pulse}) + \text{respiratory} + \text{avpu} + s(\text{temp})$$

Figure 1: Formula for GAM Model

Variable	**Estimate**	$Pr(>\|t\|)$
intercept	3.555	1.54×10^{-5}
severity 2 - Moderate	0.854	0.0112
severity 3 - Major	2.598	1.79×10^{-8}
severity 4 - Extreme	6.539	8.97×10^{-16}
risk 2 - Moderate	0.001	0.9973
risk 3 - Major	0.179	0.7005
risk 4 - Extreme	1.037	0.1952
respiratory	−0.030	0.4915
avpu (pain)	−0.346	0.7873
avpu (unresponsive)	0.640	0.5968
avpu (voice)	−0.695	0.4486

Table 5: Summary of Parametric Coefficients

Smooth Terms	**edf**	**p-value**
age	2.172	0.1743
bmi	4.521	0.4187
sp02	4.279	0.1738
sbp	1.000	0.8633
dbp	1.000	0.3054
pulse	1.000	0.2545
temp	1.000	0.0249

Table 6: Summary of Smooth Terms: edf and Significance

At a 5% significance level, the parametric coefficients show that severity significantly impacts length of stay (LOS). **Severity 4** (β=6.539) has the strongest association, indicating patients in this category stay much longer. **Severity 3** (β=2.598) and **Severity 2** (β=0.854) also show significant but smaller positive associations, confirming severity as the primary determinant of LOS.

The significant smooth term for **temperature** (edf=1.000) suggests a nonlinear relationship, where temperature variations modestly affect LOS. This highlights the potential to forecast LOS more accurately by considering both severity and temperature.

Conversely, predictors such as **risk 4** and **respiratory rate** are not statistically significant ($p>0.05$), indicating weaker associations. These results emphasise the central roles of severity and temperature in determining LOS, while other variables contribute minimally.

4 Evaluation

4.1 Logistic Lasso Regression Model

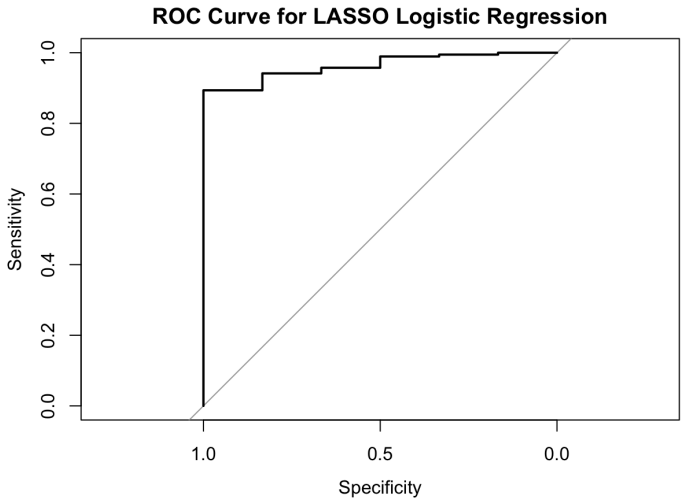

Figure 2 ROC Curve for Logistic Lasso Regression

Model	AUC
Logistic Regression	0.81
Lasso Logistic Regression	0.96

Table 4: Comparison of AUC of models

The performance of the regression model for mortality prediction can be assessed by calculating the Area Under the Curve (AUC) of the ROC curve. The ROC curve is a graphical representation of sensitivity versus specificity across a range of classification thresholds. A high AUC indicates better performance in achieving sensitivity and specificity, establishing logistic regression as an accurate tool for predicting in-hospital mortality. Implementing lasso improved the AUC by 15%, demonstrating excellent discriminatory ability between patients who survived and those who died. This suggests that the model accurately predicts mortality outcomes and enhances clinical utility.

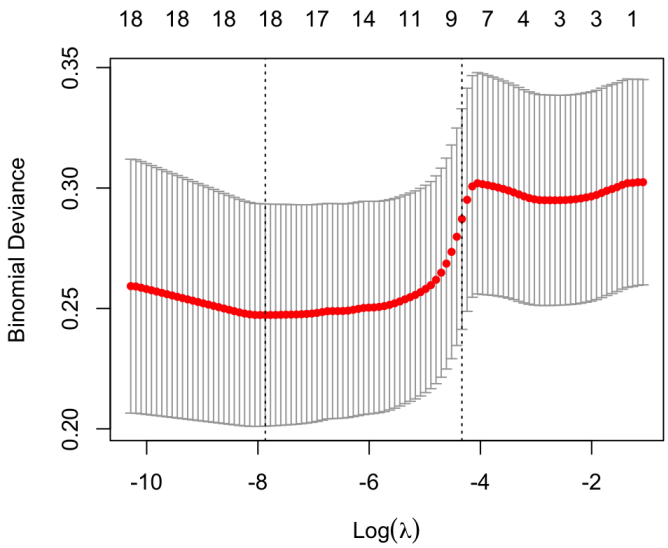

Figure 3 Lasso Regression Plot

The lasso regression plot shows the relationship between log(λ) and binomial deviance, indicating model fit. As λ increases, regularisation reduces model complexity but raises deviance. The small error bars around λmin suggest the stability and generalisability of the model, accurately predicting outcomes like mortality.

4.2 Generalised Additive Model

Metric	Value
Mean Squared Error	8.67
Root Mean Squared Error	2.94

Table 7: Summary of Model Errors

Metric	Value
Adjusted R-squared	0.226
Deviance Explained (%)	24.5%
Generalized Cross-Validation (GCV)	14.33

Table 8: Summary of GAM Model Metrics

The model's average squared error for LOS predictions is 8.67 days, with a root mean squared error (RMSE) of 2.94, indicating reasonable predictive accuracy. However, the adjusted R-squared of 0.226 and deviance explained of 24.5% suggest the model captures only a moderate portion of LOS variability, likely due to unmeasured factors such as comorbidities or hospital policies. Predictors like **respiratory rate** and **risk 4** lack statistical significance ($p>0.05$), reflecting limited explanatory power, potentially due to multicollinearity or the small sample size. An improvement to the model could be introducing interaction terms that capture joint effects. High degrees of freedom for some smooth terms (e.g., BMI, edf=4.521) raise concerns about overfitting and generalisability. Many predictors, including risk levels, contribute minimally, indicating opportunities to refine the model for greater precision in forecasting LOS. Generalised Cross-Validation (GCV) measures a model's predictive performance, with lower values indicating a better fit. However, its interpretation depends on the response variable's scale, and without a comparable figure, it provides limited insight in this instance.

4.3 Diagnostics Plot Analysis

4.3.1 Residual Plot for Mortality Model

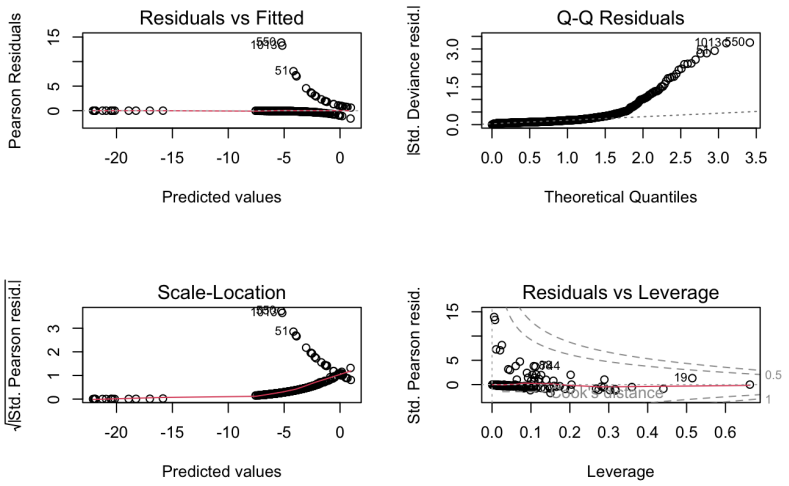

Figure 4 Diagnostic Plots of Logistic Regression Model

The residual plot for the logistic model appears random, consistent with the Gaussian assumptions of logistic regression, suggesting a good model fit with no notable outliers or deviations. The leverage plot shows no high-leverage points, indicating no undue influence from individual observations. The absence of heteroscedasticity suggests unbiased standard errors.

4.3.1 Residual Plot for LOS Model

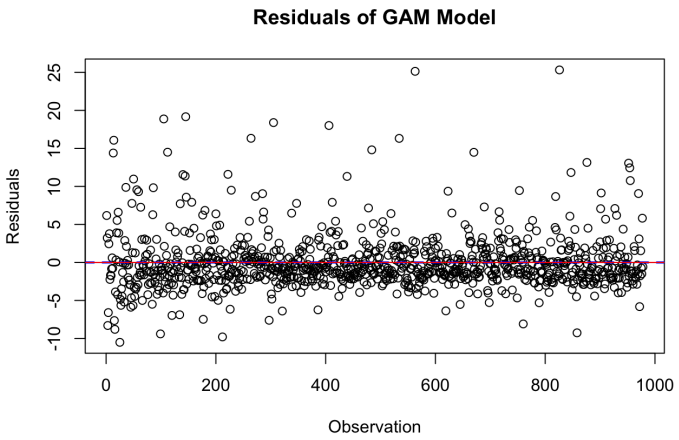

Figure 5 Residual of GAM Model

The residual plot for the LOS model reveals no discernible patterns, indicating that the model is well-specified and free from issues like non-linearity or heteroscedasticity.

5 Conclusion

Severity levels, particularly **severity 3** and **severity 4**, are the strongest predictors of both mortality and length of stay (LOS). Patients in these categories stay considerably longer than those in the baseline group and face significantly higher odds of mortality, emphasizing the critical role of severity in determining hospital outcomes. Conversely, predictors such as **risk levels** and **vital signs** (e.g., respiratory rate) showed limited statistical significance, suggesting a smaller or negligible influence in these models.

Continuous variables like **age**, **BMI**, and **spO2** exhibited modest relationships with mortality but did not significantly impact LOS. This aligns with findings that non-linear predictors, like temperature (significant in LOS), may better capture variations in hospital stay duration. However, the LOS model explained only 24.5% of the variance, highlighting the potential influence of unmeasured factors such as comorbidities, hospital policies, or social determinants of health.

Both models demonstrate reasonable predictive accuracy but face limitations. The GAM for LOS risks overfitting due to high degrees of freedom for certain smooth terms (e.g., BMI), while the logistic regression model, despite an improved AUC of 15% with LASSO, is limited by unexplained variance and weaker contributions from some predictors. Future improvement to the models can be made, for example incorporate interaction terms to the GAM would better mimic and fit the complex factors influencing patient outcomes.

6 References

McClave, S.A., Taylor, B.E., Martindale, R.G., Warren, M.M., Johnson, D.R., Braunschweig, C., McCarthy, M.S., Davanos, E., Rice, T.W., Cresci, G.A. and Gervasio, J.M., 2016. Guidelines for the provision and assessment of nutrition support therapy in the adult critically ill patient: Society of Critical Care Medicine (SCCM) and American Society for Parenteral and Enteral Nutrition (ASPEN). *JPEN. Journal of parenteral and enteral nutrition*, *40*(2), pp.159-211.

McIlvennan, C.K., Eapen, Z.J. and Allen, L.A., 2015. Hospital readmissions reduction program. *Circulation*, *131*(20), pp.1796-1803.

Pum, J.K., 2020. Evaluating sample stability in the clinical laboratory with the help of linear and non-linear regression analysis. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *58*(2), pp.188-196.