Cancer Survival Analysis Using Kaplan-Meier Curves, Log-Rank Tests, and Cox Proportional Hazards Modeling

By Vedh Bagare

Abstract:
Survival analysis is a statistical framework that is used to predict the time until an event occurs, such as death or disease recurrence. This project looks at survival in breast cancer patients using the GBSG2 (German Breast Cancer Study Group 2) dataset. Kaplan-Meier methods were used to estimate survival probabilities, log-rank tests compared survival between age groups, and a Cox proportional hazards model quantified the effect of age on survival. In the end, although older age groups showed a trend toward being riskier, the effects were not statistically significant after adjustment.

Data and Preprocessing:
The GBSG2 dataset contains 686 breast cancer patients with times for events in days. \
Below are the main steps done in preprocessing:

- Added a unique patient ID to each patient record
- Converted follow-up time to months using 30.44 days per month
- Converted the event indicator to a numeric system (0 = censored, 1 = event)
- Converted categorical variables to category datatype in pandas
- Inputted missing numeric values with the column mean and categorical variables with the mode

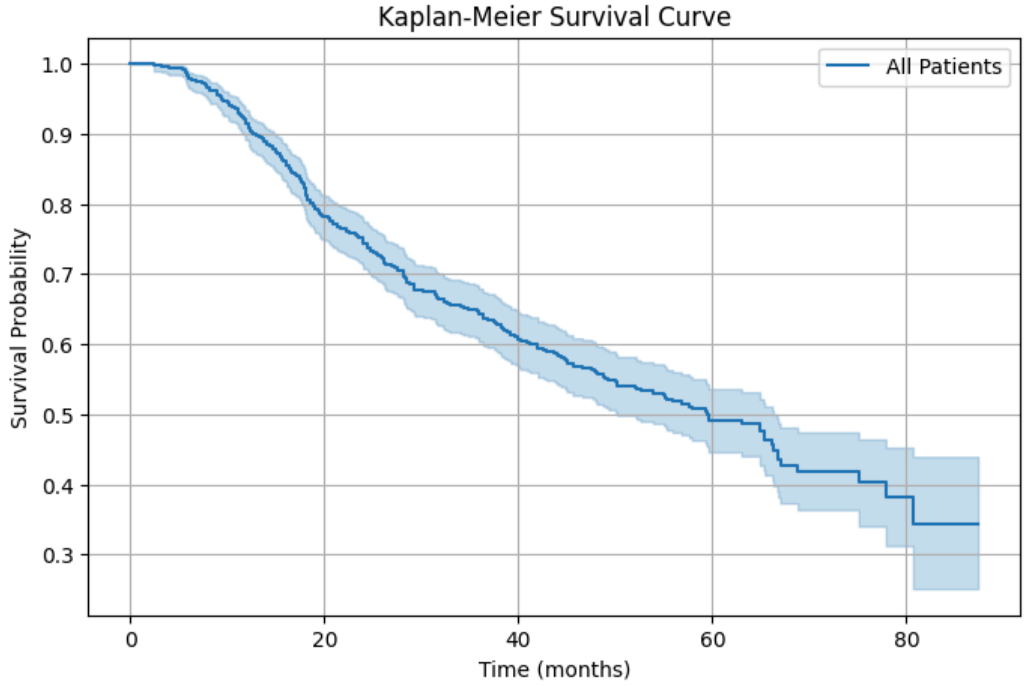
Methods:

- Kaplan-Meier Survival Estimation:
  The KM estimator basically calculates the probability of surviving past a given time, while accounting for censored observations. In this project, age groups were created by dividing patient age into bins (e.g., <50, 50-59, 60-69, 70-79, 80+). The estimator was applied to the whole group, but separate KM curves were generated for each age group listed above. 95% confidence intervals were computed for survival probabilities.
- Log-Rank Test:
  This test compares survival distributions between two or more groups. It tests the null hypothesis that the survival curves are identical across groups.
- Cox Proportional Hazards Model
  The Cox PH model estimates the effect of covariates on the hazard, which is the instantaneous risk of the event occurring. The model produces hazard ratios (relative risk compared to the reference group), 95 percent confidence intervals,

p-values for statistical significance, and a concordance index to assess the model's ability to rank patients by risk. Assumptions of proportional hazards were checked using the scaled Schoenfeld residuals.

Overall Survival Results:

The Kaplan-Meier curve for all patients is shown below.



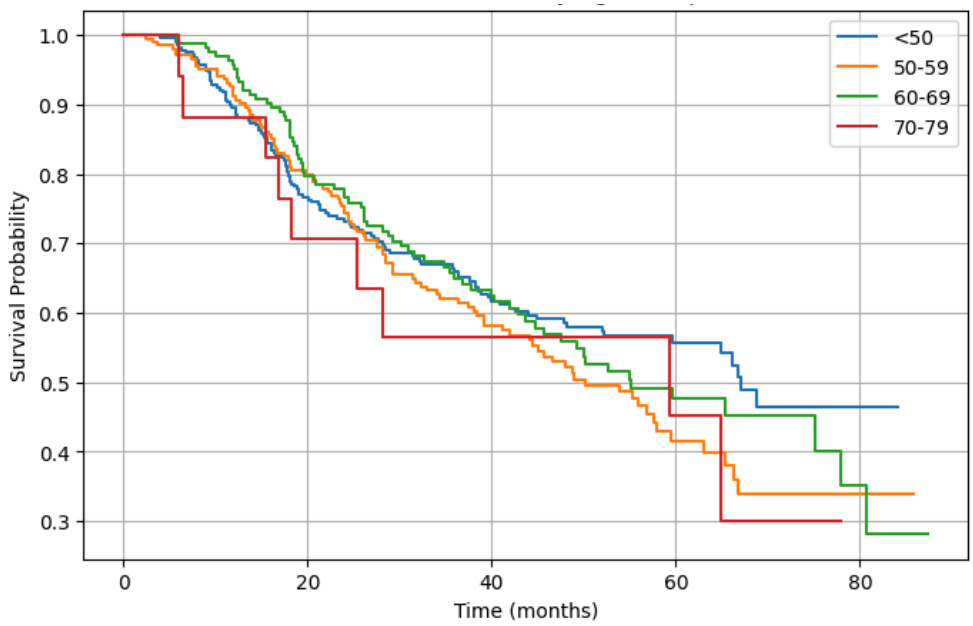
Median Survival Time: 59.4 months.

Survival probabilities at selected time points:

Time (months)	Survival Probabilities	95% CI Lower	95% CI Upper
12	0.915560	0.889994	0.933027
24	0.743005	0.707511	0.774895
36	0.642709	0.603400	0.679203
60	0.491850	0.446017	0.536003

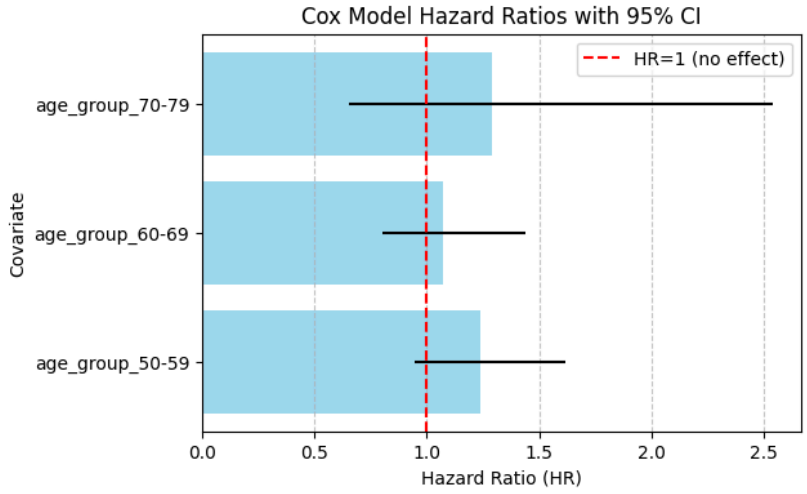
Interpretation: After the first year, around 92% of the patients survived. After 2 years, around 74% survived, with roughly half surviving for 5 years. The confidence intervals show the range of true survival. For example, at 12 months, the CI values are between 0.89-0.93, which means that we are 95% confident that the true survival probability is between 89% to 93%.

Survival by Age Group:



To clearly compare survival across age groups, a log-rank test was used. This test looked at the entire follow-up period and showed that survival differed across age groups, with older patients experiencing events earlier and more frequently than younger patients.

Cox Proportional Hazards Model Results:



Age 50-59	0.213017	1.237406	0.946480	1.617756	0.119299
Age 60-69	0.071801	1.074441	0.802623	1.438314	0.629461
Age 70-79	0.253137	1.288059	0.653294	2.539588	0.464880

The Cox Proportional Hazards model was used to quantify how the age group affects the risk of an event over time while using one group as the reference. Hazard ratios over 1 indicate a higher risk, with values below 1 meaning a lower risk. In this project, the reference group was under the age of 50. Patients aged 50-59 had a hazard ratio of 1.24, which means their estimated risk was about 24% higher, but this result is not statistically reliable because the confidence interval includes 1. Patients aged 60-69 had a ratio of 1.07, and the group aged 70-79 had a ratio of 1.29. In the end, there was no clear evidence here that these age groups had a meaningfully different risk than the reference group.

Model Diagnostics:
The Cox model says that the relative risk between age groups stays roughly constant over time. This was evaluated using tests based on Schoenfeld residuals, and no strong violations were detected. The concordance index showed the model had limited ability to perfectly rank patients by risk, which is expected because age group alone is not the best predictor of survival. Overall, the diagnostics suggest that the model is statistically acceptable, but not very predictive.

Conclusion:
This analysis demonstrates how survival analysis techniques can be used to study time-to-event data while accounting for censoring. While age group alone was not a strong predictor of survival in this dataset, future analysis using additional clinical factors would improve predictive performance.