Group Members: **DS-A22-cohort**

Kevin SIMO

Nathanaël DEFO TOTOUOM

Webana T. Julien AGA

M amadou Saliou DIALLO

Survival Analysis Project Report

Introduction

This report outlines the statistical survival analysis conducted on a healthcare dataset. The purpose is to understand how different variables, such as age, gender, and medical condition, influence the survival time of patients. The analysis includes nonparametric estimation of survival for different medical conditions, nonparametric comparison between groups, and a semi-parametric Cox regression to evaluate the effects of various factors on survival time.

I. Data preprocessing

The dataset contains columns such as Survival.Time, Date.of.Admission, Discharge.Date, Gender, Medical.Condition and Age and there were no missing values identified.

a. Validation of the variable Survival.Time:

- **Negative Values**: No negative values were found in the Survival.Time column.
- **Unrealistic Values**: Survival times greater than 365 days were considered unrealistic. However, no such values were found in the dataset.

- **Inconsistencies**: A comparison between the calculated survival time (difference between Discharge.Date and Date.of.Admission) and the provided Survival.Time showed no discrepancies, indicating that the survival times are accurate.

b. **Data Type Conversion :**

Key variables, such as survival time (Survival.Time) and medical condition (Medical.Condition) and and Admission.Type are converted into appropriate data formats for analysis. These preparations are essential to ensure the accuracy of the statistical analyses.

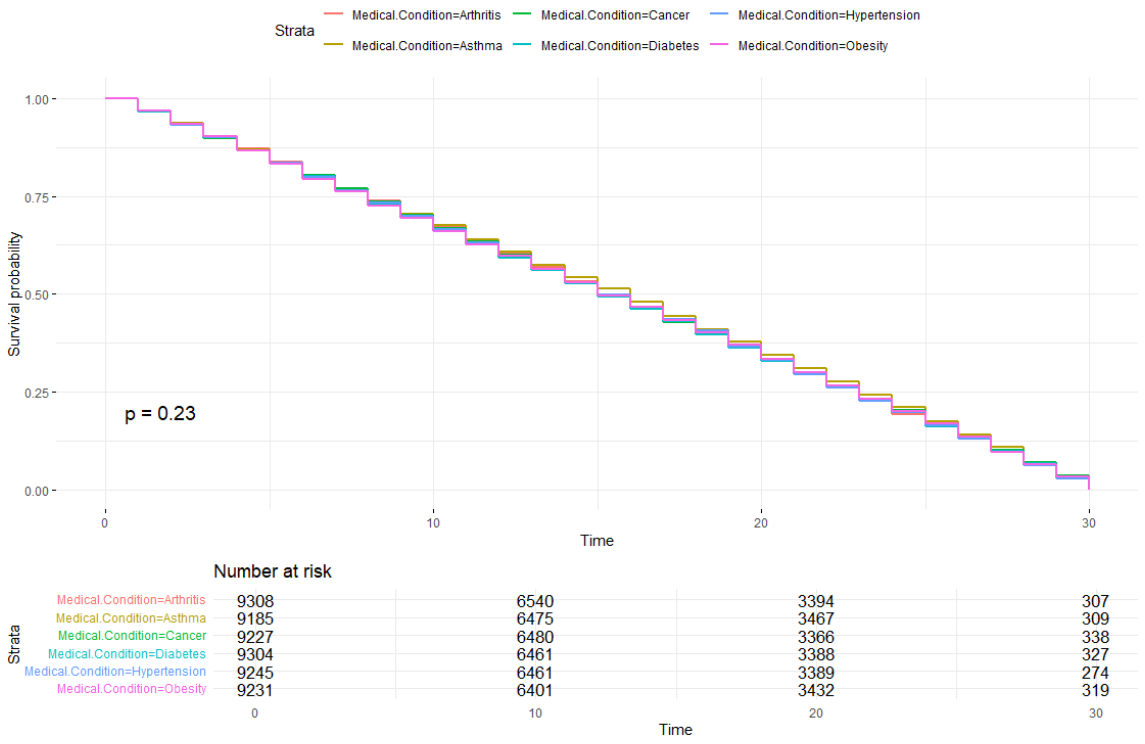
II. Data Modeling

In this section, we will discuss, we will use Non-Parametric Estimation of Survival for One or More Groups, Non-Parametric Comparison of Two or More Groups and Cox Proportional Hazards Model.

1. Non-Parametric Estimation of Survival for One or More Groups

Non-parametric survival estimation was performed using the Kaplan-Meier method. This method is appropriate for estimating the survival function of patients without assuming a specific parametric distribution for survival time. It allows visualization of the probability of survival at different time points for different groups of patients based on their medical condition.

The Kaplan-Meier curve provides a visual representation of survival functions.



The survival curves for different medical conditions (Arthritis, Asthma, Cancer, Diabetes, Hypertension, and Obesity) were plotted.

Log-Rank Test: The p-value from the log-rank test was 0.2, indicating no statistically significant difference in survival times across the different medical conditions.

2. Non-Parametric Comparison of Two or More Groups

Log-Rank Test for Group Comparison

The log-rank test was used to compare survival distributions across different medical conditions. The test results are as follows:

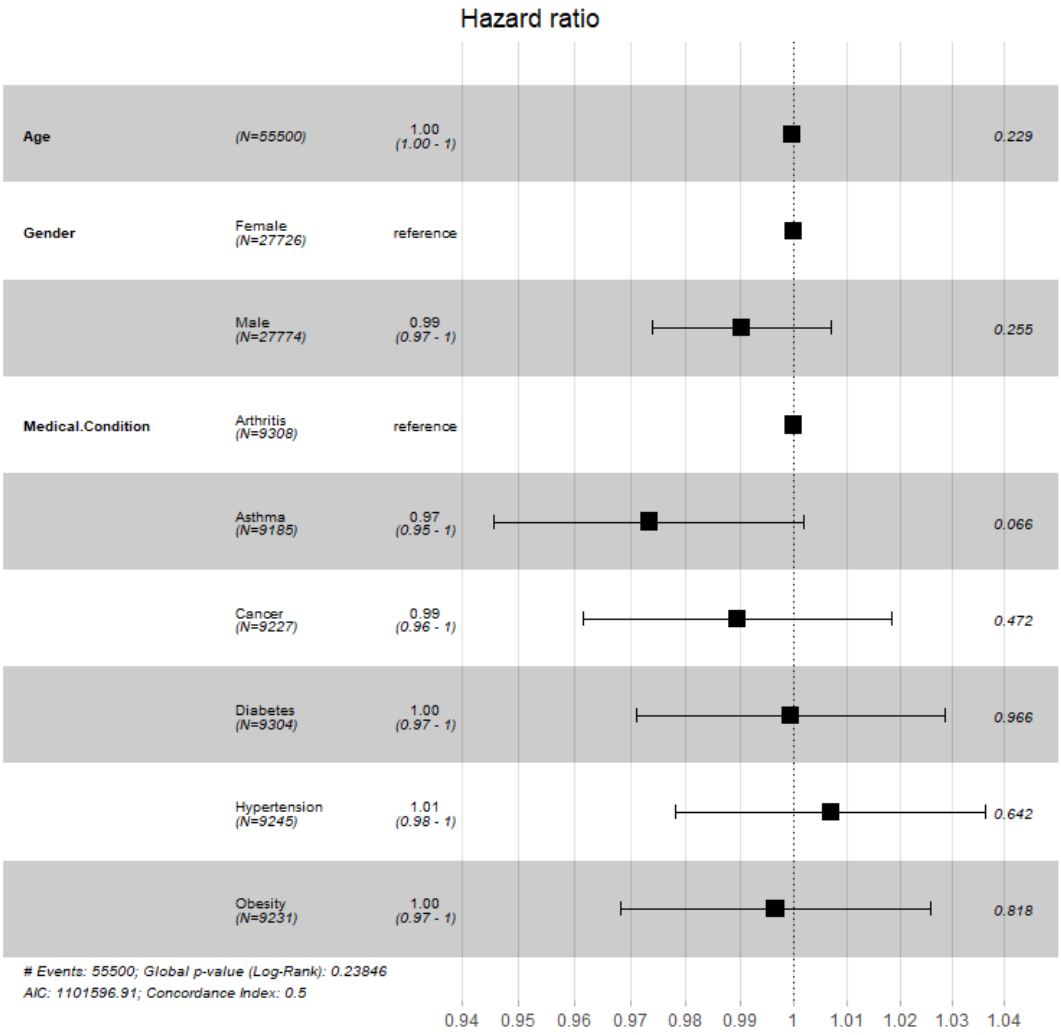
```
> print(logrank_test)
Call:
survdiff(formula = Surv(Survival.Time) ~ Medical.Condition, data = df)

                                  N Observed Expected (O-E)^2/E (O-E)^2/V
Medical.Condition=Arthritis    9308     9308     9257    0.2777    0.3840
Medical.Condition=Asthma       9185     9185     9380    4.0725    5.6598
Medical.Condition=Cancer       9227     9227     9260    0.1192    0.1656
Medical.Condition=Diabetes     9304     9304     9250    0.3178    0.4407
Medical.Condition=Hypertension 9245     9245     9145    1.0854    1.4935
Medical.Condition=Obesity      9231     9231     9207    0.0632    0.0874

 Chisq= 6.8  on 5 degrees of freedom, p= 0.2
>
```

- **Chi-Square Statistic**: 6.8 on 5 degrees of freedom.
- **P-Value**: 0.2, indicating that there is no statistically significant difference in survival across the groups.

3. Cox Proportional Hazards Model

A Cox Proportional Hazards Model was fitted to evaluate the effect of age, gender, and medical conditions on the survival time.

```
> summary(cox_fit)
Call:
coxph(formula = Surv(Survival.Time) ~ Age + Gender + Medical.Condition,
    data = df)

  n= 55500, number of events= 55500

                                    coef  exp(coef)   se(coef)      z Pr(>|z|)
Age                           -0.0002607  0.9997394  0.0002166 -1.203   0.2288
GenderMale                    -0.0096686  0.9903780  0.0084902 -1.139   0.2548
Medical.ConditionAsthma       -0.0269945  0.9733666  0.0147083 -1.835   0.0665 .
Medical.ConditionCancer       -0.0105570  0.9894986  0.0146912 -0.719   0.4724
Medical.ConditionDiabetes     -0.0006172  0.9993830  0.0146603 -0.042   0.9664
Medical.ConditionHypertension  0.0068347  1.0068581  0.0146841  0.465   0.6416
Medical.ConditionObesity      -0.0033796  0.9966261  0.0146894 -0.230   0.8180
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                              exp(coef) exp(-coef) lower .95 upper .95
Age                              0.9997     1.0003    0.9993     1.000
GenderMale                       0.9904     1.0097    0.9740     1.007
Medical.ConditionAsthma          0.9734     1.0274    0.9457     1.002
Medical.ConditionCancer          0.9895     1.0106    0.9614     1.018
Medical.ConditionDiabetes        0.9994     1.0006    0.9711     1.029
Medical.ConditionHypertension    1.0069     0.9932    0.9783     1.036
Medical.ConditionObesity         0.9966     1.0034    0.9683     1.026

Concordance= 0.504  (se = 0.001 )
Likelihood ratio test= 9.2  on 7 df,   p=0.2
Wald test            = 9.18  on 7 df,   p=0.2
Score (logrank) test = 9.18  on 7 df,   p=0.2

>
```



- **Model Summary**:
  - The hazard ratios (exp(coef)) for each factor were close to 1, indicating a minimal effect on the survival time.
  - **Age**: Hazard ratio of 0.9997 suggests that age has a negligible impact on survival time.
  - **Gender**: Being male (reference category for Gender) has a hazard ratio of 0.9904, also indicating a minimal effect.
  - **Medical Conditions**: None of the medical conditions showed a statistically significant impact on survival time, with p-values greater than 0.05 for all conditions.
- **Concordance**: The model's concordance index was 0.504, suggesting that the model has limited predictive power.

- **Overall Model Significance**:
    - Likelihood Ratio Test: p = 0.2
    - Wald Test: p = 0.2
    - Score (log-rank) Test: p = 0.2

All tests indicate that the predictors in the model do not significantly explain the variation in survival times.

4. Visual Representation

- **Kaplan-Meier Curve**: The curve shows similar survival probabilities across different medical conditions, reinforcing the results from the log-rank test.
- **Cox Model Plot**: A forest plot of the Cox model coefficients was generated to visualize the effect sizes of different factors.

Conclusion

The survival analysis on this dataset shows that there is no statistically significant difference in survival times across different medical conditions. The Cox Proportional Hazards Model also suggests that age, gender, and medical conditions do not significantly impact survival time.

The results indicate that other factors not captured in this dataset might be influencing survival, or the dataset may lack sufficient variation in these variables to detect differences.