A Survival Analysis Study on German Breast Cancer Data

Mehak Malik (mm3646)
Shubham Patil (sp2484)

22 April 2024

Abstract

This study focuses on analyzing data from a German breast cancer study conducted in 1989 to understand factors influencing time to death and cancer recurrence in patients. The dataset comprises 686 observations with 16 variables, including patient demographics, tumor characteristics, and treatment information. Survival analysis techniques, including Kaplan-Meier estimation and Cox proportional hazards regression, were employed to analyze the data. The study aimed to identify covariates predictive of survival time and recurrence risk while ensuring the validity of assumptions underlying the chosen methods. Key findings indicate that age, tumor grade, tumor size, number of positive lymph nodes, and progesterone receptor count significantly influence survival time. Additionally, the study confirms the importance of validating assumptions in survival analysis models for robust results. Overall, the Cox proportional hazards model emerged as the most suitable for this dataset, providing valuable insights into breast cancer prognosis and informing future research and treatment strategies.

Introduction

Breast cancer remains a significant global health challenge, necessitating ongoing research to understand its complexities and improve patient outcomes. Survival analysis techniques offer valuable insights into factors influencing both time to death and cancer recurrence, providing clinicians and researchers with essential prognostic information. This study contributes to this body of knowledge by analyzing data from a German breast cancer study conducted in 1989.

In the context of breast cancer research, numerous studies have explored the relationship between various patient characteristics, tumor features, and treatment modalities with survival outcomes. The literature highlights the importance of factors such as age, tumor grade, lymph node involvement, and hormone receptor status in predicting prognosis and guiding treatment decisions. However, the specific impact of these factors within the context of the German breast cancer study remains to be fully elucidated.

The purpose of this work is to investigate the covariates predictive of time to death and breast cancer recurrence within the studied population. We hypothesize that patient demographics, tumor characteristics, and treatment variables will significantly influence survival outcomes. By employing survival analysis techniques, including Kaplan-Meier estimation and Cox proportional hazards regression, we aim to identify key prognostic factors and assess their impact on patient survival.

The rationale behind this study lies in the need to improve our understanding of breast cancer prognosis and tailor treatment strategies accordingly. By elucidating the factors associated with prolonged survival and reduced recurrence risk, clinicians can better personalize patient care and optimize treatment outcomes. Furthermore, by validating the assumptions underlying survival analysis models, we ensure the robustness and reliability of our findings.

Possible outcomes of this study include the identification of specific patient demographics, tumor characteristics, and treatment modalities associated with favorable or adverse survival outcomes. Additionally, by validating the assumptions of survival analysis techniques, we can enhance the credibility of our findings and provide valuable insights for future research endeavors. Ultimately, this work aims to contribute to the growing body of knowledge on breast cancer prognosis and inform evidence-based clinical practice.

Materials and Methods

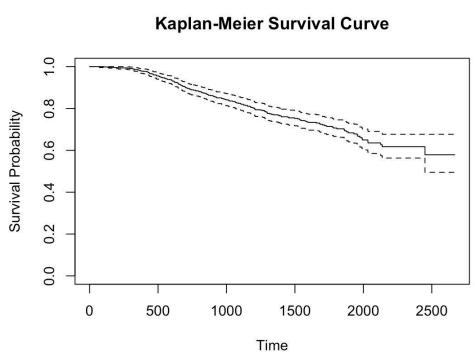
1. Study Design:
    - The study utilized data from a German breast cancer study conducted in 1989.
    - A retrospective observational design was employed to analyze survival outcomes and recurrence risk among breast cancer patients.
2. Data Collection Protocol:
    - The dataset comprised 686 observations with 16 variables, including patient demographics, tumor characteristics, and treatment information.
    - Variables collected included age at diagnosis, menopausal status, hormone therapy, tumor size, tumor grade, number of positive lymph nodes, progesterone receptor count, estrogen receptor count, time to recurrence, and time to death.

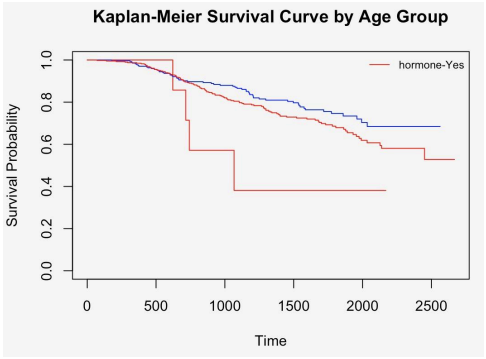
- Data on each patient's diagnosis date, recurrence date (if applicable), and death date (if applicable) were recorded.
- Covariates were categorized and coded according to predefined criteria.

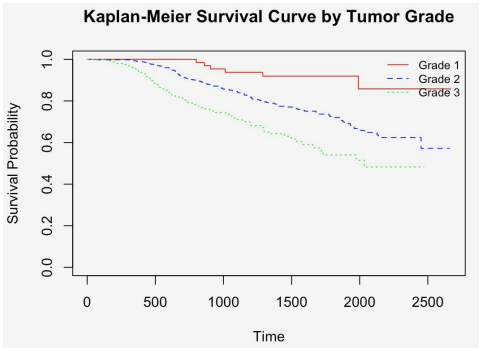
3. Data Analysis:
   a. Survival Analysis Techniques:
      - Kaplan-Meier estimation: Used to visualize survival probabilities over time and assess factors influencing survival.
      - Cox Proportional Hazards Regression: Employed to identify covariates predictive of time to death and breast cancer recurrence. Assumptions of the Cox PH model were validated.
   b. Parametric Regression Models:
      - Exponential, Weibull, and Log-normal distributions were applied to assess the impact of categorical variables (e.g., tumor grade, hormone therapy) on survival outcomes.
      - Akaike Information Criterion (AIC) was used to compare the performance of different parametric models.
4. Statistical Analysis:
   - Descriptive statistics were computed for continuous variables (e.g., age, tumor size) and categorical variables (e.g., menopausal status, hormone therapy).
   - Correlation analysis was performed to assess relationships between variables.
   - Cox proportional hazards models were fitted to the data using the 'coxph' function in R, adjusting for relevant covariates.
   - Assumptions of the Cox PH model were assessed through Kaplan-Meier plots and Schoenfeld Residuals analysis.
   - Parametric regression models were implemented using appropriate functions in R, and model fit was evaluated using AIC.
5. Ethical Considerations:
   - Ethical approval for the original breast cancer study and subsequent data analysis was obtained from relevant institutional review boards.
   - Patient confidentiality and data privacy were maintained throughout the study.
6. Software:
   - Data analysis was performed using statistical software R, utilizing packages such as 'survival', 'ggplot2', and 'survminer' for survival analysis and visualization.

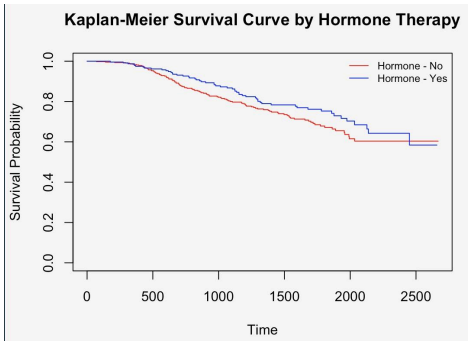
Results

1. Survival Analysis:



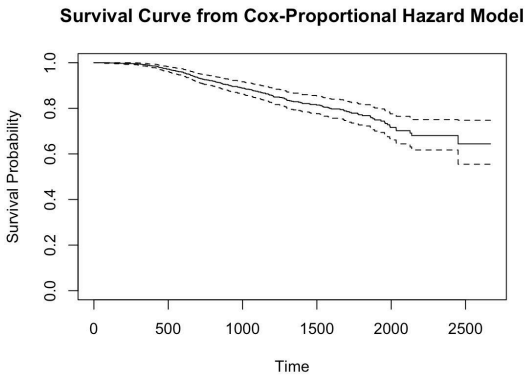


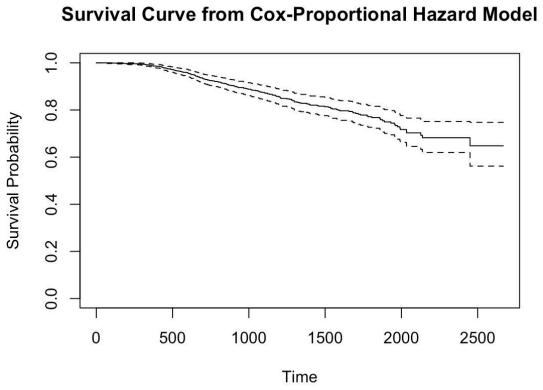


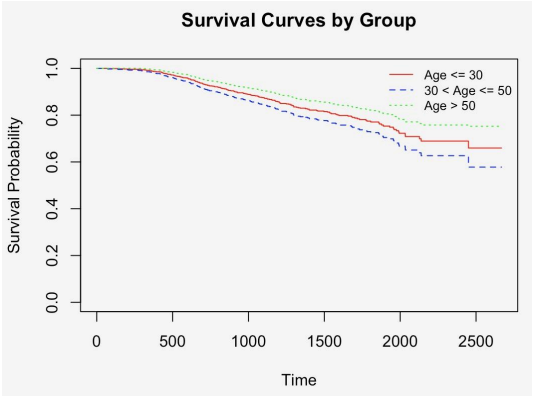


- Kaplan-Meier analysis revealed that patients who underwent hormonal therapy had a higher probability of surviving for at least 1000 days compared to those who did not receive hormonal therapy.
- The Kaplan-Meier curve showed a declining survival rate over time, with approximately 58% of patients surviving after 2600 days.
- Factors such as age, tumor size, tumor grade, number of positive lymph nodes, and progesterone receptor count were found to significantly influence survival outcomes.

2. Cox Proportional Hazards Analysis:

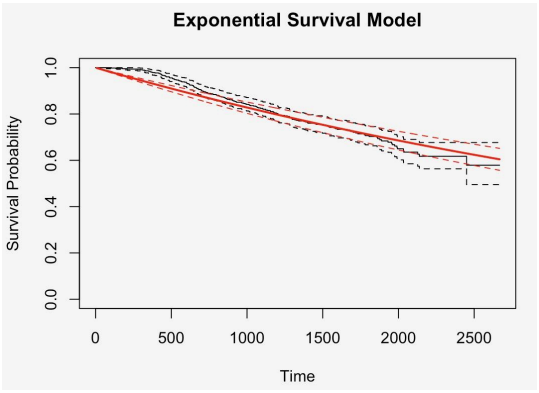


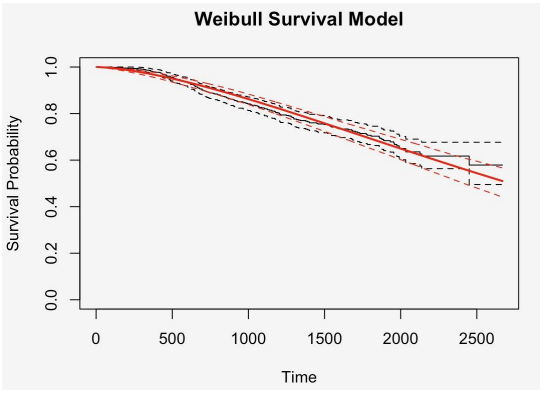


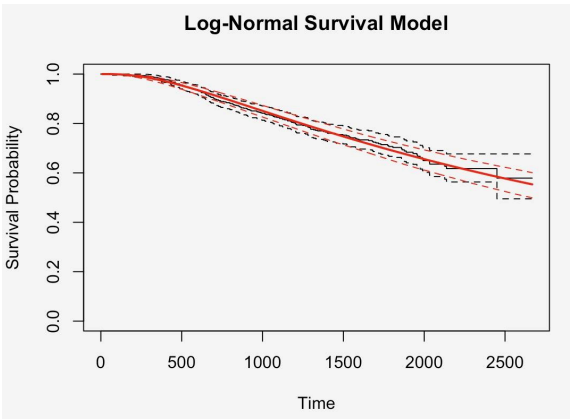


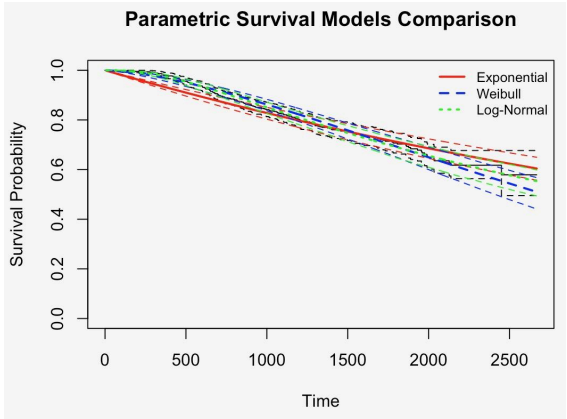
- Cox proportional hazards regression identified several covariates as significant predictors of time to death and breast cancer recurrence.
- Age, tumor grade, number of positive lymph nodes, and progesterone receptor count were associated with increased hazard (likelihood of death), while hormone therapy was associated with a decreased hazard after adjusting for other covariates.
- The final Cox PH model highlighted the importance of age, tumor size, tumor grade, progesterone receptor count, and number of positive lymph nodes in explaining survival time.

3. Parametric Regression Models:



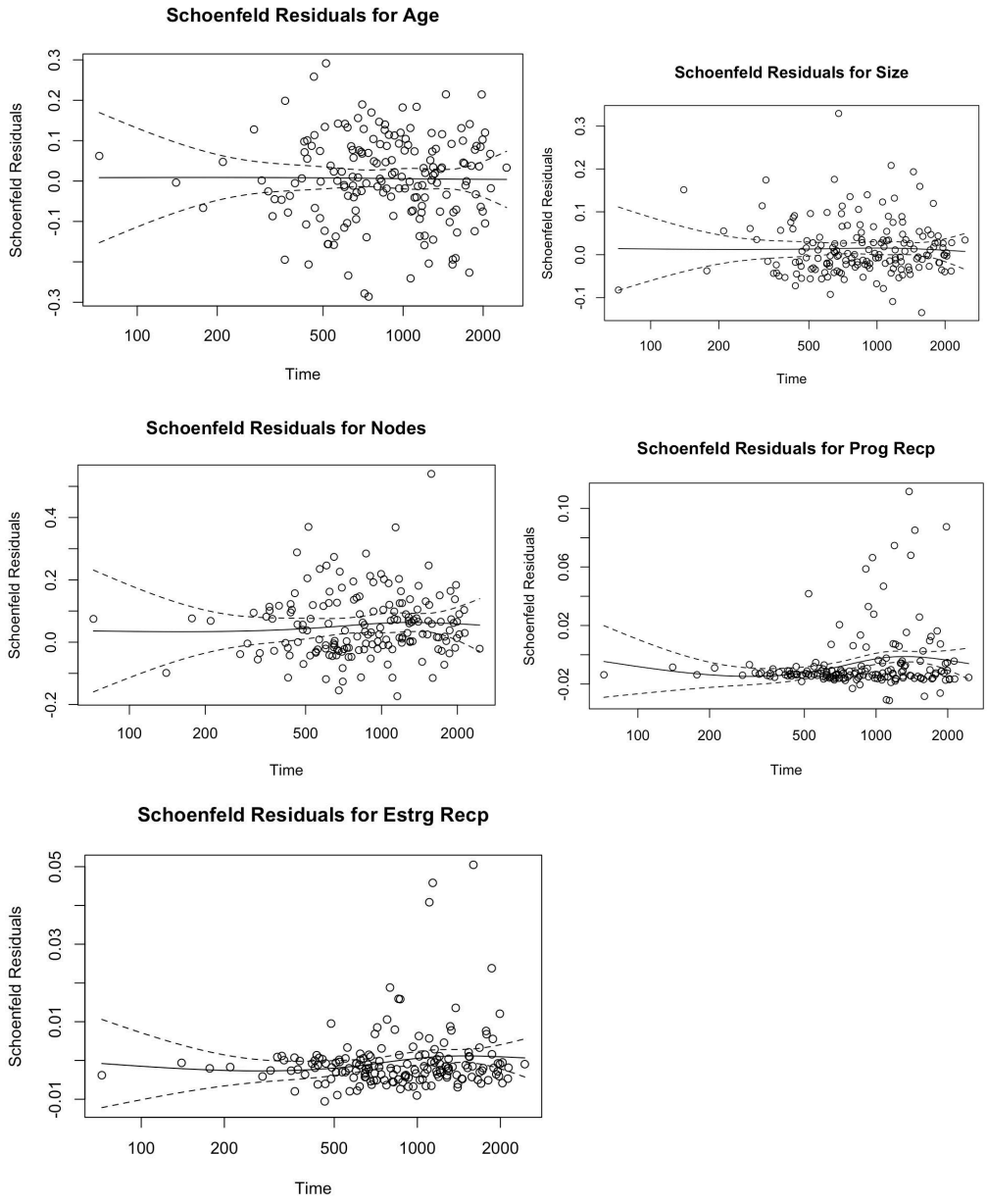






- Parametric regression models, including exponential, Weibull, and log-normal distributions, were employed to assess the impact of categorical variables on survival outcomes.
- The log-normal model provided the best fit to the data based on Akaike Information Criterion (AIC), closely followed by the exponential and Weibull models.

4. Validation of Assumptions:



- Kaplan-Meier plots demonstrated the proportional hazard assumption for tumor grade, with consistent survival probabilities across different tumor grades.
- Schoenfeld Residuals analysis revealed no violation of the proportional hazards assumption, supporting the validity of the Cox PH model.

5. Overall Findings:
    - Age, tumor grade, tumor size, progesterone receptor count, and number of positive lymph nodes emerged as significant predictors of breast cancer survival.
    - Hormonal therapy was associated with a reduced hazard of death, highlighting its potential benefits in improving patient outcomes.
    - The Cox proportional hazards regression model provided valuable insights into the factors influencing survival time and recurrence risk in breast cancer patients.

Acknowledgements

I would like to extend my sincere gratitude to Professor Jack for their exceptional classroom teaching and mentorship throughout my academic journey. Their dedication to fostering a supportive learning environment and their expertise in the field of Survival Analysis have been instrumental in shaping my understanding and passion for research.

I am also thankful for the wealth of knowledge and resources available on the internet, which have served as invaluable supplements to classroom learning. Websites, online forums, and academic platforms have provided me with access to a vast array of study materials, research articles, and tutorials, enriching my educational experience and broadening my perspective on various topics.

Additionally, I would like to acknowledge the contributions of researchers, educators, and professionals who have generously shared their expertise and insights online. Their willingness to disseminate information and engage in scholarly discourse has greatly facilitated my learning process and inspired me to pursue excellence in my own academic endeavors.

Literatures

https://paperswithcode.com/paper/sensitivity-of-survival-analysis-metrics

https://paperswithcode.com/paper/survival-analysis-algorithms-based-on

https://paperswithcode.com/paper/adaptive-sampling-for-weighted-log-rank

https://paperswithcode.com/paper/towards-flexible-time-to-event-modeling

https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/

https://www.reddit.com/r/rstats/comments/njf52n/timevarying_covariates_in_survival_analysis/

https://www.statmethods.net/advstats/glm.html

https://www.openintro.org/go/?id=survival_analysis_in_R&referrer=/book/surv_in_r/index.php