Individual Coursework Submission Form

Specialist Masters Programme

Surname: Lai	**First Name: Hoang Duong**
MSc in: MSc Business Analytics	**Student ID number: 240034730**
Module Code: SMM634	
Module Title: Analytics Method for Business	
Lecturer:Dr. Rosalba Radice	**Submission Date: 09/12/2024**

Declaration:

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:		**Final Mark:**	%

Table of Content

1. Model Selection for Mortality

a. Data inspection

The first part of this report aims to analyse and fit a model to estimate the probability that a patient died based on certain health factors. The data that this model was built on is taken from 978 randomly selected patients admitted at an over-500-bed medical center in the state of Virginia between January and September 2014.

The table below shows the variables and their description. The name of each variable will also be used as abbreviations throughout the report.

Variables	Variable type	Description
los	Continuous	Length of stay
died	Binary	1 for died 0 for alive
age	Continuous	Age of patient
gender	Binary	"male" or "female"
bmi	Continuous	Body Mass Index
severity	Categorical	4 – Extreme 3 – Major 2 – Moderate 1 – Minor
sp02	Continuous	Oxygen saturation level
sbp	Continuous	Systolic blood pressure.
dbp	Continuous	Diastolic blood pressure
pulse	Continuous	Pulse rate
respiratory	Continuous	Respiratory rate
avpu	Categorical	A: Alert V: responding to Voice P: responding to Pain U: Unconscious
temp	Continuous	Body temperature

As the goal of the model is to predict the probability that a patient will die or not, the dependent variable is a binary variable. As the two most common models to calculate the probability of a binary dependent variable are probit and logit model, this report will proceed to find the most suitable model among them.

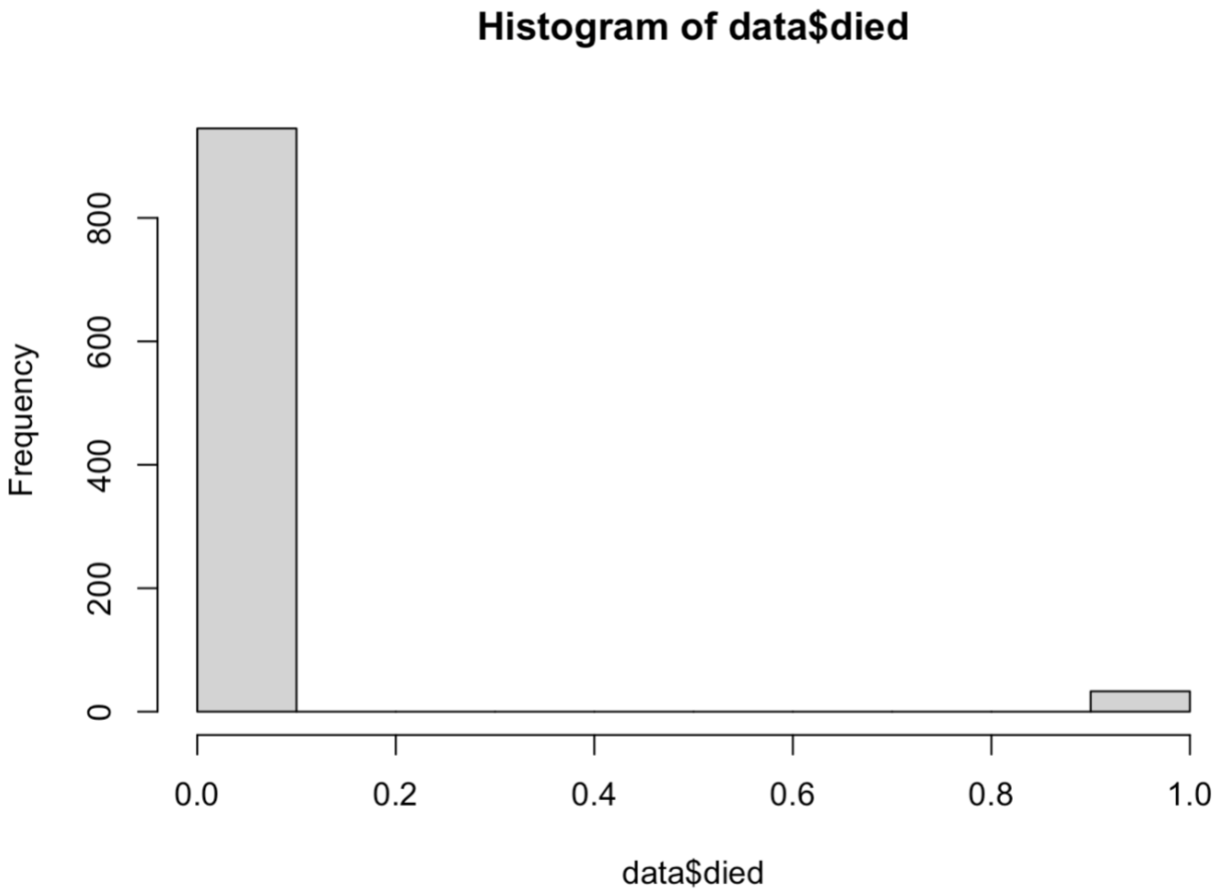
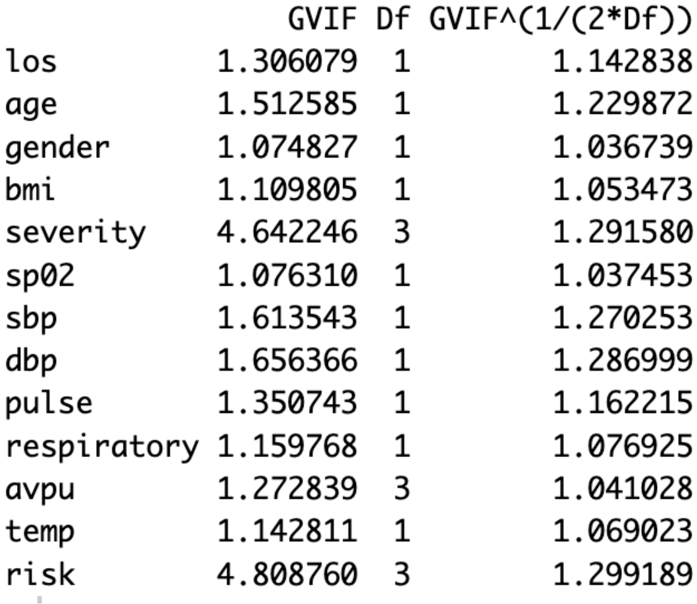


Figure 1. Distribution of mortality rate

While inspecting the distribution of observations for people who died and not, we can easily spot a highly imbalanced structure, with a very large proportion of 0s (survivors) and only a small proportion of 1s (deaths). With this distribution, most observations are concentrated at 0 while very few at 1, the logit model can better predict these small probabilities without being overly sensitive to imbalanced data. Combined with its ease of coefficients interpretation, logit model will be the focus of this report to generate a prediction model.

b. Model fitting:

Before fitting the model, it is important to assess potential collinearity among the variables. A Variance Inflation Factor (VIF) test was conducted using a linear model, revealing a possible correlation between the risk and severity factors. This indicates the need for caution when including these variables in the final model. Apart from this, no significant signs of collinearity were observed among the other variables.



```
                GVIF Df GVIF^(1/(2*Df))
los         1.306079  1        1.142838
age         1.512585  1        1.229872
gender      1.074827  1        1.036739
bmi         1.109805  1        1.053473
severity    4.642246  3        1.291580
sp02        1.076310  1        1.037453
sbp         1.613543  1        1.270253
dbp         1.656366  1        1.286999
pulse       1.350743  1        1.162215
respiratory 1.159768  1        1.076925
avpu        1.272839  3        1.041028
temp        1.142811  1        1.069023
risk        4.808760  3        1.299189
```

Figure 2. VIF score for all variables

After testing for collinearity in the variables, the logit model was fit and refined to find the most significant variables. The final model chosen was a logit model with the following formula:

$$P(died\ =\ 1\,|\,los,\ sp02,\ risk|)\ =\ \frac{1}{1+e^{-(\beta 0\,+\,\beta 1.los\,+\,\beta 2.sp02\,+\,\beta 3.risk\ 1\,+\,\beta 4.\ risk\ 2\,+\,\beta 5.risk\ 3\,+\,\beta 6.risk 4)}}$$

Among the all the variables, length of stay, sp02 and risk show the most significant impact on the probability that a patient will die.

2. Model selection for Length of Stay

a. Data Inspection

Length of Stay represents the number of days a patient spends in a hospital. As a positive count variable greater than zero, a Poisson model is a natural choice. However, since Length of Stay can also be viewed as a continuous positive variable, a Gamma model may also be appropriate. To identify the model that offers the best goodness of fit and is most interpretable, both models were evaluated, and the most suitable option was selected.

Further analysis of the distribution of Length of Stay revealed that the number of hospital-stay days is over dispersed, as indicated by the histogram and the ratio of the variable's variance and mean. This prompted a shift in modeling focus toward Quasi-Poisson, Negative Binomial, and Gamma models. The skewness of the LOS distribution further supported the consideration of the Negative Binomial and Gamma models, as both account for dispersion through the parameters α α (for Negative Binomial) and ϕ (for Gamma).

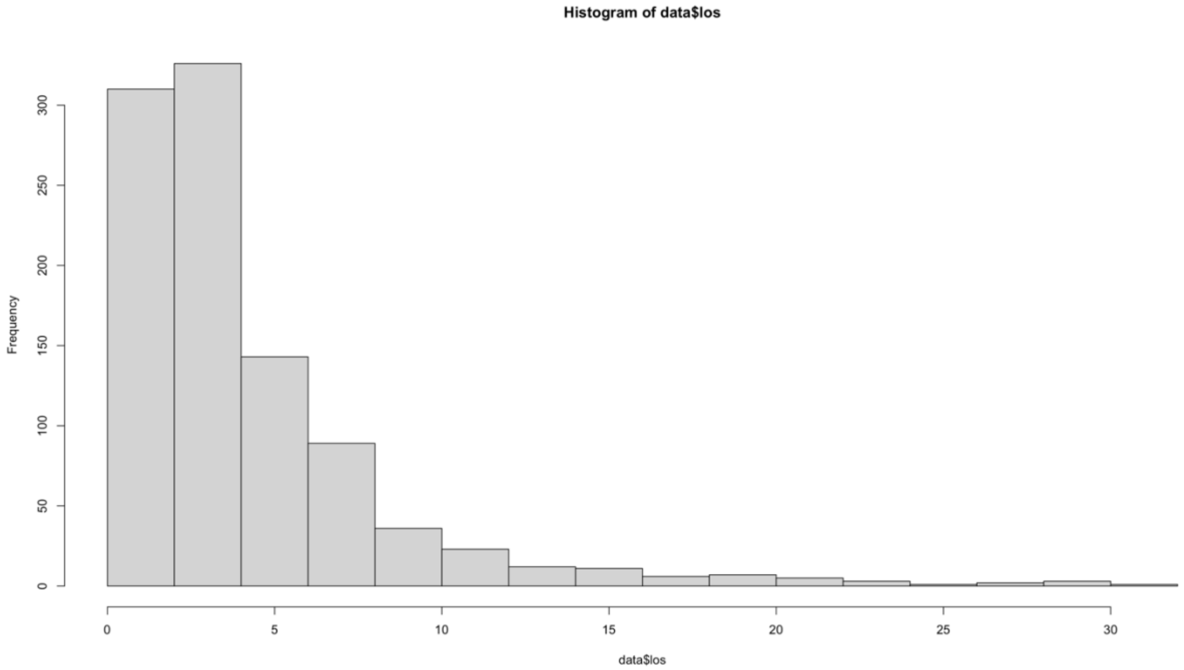


Figure 3. Distribution of days spent at hospital

$$Dispersion\ ratio\ (los) = \frac{Variance\ of\ los}{Mean\ of\ los} = 3.84$$

b. Model fitting

After analysing and testing both the Negative Binomial and Gamma models for predicting Length of Stay, the Negative Binomial model emerged as the preferred choice. This decision was based on its ability to strike a balance between ease of interpretation and statistical properties fit. The final fitted model is as follows:

$$\log(\mu) = \beta 0 + \beta 1.died + \beta 2.genderM + \beta 3.severity2 + \beta 4.severity3 + \beta 5.severity4$$

Where:
μ = expected value of length of stay of a patient

This model indicates that, the gender of the patients, his or her severity of illness and whether that patient died or not have an impact on his or her length of stay at the hospital.

3. Summary and Interpretation

a. Mortality probability model

The collinearity between risk and severity underscores the relationship between a patient's severity level and their likelihood of mortality. Specifically, patients admitted under extreme severity conditions face a significantly higher probability of dying. This observation implies that higher severity levels correspond to an increased risk of mortality.
The following table presents the results of fitting a Logit model to estimate the probability of mortality for patients. It includes both the estimated coefficients and their exponentiated values, which provide a more intuitive interpretation of the impact of each predictor on the odds of mortality.

Variables	Coefficients	Exponentiated
(Intercept)	5.98013	
los	-0.13818	0.87
sp02	-0.11279	0.89
Risk2 - Moderate	0.40491	1.50
Risk3 - Major	2.49690	12.14
Risk4 - Extreme	5.11816	167.03

The table of coefficients highlights the key factors influencing the probability of patient mortality, including Length of Stay (**los**), Oxygen Saturation (**sp02**), and Risk level (**risk**). By exponentiating the coefficients, we obtain the odds ratios, which show how the odds of mortality change with a 1-unit increase in each predictor.

- **Length of Stay (los)**: A 1-day increase in **los** is associated with a decrease in the odds of mortality, indicating that patients who stay longer are less likely to die. This might reflect successful treatment and recovery over time.
- **Oxygen Saturation (sp0$2_2$)**: A 1-unit increase in **sp02** reduces the odds of mortality, as higher oxygen saturation levels are associated with better health outcomes. This suggests that patients with higher oxygen saturation are more likely to survive.
- **Risk Levels (risk)**: The relationship between Risk and mortality odds is exponential. As patients move from Risk1 (Minor) to Risk4 (Extreme), the odds of mortality increase dramatically. For example, a patient in the Risk4 (Extreme) category has odds of dying that are 167 times greater than a patient classified as Risk1 (Minor). This highlights the critical

importance of close monitoring and intervention for patients in the Risk4 category, as they face a significantly higher likelihood of death compared to those with lower risk levels.

While numerous published studies have established relationships between BMI, age, gender, severity, and mortality rate (Burns and Wholey, 1991; Engeland et al., 2003; Hauck and Hollingsworth, 2010; Kvamme et al., 2012), this model chose to exclude these variables as explanatory factors. Despite being initially considered, these variables were found to be statistically insignificant in explaining patient mortality within the context of this analysis.

b. Length of Stay model

After fitting a Negative Binomial model to estimate the impact of various predictors on Length of Stay (**los**), the key statistically significant variables were identified. The results are summarized in the table below.

The table includes:

- Coefficients from the model, which represent the effect of a 1-unit change in the predictor on the log of the expected **los**.
- Exponentiated coefficients, which provide a more intuitive interpretation as multiplicative changes in the expected **los** for a 1-unit change in the predictor.
- Effect of a 1-unit change in each variable, indicating whether the variable increases or decreases the expected length of stay.

This summary highlights the most influential factors affecting the number of days a patient stays in the hospital and provides a clear understanding of how each variable affects the predicted LOS.

Variables	Coefficients	Exponentiated	Effect
(Intercepts)	1.10657	3.02	Baseline
died	-0.53752	0.58	Decrease 42%
genderM	-0.13912	0.87	Decrease 13%
Severity2 – Moderate	0.29344	1.34	Increase 34%
Severity3 – Major	0.74725	2.11	Increase 111%
Severity4 - Extreme	1.41446	4.11	Increase 311%

From the results above, we can see that on average, when every other variable is zero, the number of days spent at hospital is 3.02. If a patient died, it would reduce the number of days stayed at the hospital by 42%. The result also showed that a male patient on average also spend less time staying at the hospital than a female patient by 13%. The severity level in this case is also having significant impact on the length of stay, as the severity level increases, the length of stay also increases exponentially.

4. Analysis evaluations

Both models presented variables that show statistically significance and contribute directly to the probability that a patient will die as well as the duration that a patient will stay at a hospital.

a. Mortality probability model

For the Mortality probability model, both Confusion Matrix and ROC curve, tools to assess the performance of a model, show good results:

Predicted \ Actual	**Actual Positive (1)**	**Actual Negative (0)**
Predicted Positive (1)	7	1
Predicted Negative (0)	26	944

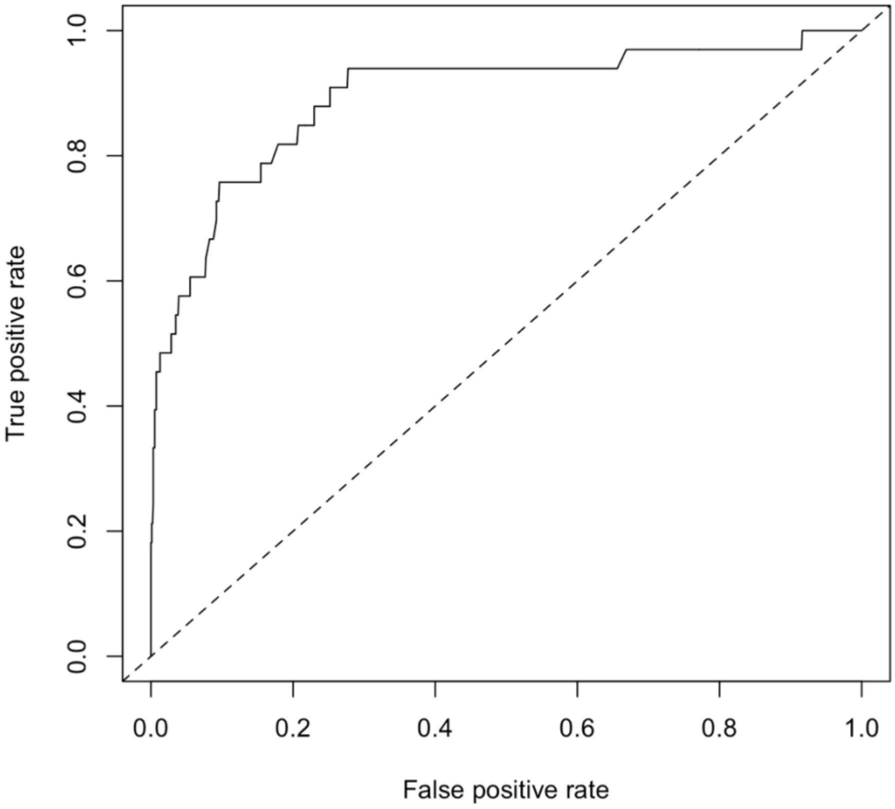


Figure 4. Mortality probability model ROC curve

The ROC curve here is well above the random classifier line (the dotted line) and curved towards the top left corner of the graph. This suggests that the model has successfully differentiated between positive and negative outcomes, which indicates high predictive power. One limitation of this model is that, while it performs well in predicting outcome for true negative, it is not doing so good for the true positive. Further tuning of the model as well as increase data collection on patient that died will help the model perform better at predicting positive outcome.

Further exploration into polynomial regression model could provide deeper insights into both the mortality probability of patients as well as the impacts of patients' health indicators on the length of hospital stay.

Other variables: temperature, blood pressure, respiratory, pulse rate, AVPU score are removed from the model due to statistically insignificant. However, consider that Body Temperature, Pulse Rate, Respiration Rate, Blood Pressure are vital signs of a patient and usually used to monitor or detect medical problems that may lead to fatality (Johns Hopkins Medicine, 2024), further exploration on their interaction terms could potentially provide more information on its effect on the probability of mortality.

b. Length of Stay:

The Negative Binomial model’s post diagnostic testing performed quite well showing little abnormality.

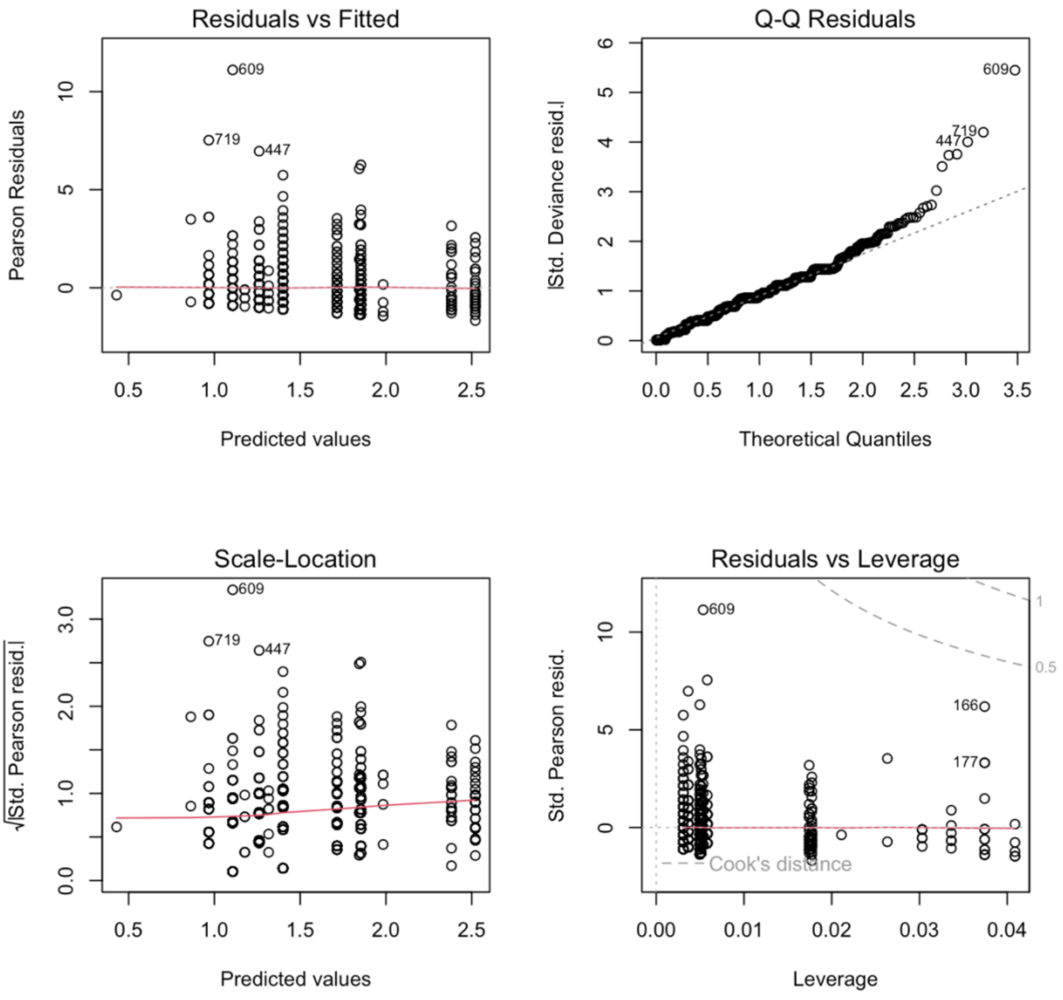


Figure 5. Post Diagnostic testing

The Residual vs Fitted plot shows no visible trends and the data is scattered around zero with random variance.

The Q-Q Residuals plot show a little deviation at the tail, but overall fits quite well with the normality distribution line

The Scale-Location and the Residuals vs Leverage plot also shows good results, with no influential points outside the Cook's distance that is impacting the analysis.

There are outliers spotted in the plots, which if removed, can show significant improvement of the model's fit.

It is important to note that several variables identified in previous published literatures as influencing the length of stay were excluded from this analysis due to their statistical insignificance. Future research could explore the potential relationships by incorporating interaction terms or utilizing a polynomial regression model to capture more complex dynamics.

References:

- Burns, L.R. and Wholey, D.R. (1991). The Effects of Patient, Hospital, and Physician Characteristics on Length of Stay and Mortality. *Medical Care*, [online] 29(3), pp.251–271. doi:https://doi.org/10.2307/3766013.
- Engeland, A., Bjørge, T., Selmer, R.M. and Tverdal, A. (2003). Height and Body Mass Index in Relation to Total Mortality. *Epidemiology*, [online] 14(3), pp.293–299. doi:https://doi.org/10.2307/3703849.
- Hauck, K. and Hollingsworth, B. (2010). The Impact of Severe Obesity on Hospital Length of Stay. *Medical Care*, [online] 48(4), pp.335–340. doi:https://doi.org/10.2307/27798453.
- Horn, S.D., Sharkey, P.D., Buckle, J.M., Backofen, J.E., Averill, R.F. and Horn, R.A. (1991). The Relationship between Severity of Illness and Hospital Length of Stay and Mortality. *Medical Care*, [online] 29(4), pp.305–317. doi:https://doi.org/10.2307/3765825.
- Johns Hopkins Medicine (2024). *Vital signs (body temperature, pulse rate, respiration rate, blood pressure)*. [online] Johns Hopkins Medicine Health Library. Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/vital-signs-body-temperature-pulse-rate-respiration-rate-blood-pressure [Accessed 5 Dec. 2024].
- Kvamme, J., Holmen, J., Wilsgaard, T., Florholmen, J., Midthjell, K. and Jacobsen, B.K. (2012). Body mass index and mortality in elderly men and women: the Tromsø and HUNT studies. *Journal of Epidemiology and Community Health (1979)*, [online] 66(7), pp.611–617. doi:https://doi.org/10.2307/23216030.
- Lindsey, H. (2022). *What Are Vital Signs, and Why Are They Important?* [online] Healthline. Available at: https://www.healthline.com/health/what-are-vital-signs [Accessed 5 Dec. 2024].