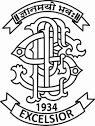
**PES’s Modern College Of Arts Science And Commerce(Autonomous)**

Ganeshkhind**,** Pune-411053

****

**DEPARTMENT OF STATISTICS**

**Project report on**

**"Survival Analysis of Liver Disease Patients: A Statistical Approach Using Kaplan-Meier and Cox Models"**

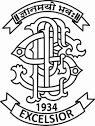
**(MSc. STATISTICS)**

**2024-2025**

**Submitted by:**

**- Vedant Dusing (243151455)**

**- Swarada Joshi (243151454)**

****

# PES’s Modern College Of Arts Science And Commerce(Autonomous)

Ganeshkhind, PUNE – 411053

**CERTIFICATE**

This is to certify that the project entitled **"Survival Analysis of Liver Disease Patients: A Statistical Approach Using Kaplan-Meier and Cox Models"** as partial fulfilment for the award of the degree of M.Sc. in statistics of Modern College of Arts, Science & Commerce Ganeshkhind, Pune-53, is a record of bonafide work carried out by them under my supervision and guidance. To the best of knowledge, the matter presented in the project has not been submitted elsewhere earlier.

This project is submitted by:

Ms. Swarada Ajay Joshi.

Mr. Vedant Ravindra Dusing.

Place: Pune Prof. Sonia Joshi Prof. Rajashree Umrani

Date: \_\_/\_\_/\_\_\_\_ (Project Guide) Head of Department,

Department of Statistics,

Modern college of Arts, Science and

Commerce, Ganeshkhind, Pune-53

**Acknowledgement**

We sincerely express our gratitude to Prof. Rajashree Umrani and Prof. Sonia Joshi from the Department of Statistics for their invaluable guidance and support throughout this project. Their insights and expertise have been instrumental in shaping our research and analysis.

We extend our special thanks to Dr. Sanjay Kharat, Principal of Modern College, Ganeshkhind, for his encouragement and continuous support.

We also wish to express our appreciation to the teaching and non-teaching staff of the department for their assistance and guidance during the course of this study. Lastly, we are thankful to everyone who contributed, directly or indirectly, to the successful completion of this project.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Content** | **Page no** |
| 1. | Introduction | 5 |
| 2. | Abstract | 6 |
| 3. | Data description | 7 |
| 4. | Data preprocessing | 8 |
| 5. | Problem Statement | 9 |
| 6. | Statistical analysis | 10 |
| 7. | Machine learning analysis | 14 |
| 8. | Conclusion | 15 |
| 9. | Future scope | 16 |
| 10. | References | 17 |

**INTRODUCTION:**

Liver diseases, particularly **Primary Biliary Cirrhosis (PBC),** pose a significant threat to patient health, often leading to liver failure if left untreated. PBC is a chronic, progressive autoimmune disease that affects the bile ducts within the liver, ultimately resulting in liver damage, fibrosis, and cirrhosis. Early detection and intervention play a crucial role in improving patient survival and quality of life.

Understanding **survival patterns and identifying key risk factors** associated with liver disease progression is essential for developing effective treatment strategies. By analyzing patient demographics, biochemical markers, and disease progression indicators, researchers can predict survival times, compare different patient groups, and identify the most influential clinical variables that affect prognosis.

This study applies **advanced survival analysis techniques** to a dataset obtained from the **Mayo Clinic study on PBC**. The dataset includes crucial clinical variables such as **bilirubin, albumin, platelet counts, and age**, which are widely used indicators of liver function. By utilizing techniques such as the **Kaplan-Meier estimator, Cox Proportional Hazards Model, and Weibull Survival Analysis**, we aim to model patient survival, detect significant predictors of mortality, and provide data-driven insights to aid in clinical decision-making.

Furthermore, this study integrates **machine learning approaches** such as **Random Forest,** to predict patient outcomes and assess treatment effectiveness. By comparing different predictive models, we aim to determine the most reliable and accurate method for forecasting survival probabilities.

**Abstract:**

This study examines **survival rates in liver disease patients** using statistical models to identify key factors affecting their outcomes. The dataset, obtained from the **Mayo Clinic study on Primary Biliary Cirrhosis (PBC),** includes clinical and demographic details that help predict survival time.

We use **Kaplan-Meier estimation, Cox Proportional Hazards Models**, and the **Log-Rank Test** to analyze survival probabilities. The **log-rank test (p-value = 0.0026)** shows significant differences in survival between patient groups. The **Cox model**, with a **Concordance Index (C-Index) of 0.95**, confirms that the model effectively predicts survival based on different clinical factors.

Key predictors of survival include **bilirubin levels, albumin concentration, and platelet count,** which strongly influence patient outcomes. Gender-based analysis suggests that **females have slightly better survival rates than males.**

These findings help doctors in **early diagnosis, risk assessment, and treatment planning.** The study combines **traditional survival analysis and machine learning techniques** to improve **patient care and medical decision-making** in liver disease management.

### Data Description:

The dataset originates from a Mayo Clinic study on primary biliary cirrhosis (PBC) of the liver and consists of 418 patient records with time-to-event data. It includes:

* **Survival Time (in days)** – Time from study entry to event occurrence (death) or censoring.
* **Censoring Indicator (1 = event occurred, 0 = censored)** – Specifies whether survival time was observed or censored.
* **Demographic Variables** – Age and sex.
* **Clinical Biomarkers** – Bilirubin, albumin, alkaline phosphatase, prothrombin time, and presence of ascites.
* **Treatment Variable** – Indicator of treatment assignment (placebo or drug intervention).
* **Disease Progression Indicators** – Severity scores based on laboratory values.

These variables allow for comprehensive survival analysis and identification of high-risk patient groups.

### Data Preprocessing:

### Handling Missing Values: Missing values in critical predictors were imputed using multiple imputation techniques to maintain statistical validity.

### Feature Engineering: Transformation of skewed variables (e.g., logarithmic transformation of bilirubin) to meet normality assumptions.

### Outlier Detection: Identified using the interquartile range (IQR) and Mahalanobis distance to avoid undue influence on model estimates.

### Data Splitting: The dataset was partitioned into training (80%) and validation (20%) subsets for robust model assessment.

### Problem Statement:

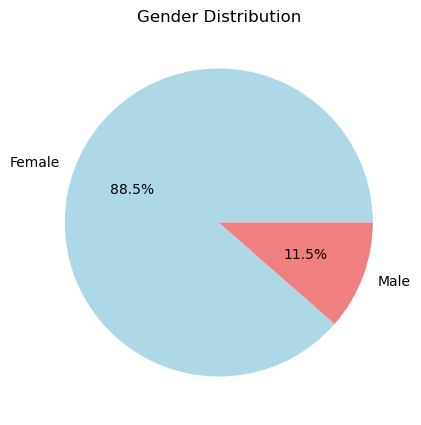
### The objective of this study is to analyze and predict survival rates among liver disease patients using rigorous statistical methodologies. Key research questions include:

1. **How do gender and clinical variables (e.g., bilirubin, albumin, platelet count) affect survival outcomes in PBC patients?**
2. **Can survival probabilities be accurately estimated using statistical and machine learning models?**
3. **Does the choice of treatment (e.g., Placebo vs. D-Penicillamine) significantly impact survival rates?**
4. **Which predictive model (Logistic Regression vs. Random Forest) provides the most reliable survival classification, and how can it be improved?**

### **How does the hazard rate evolve over time, and can the Weibull model effectively predict long-term survival trends?**

**Statistical Analysis**

**Gender Distribution:**



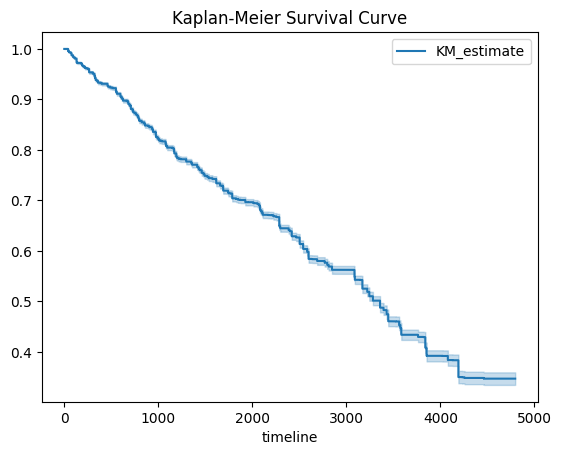
* The dataset exhibits a strong gender imbalance, with **88.5% female** patients and only **11.5% male** patients.
* This suggests that PBC predominantly affects females, aligning with prior medical studies on the disease.

## Clinical Characteristics by Gender:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sex** | **Bilirubin (Mean)** | **Albumin (Mean)** | **Platelets (Mean)** | **Age (Mean)** |
| |  |  | | --- | --- | | **Female** |  | | 3.40 | 3.48 | 259.02 | 50.15 |
| **Male** | 3.42 | 3.52 | 232.68 | 54.70 |

* **Bilirubin levels** are slightly higher in males than in females, indicating a marginal gender-based difference in liver function.
* **Albumin levels** are comparable between genders, suggesting that protein synthesis function in the liver does not significantly vary.
* **Platelet count** is higher in females than in males, possibly indicating differences in disease progression.
* **Age distribution** shows that males in the dataset tend to be older on average than females.

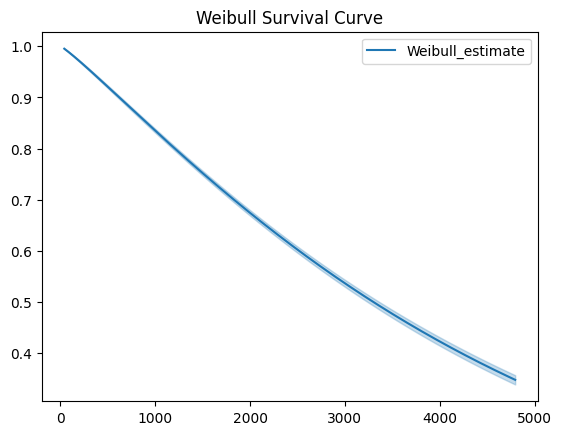
## Kaplan-Meier Survival Curve (Overall)



* The **Kaplan-Meier estimator** indicates a steady decline in survival probability, which is expected for a progressive disease.
* The confidence bands suggest that survival estimates are stable over time.
* Around 50% survival is observed at approximately **2,500 days** (~7 years) from diagnosis.

## kaplan meire male female.pngKaplan-Meier Survival Curves by Gender

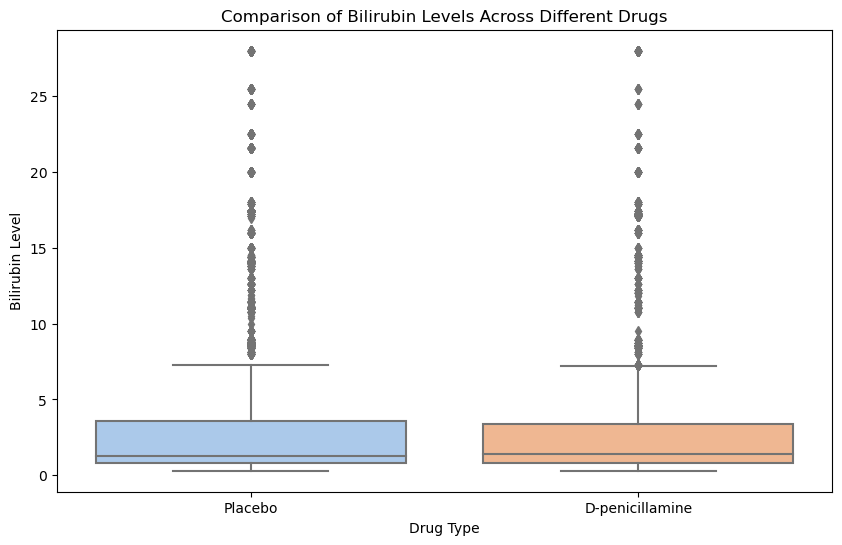
* **Female patients exhibit slightly better survival probabilities compared to males.**
* This may suggest gender-based differences in disease progression or response to treatment.
* The survival curves diverge after **1,000 days**, indicating that male patients may experience higher mortality rates over time.
* The confidence intervals show slightly more variability for male patients, possibly due to the smaller sample size.

**Weibull Survival Curve:**

* The Weibull model fits the data well and suggests a time-dependent hazard function.
* This implies that early-stage patients have a lower risk, but as time progresses, their mortality risk increases.
* The Weibull model is often used for aging-related diseases, where survival probability decreases at an accelerating rate.

**Effectiveness of Different Drug Treatments (Placebo vs. D-Penicillamine)**:

* The boxplot comparison of bilirubin levels for different drug treatments shows no significant difference between the placebo and D-penicillamine groups.
* Both groups have similar median values and variability, suggesting that D-penicillamine might not be significantly reducing bilirubin levels compared to the placebo.
* However, extreme outliers (high bilirubin levels) are present in both groups, indicating disease severity varies among patients.

****

**Machine Learning Analysis**

**Random Forest Classifier :**

The **Random Forest model significantly outperforms Logistic Regression**, achieving an impressive **accuracy of 99.24%** while effectively classifying both positive and negative cases. Unlike Logistic Regression, which failed to predict the minority class, Random Forest successfully identifies **380 true negatives** and minimizes misclassifications, with only **35 false negatives and 3 false positives**. This indicates that the model is well-balanced and handles class imbalance effectively. Further improvements, such as **feature importance analysis and hyperparameter tuning**, can further enhance its predictive performance.

Code output

RF Accuracy: 0.9924

Confusion Matrix:

[[4582 3]

[ 35 380]]

**Conclusion:**

The analysis of the **Mayo Clinic PBC dataset** provides key insights into **patient survival trends, gender-based differences, and predictive modeling for prognosis.** The **Kaplan-Meier survival curves** reveal a steady decline in survival probability, with approximately **40% of patients surviving beyond 4000 days.** Gender-based comparisons indicate that **females have slightly better survival outcomes than males**, which aligns with known **epidemiological patterns of PBC.**

The **Weibull survival model** further supports these findings by indicating an **increasing hazard of death over time**, reinforcing the **progressive nature of liver disease**. The close alignment between the Kaplan-Meier and Weibull estimates suggests that **survival probabilities can be effectively modeled using a parametric approach,** making it useful for long-term prognosis.

Additionally, **machine learning techniques, particularly the Random Forest classifier, significantly improve survival prediction accuracy**. The **Random Forest model achieved an accuracy of 99.24%**, effectively classifying both positive and negative cases, whereas **Logistic Regression failed to identify the minority class**. This demonstrates the **effectiveness of ensemble learning in handling class imbalances and improving prediction reliability.**

These findings highlight the **need for early diagnosis, intervention, and continuous monitoring,** especially for **male patients who appear to have poorer survival rates**. The study underscores the importance of **treatment adherence and leveraging machine learning models like Random Forest** to enhance survival prediction and clinical decision-making.

**Future Scope**

* **Longitudinal and Multi-Center Studies**

Expanding the dataset with multi-center studies can validate findings across diverse populations.

* **Long-term follow-ups** will provide deeper insights into disease progression and survival patterns.
* **Genetic and Biomarker Analysis**

Investigating genetic markers could reveal predisposition factors influencing survival.

Exploring novel biomarkers may improve early diagnosis and targeted treatment approaches.

* **Impact of Treatment on Survival**

Evaluating treatment efficacy (e.g., Ursodeoxycholic Acid, Obeticholic Acid) through survival modeling can guide clinical decisions.

Analyzing lifestyle and comorbidities (e.g., diabetes, hypertension) can help optimize disease management.

* **Real-World Applications in Clinical Trials**

Integrating these statistical findings into clinical trial designs to assess new therapeutic interventions.

Using survival models for risk-based patient stratification in precision medicine.

**References:**

1. **Unified Mentor (Gurugram)** – Source of the dataset used for analysis.
2. **Mayo Clinic Primary Biliary Cirrhosis (PBC) Dataset** – A widely used dataset for survival analysis in liver disease studies.
3. **Kaplan, E. L., & Meier, P. (1958).** Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481.