*Create a title page here*

**Analyzing the Hospital Discharge Status of Patients diagnosed with Burn Injury**

**1. Introduction**

In the realm of healthcare, where each decision can be a matter of life and death, the ability to predict patient outcomes accurately holds profound significance. This study is dedicated to the critical task of forecasting the likelihood of survival among patients who have endured the harrowing experience of burn injuries. The primary objective is to construct a predictive model that can effectively estimate the probability of a patient surviving a burn injury based on the selected set of variables. In essence, we aim to answer the pivotal question: What are the determinants and indicators that influence a patient's chances of survival following a burn injury? By employing sophisticated analytical techniques, we intend to unearth patterns, relationships, and insights that can assist healthcare professionals in making informed decisions and optimizing patient care strategies.

Burn injuries, often associated with excruciating pain and profound physical and psychological trauma, are a widespread occurrence across the globe. These injuries can result from various sources, including thermal, chemical, electrical, and radiation exposures. In our pursuit to predict patient survival following burn injuries, we are delving into a problem domain that intersects with a multitude of challenges. Burn injuries not only inflict severe pain but can also lead to complications like infections, organ failure, and long-term disabilities. Moreover, they impose a substantial economic burden on healthcare systems and societies, affecting millions of lives annually.

According to the World Health Organization (WHO), an estimated 265,000 deaths occur each year due to fires alone, with millions more experiencing non-fatal burn injuries, often resulting in varying degrees of disability. While advancements in burn care have improved survival rates, the precise determination of a patient's likelihood of survival remains a complex and vital matter. This analysis seeks to contribute to the ongoing efforts to enhance burn injury management by providing a data-driven approach to predict patient outcomes.

The chosen analysis approach, predictive modeling, is well-suited to the problem at hand. Predictive modeling allows us to harness the power of data to discern hidden patterns and relationships that might elude human observation. By training a predictive model on historical patient data, we can learn from past cases and extrapolate that knowledge to make informed predictions about future patients.

This approach is particularly relevant for burn injuries because it can take into account a multitude of variables – from patient demographics and injury severity to treatment protocols and comorbidities – to build a comprehensive understanding of survival factors. Furthermore, predictive modeling offers a quantitative and probabilistic framework, allowing us to quantify the uncertainty associated with our predictions, a critical aspect in clinical decision-making.

**2. Understanding the Data**

The dataset for this was selected from the module’s recommended datasets list. Here is the Definitions of the columns of the data.

**Table 1.0: Description of Data Variables**

|  |  |  |
| --- | --- | --- |
| **Variable names** | **Codes/Values** | **Description** |
| FACILITY | 1 - 40 | Burn Facility |
| DEATH | 0 = Alive (Survive)  1 = Dead (Not Survive) | Hospital Discharge Status |
| AGE | years | The infected person/patient’s age in years at admission |
| GENDER | 0 = Female  1 = Male | Gender of infected person/patient |
| RACE | 0 = Non-White  1 = White | Race of infected person/patient |
| TBSA | 0 to 100% | Total burn surface area |
| INH\_INJ | 0 = No  1 = Yes | Burn involved inhalation injury |
| FLAME | 0 = Flame absent  1 = Flame present | Flame involved in burn |

The target variable “DEATH” refers to the hospital discharge status of the patient. It is integer valued from 0 (Alive) and 1 (Dead). DEATH is the dependent variable and rests all the variable are the independent variables.

**2.1 Correlation**

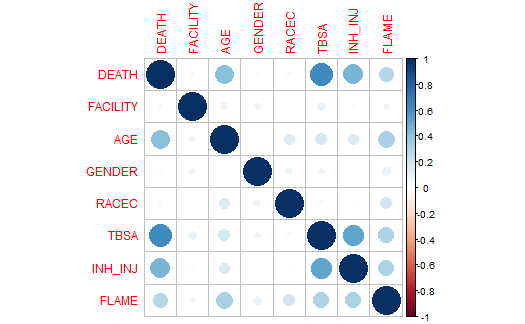
Correlations is used to determine whether and to what extent two variables are related linearly. Pearson's Correlation Coefficient, one of the most commonly used correlation measures, provides a numerical value that represents the strength and direction of this linear relationship. The correlation matrix in the figure below shows that the dependent variable is not/least related with FACILITY, GENDER, RACE as they lies between [0.2, 0]. Hence, we remove those variable from the dataset for further analysis.

There is a moderate positive correlation of 0.411 between DEATH and AGE. This suggests that older patients tend to have a higher likelihood of not surviving the burn injury.

There is a strong positive correlation of 0.629 between DEATH and TBSA. This suggests that as the percentage of total body surface area burned (TBSA) increases, the likelihood of not surviving the burn injury also increases.

DEATH and INH\_INJ have a strong positive correlation of 0.460, indicating that patients with inhalation injuries tend to have a higher likelihood of not surviving the burn injury.

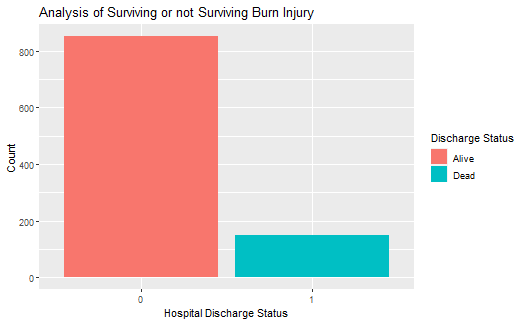
The correlation is 0.284, indicating a moderate positive relationship between DEATH and FLAME. This suggests that injuries caused by flame burns may be associated with a higher likelihood of not surviving the burn injury.



**Figure 2.1: Correlation Matrix of study variables**

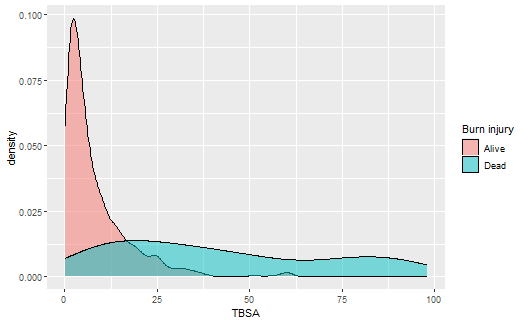
**2.2 Variable Analysis**

The number of people that survive the burn injury is more than with a number of people that did not survive it. We will perform further analysis to find out more about the relevant parameter that contributes to patients surviving the burn injury.



**Figure 2.2: Bar plot for hospital discharge status**

From the density graph below, we can observe that patients' chances of surviving the burn injury is significantly higher than and their chance of not surviving it base on the total burn surface area. This observation is inline with the association of this variable and the target variable observed earlier.



**Figure 2.2: Density graph for Total burn surface area**

**3. Results**

**3.1 Logistics Regression Model**

This regression model requires a set of covariates *X = X1, X2, …, Xk* with parameters *β = β1, β2, …, βk* in the model. Thus, the model is specified as,



where *pi* represents the probability of surviving the burn injury in a dichotomized variable setting with 0 “alive” and 1 “dead”. Then, *1 - Pi* is the probability of not surviving the burn injury. The log of the ratio  is the logit, which serves as the dependent variable in this modelling framework. Taking the exponent of both sides of the equation, it is very straightforward to conduct a maximum likelihood estimation in determining the estimates of the model. Thus, exp, exp, exp, …, exp are estimates of risk (odd-ratio) of the logit model. The regression modelling results are presented in Table 2.0.

**3.2 Model Implementation**

The result form the regression model shows the following;

* The coefficient estimate of the variable AGE is b = 0.0738, which is positive. This means that an increase in AGE is associated with increase in the probability of not surviving a burn injury.
* The coefficient estimate of the variable TBSA is b = 0.0731, which is positive. This means that an increase in TBSA is associated with increase in the probability of not surviving a burn injury.
* The coefficient estimate of the variable INH\_INJ is b = 1.5263, which is positive. This means that an increase in INH\_INJ is associated with increase in the probability of not surviving a burn injury.
* The coefficient estimate of the variable FLAME is b = 0.7646, which is positive. This means that an increase in FLAME is associated with increase in the probability of not surviving a burn injury.
* The age of infected person/patient’s at admission and the total burn suface area are the most significant variables for predicting if a patient will survive the burn injury or not.
* The AIC value is used to state the goodness of the model. The smaller the value of AIC the better the model performing.

**Table 2.0: Logistic regression results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **β** | **S.E (β)** | **Exp (β)** | **P value** |
| Intercept | -7.7548 | 0.8868 | -8.7452 | < 0.0000\*\*\* |
| AGE | 0.0738 | 0.0110 | 1.0766 | 0.0000\*\*\* |
| TBSA | 0.0731 | 0.0107 | 1.0759 | 0.0000\*\*\* |
| INH\_INJ | 1.5263 | 0.5000 | 4.6011 | 0.0023\*\* |
| FLAME | 0.7646 | 0.5183 | 2.1481 | 0.1402 |

AIC: 187.126

**3.3 Evaluation of the Model**

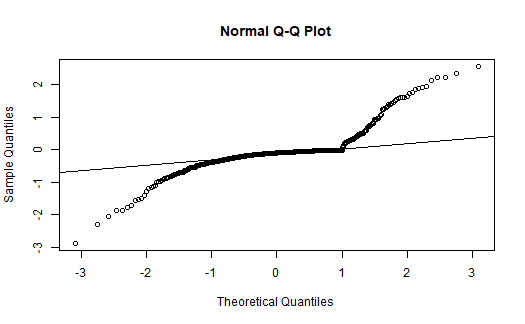
The model is observed to be 92% accurate, with 98% and 58% Sensitivity and Specificity respectively. 421 values of o's predicted correctly and 41 values of 1's is predicted correctly out of total 500 values. We can analyse from the above output that the number of patients surviving a burn injury is more than those who do not survive it.

**Table 3.0: Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | *Actual* | |
|  |  | 0 | 1 |
| *Prediction* | 0 | 421 | 30 |
| 1 | 8 | 41 |

**3.4 Examining Predicted vs. Residual (The Residual Plot)**

The model for the charts below display a good predictions.



**Figure 3.0: Normal Q-Q Residual Plots**

**4. Conclusions**

**4.1 Summary**

In summary, the regression model results indicate several key findings: First, positive coefficient estimates for variables AGE, TBSA, INH\_INJ, and FLAME suggest that an increase in these factors corresponds to an increased probability of not surviving a burn injury. Second, it's observed that the age of the patient at admission and the total burn surface area (TBSA) are the most influential variables for predicting burn injury survival. Finally, the use of the AIC (Akaike Information Criterion) value as a model assessment metric highlights the model's performance, with a smaller AIC indicating a better-performing model.

**4.2 Limitations of the study**

Some of the limitations of the study are:

* Limited number of variables in the data**:** This may result in a loss of valuable information that could have contributed to a more comprehensive understanding of the burn study.
* Due to no association or variable correlation in the burn study data, some variables are not included in the analysis.

**4.3 Recommendation/Improvement Areas:**

* There should be increased efforts to ensure adequate keeping of information or updating of database of burn injury diagnosis for future analysis.
* Efforts should be made to increase awareness and proper documentation of other factors that may increase or reduce the risk of burn injury.

**References**

Achia T, Wangombe A, Khadioli N 2010. A logistic regression model to identify key determinants of poverty using demographic and health survey data. European Journal of Social Sciences, 13(1): 38- 45.

**Appendix**

library('corrplot')

## corrplot 0.92 loaded

library('ggplot2')  
library('tidyverse')

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.2 ✔ tibble 3.2.1  
## ✔ purrr 1.0.2 ✔ tidyr 1.3.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library('caret')

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library('caTools')

# Importing the data file  
data <- read.csv('burn.csv')

# DISPLAY THE FIRST FEW ROWS OF DATA  
head(data)

## DEATH FACILITY AGE GENDER RACEC TBSA INH\_INJ FLAME  
## 1 0 11 26.6 1 1 25.3 0 1  
## 2 0 1 2.0 0 0 5.0 0 0  
## 3 0 12 22.0 0 0 2.0 0 0  
## 4 0 1 37.3 1 1 2.0 0 0  
## 5 0 1 52.1 1 1 6.0 0 1  
## 6 0 6 50.2 1 1 7.0 0 0

# Understanding the datatype of dataset  
str(data)

## 'data.frame': 1000 obs. of 8 variables:  
## $ DEATH : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ FACILITY: int 11 1 12 1 1 6 22 1 1 1 ...  
## $ AGE : num 26.6 2 22 37.3 52.1 50.2 2.5 53.8 31.9 41.1 ...  
## $ GENDER : int 1 0 0 1 1 1 0 0 1 1 ...  
## $ RACEC : int 1 0 0 1 1 1 0 1 1 1 ...  
## $ TBSA : num 25.3 5 2 2 6 7 7 0.9 2 22 ...  
## $ INH\_INJ : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ FLAME : int 1 0 0 0 1 0 0 1 0 1 ...

# DISPLAY THE SUMMARY  
summary(data)

## DEATH FACILITY AGE GENDER RACEC   
## Min. :0.00 Min. : 1.00 Min. : 0.10 Min. :0.000 Min. :0.000   
## 1st Qu.:0.00 1st Qu.: 2.00 1st Qu.:10.85 1st Qu.:0.000 1st Qu.:0.000   
## Median :0.00 Median : 8.00 Median :31.95 Median :1.000 Median :1.000   
## Mean :0.15 Mean :11.56 Mean :33.29 Mean :0.705 Mean :0.589   
## 3rd Qu.:0.00 3rd Qu.:18.25 3rd Qu.:51.23 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :1.00 Max. :40.00 Max. :89.70 Max. :1.000 Max. :1.000   
## TBSA INH\_INJ FLAME   
## Min. : 0.10 Min. :0.000 Min. :0.000   
## 1st Qu.: 2.50 1st Qu.:0.000 1st Qu.:0.000   
## Median : 6.00 Median :0.000 Median :1.000   
## Mean :13.54 Mean :0.122 Mean :0.529   
## 3rd Qu.:16.00 3rd Qu.:0.000 3rd Qu.:1.000   
## Max. :98.00 Max. :1.000 Max. :1.000

cor <- round(cor(data[1:8]),3)  
cor

## DEATH FACILITY AGE GENDER RACEC TBSA INH\_INJ FLAME  
## DEATH 1.000 0.033 0.411 -0.029 0.038 0.629 0.460 0.284  
## FACILITY 0.033 1.000 0.078 0.062 -0.005 0.098 0.024 0.051  
## AGE 0.411 0.078 1.000 -0.012 0.146 0.173 0.156 0.325  
## GENDER -0.029 0.062 -0.012 1.000 0.079 0.062 0.007 0.092  
## RACEC 0.038 -0.005 0.146 0.079 1.000 0.038 0.020 0.181  
## TBSA 0.629 0.098 0.173 0.062 0.038 1.000 0.522 0.310  
## INH\_INJ 0.460 0.024 0.156 0.007 0.020 0.522 1.000 0.315  
## FLAME 0.284 0.051 0.325 0.092 0.181 0.310 0.315 1.000

# Displaying the correlation matrix  
corr <- cor(data)  
# Visualize the correlation matrix  
corrplot(corr)

A diagram of a number of people

Description automatically generated with medium confidence

# Deleting not related variables  
data = subset(data, select = c(-FACILITY,-GENDER,-RACEC))

# Converting the integer data to factor  
data$DEATH <- as.factor(data$DEATH)  
data$INH\_INJ <- as.factor(data$INH\_INJ)  
data$FLAME <- as.factor(data$FLAME)

# Bar plot for DEATH (hospital discharge status)   
data$DEATH <- as.factor(data$DEATH)  
ggplot(data, aes(x=data$DEATH, fill=data$DEATH)) +   
 geom\_bar() +  
 xlab("Hospital Discharge Status") +  
 ylab("Count") +  
 ggtitle("Analysis of Surviving or not Surviving Burn Injury") +  
 scale\_fill\_discrete(name = "Discharge Status", labels = c("Alive", "Dead"))

## Warning: Use of `data$DEATH` is discouraged.  
## ℹ Use `DEATH` instead.  
## Use of `data$DEATH` is discouraged.  
## ℹ Use `DEATH` instead.

A graph of a number of patients

Description automatically generated

## Age variable Analysis

# Group the different ages in three groups (young, middle, old)  
young <- data[which((data$AGE<10)), ]  
middle <- data[which((data$AGE>=10)&(data$AGE<50)), ]  
elderly <- data[which(data$AGE>50), ]  
groups <- data.frame(age\_group = c("young","middle","elderly"), group\_count = c(NROW(young), NROW(middle), NROW(elderly)))  
  
#ploting different age groups  
ggplot(groups, aes(x=age\_group, y=group\_count, fill=age\_group)) +   
 ggtitle("Age Analysis") +  
 xlab("Age Group") +  
 ylab("group Count") +  
 geom\_bar(stat="identity") +  
 scale\_fill\_discrete(name = "Age Group", labels = c("Elderly", "Middle", "Young"))

A graph with different colored squares

Description automatically generated

# Density graph for TBSA (Total burn surface area)  
ggplot(data, aes(x = TBSA, fill = DEATH)) +  
 geom\_density(alpha=0.5) +  
 scale\_fill\_discrete(name = "Burn injury", labels = c("Alive", "Dead"))

A graph of a graph showing the amount of burn fat

Description automatically generated with medium confidence

data <- data[, c(1, 2, 3, 4, 5)]  
  
# Dividing the data set in train and test datasets  
dataSample <- sample.split(data[,ncol(data)-1], SplitRatio=0.50)  
trainSet = subset(data,dataSample == TRUE)  
testSet = subset(data,dataSample == FALSE)

# Creating a logistic model  
logisticmodel <- glm(DEATH~.,data = trainSet, family = "binomial")  
  
# Summary of the created model  
summary(logisticmodel)

##   
## Call:  
## glm(formula = DEATH ~ ., family = "binomial", data = trainSet)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.98840 0.90259 -8.851 < 2e-16 \*\*\*  
## AGE 0.07421 0.01105 6.717 1.85e-11 \*\*\*  
## TBSA 0.09003 0.01228 7.331 2.28e-13 \*\*\*  
## INH\_INJ1 1.17303 0.50635 2.317 0.0205 \*   
## FLAME1 0.90757 0.50157 1.809 0.0704 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 422.71 on 499 degrees of freedom  
## Residual deviance: 171.00 on 495 degrees of freedom  
## AIC: 181  
##   
## Number of Fisher Scoring iterations: 7

expb <- exp(coef(logisticmodel))  
expb

## (Intercept) AGE TBSA INH\_INJ1 FLAME1   
## 0.0003393762 1.0770283616 1.0942051756 3.2317803407 2.4782852979

# Making prediction with the above model  
predictdata <- predict(logisticmodel, newdata = testSet[, -1], type="response")  
pred <- ifelse(predictdata>=0.5,1,0)  
pred <- as.factor(pred)  
observed <- testSet[,1]

# Checking the accuracy of the model  
confusionMatrix(pred, observed)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 412 23  
## 1 13 52  
##   
## Accuracy : 0.928   
## 95% CI : (0.9017, 0.9491)  
## No Information Rate : 0.85   
## P-Value [Acc > NIR] : 7.327e-08   
##   
## Kappa : 0.7012   
##   
## Mcnemar's Test P-Value : 0.1336   
##   
## Sensitivity : 0.9694   
## Specificity : 0.6933   
## Pos Pred Value : 0.9471   
## Neg Pred Value : 0.8000   
## Prevalence : 0.8500   
## Detection Rate : 0.8240   
## Detection Prevalence : 0.8700   
## Balanced Accuracy : 0.8314   
##   
## 'Positive' Class : 0   
##

# Residual plot

#get list of residuals  
res <- resid(logisticmodel)

#produce residual vs. fitted plot  
plot(fitted(logisticmodel), res)

A graph of a function

Description automatically generated with medium confidence

#create Q-Q plot for residuals  
qqnorm(res)  
qqline(res)

A graph with a line

Description automatically generated