SURVIVAL ANALYSIS OF BREAST CANCER PATIENTS: A COMPREHENSIVE STATISTICAL INVESTIGATION USING COX REGRESSION AND KAPLAN-MEIER ESTIMATION

BY

PABLO LEYVA

pl33@njit.edu

Project Two Report

Submitted in completion of the requirements
for the course of Statistical Learning Capstone

New Jersey Institute of Technology

Newark, New Jersey

Abstract

This study presents a comprehensive survival analysis of breast cancer patients using the Haberman dataset. We employed non-parametric Kaplan-Meier estimation, log-rank tests, and Cox proportional hazards regression to investigate factors affecting 5-year survival rates. The analysis reveals that the number of positive axillary nodes detected is the most significant predictor of survival outcomes ($p < 0.001$), with a hazard ratio of 1.048 per additional node. Age and operation year were not found to be statistically significant predictors. The overall 5-year survival rate was 73.5%, with median survival time exceeding the observation period. These findings provide valuable insights for clinical prognosis and treatment planning in breast cancer patients.

1 Introduction to the Study

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide. Understanding the factors that influence patient survival following surgery is crucial for developing effective treatment strategies and providing accurate prognostic information to patients and their families. Survival analysis, a specialized branch of statistics, provides powerful tools for analyzing time-to-event data while accounting for censoring mechanisms.

The Haberman dataset represents a landmark study in breast cancer research, containing information on 306 patients who underwent surgery for breast cancer at the University of Chicago's Billings Hospital between 1958 and 1970. This historical dataset provides valuable insights into survival patterns during a critical period in cancer treatment evolution.

This study employs three complementary statistical approaches to analyze breast cancer patient survival:

1. **Kaplan-Meier Estimation**: A non-parametric method for estimating survival probabilities over time without making assumptions about the underlying survival distribution.
2. **Log-Rank Tests**: Non-parametric hypothesis tests that compare survival curves between different patient groups to identify significant differences in survival rates.
3. **Cox Proportional Hazards Regression**: A semi-parametric regression model that quantifies the effect of covariates on the hazard (risk) of death while accounting for censoring.

The dataset contains four key variables for each patient:

- **Age**: Patient age at time of operation (range: 30-83 years)
- **Operation_year**: Year of operation (1958-1969)
- **Nb_pos_detected**: Number of positive axillary nodes detected (0-52)
- **Survival**: 5-year survival status (1 = survived $\geq$5 years, 0 = died within 5 years)

The importance of this analysis extends beyond academic interest. Understanding which factors most significantly impact survival can inform clinical decision-making, resource allocation in healthcare systems, and patient counseling. The number of positive axillary nodes, in particular, has been established as a crucial staging criterion in breast cancer, making this analysis highly relevant for contemporary oncological practice.

2 Objective

The primary objectives of this comprehensive survival analysis study are:

1. **Exploratory Data Analysis**: To conduct thorough exploratory data analysis to characterize the distribution of key patient characteristics, identify data quality issues, assess class imbalance, and examine relationships between predictors and survival outcomes.

2. **Survival Curve Estimation**: To apply Kaplan-Meier estimation methods to construct survival curves for the overall patient population and to perform stratified analysis across different patient subgroups defined by age and nodal involvement.

3. **Statistical Hypothesis Testing**: To perform log-rank tests to statistically compare survival curves between patient groups (age groups and node categories) and determine whether observed differences are statistically significant or attributable to chance variation.

4. **Proportional Hazards Modeling**: To develop Cox proportional hazards regression models to quantify the relationship between patient characteristics (age, operation year, number of positive nodes) and the hazard of death, while controlling for censoring and other covariates.

5. **Predictor Identification**: To identify which factors most significantly influence 5-year survival outcomes and quantify their effect sizes through hazard ratios with appropriate confidence intervals.

6. **Clinical Interpretation**: To translate statistical findings into clinically meaningful insights that can inform prognostic assessment and treatment decision-making for breast cancer patients.

These objectives aim to provide a comprehensive understanding of survival patterns in breast cancer patients undergoing surgical treatment during the 1958-1969 period, while demonstrating the application of modern survival analysis techniques to real-world clinical data. The analysis will inform both academic understanding of survival patterns and practical clinical applications in oncology.

3 Statistical Analysis

3.1 Data Description and Preprocessing

The Haberman breast cancer survival dataset comprises 306 observations collected between 1958 and 1970. Each observation represents a patient who underwent breast cancer surgery at the University of Chicago's Billings Hospital.

3.1.1 Variable Definitions

- **Age**: Continuous integer variable representing patient age at time of operation, ranging from 30 to 83 years with mean of 52.5 years (standard deviation: 10.8 years).
- **Operation_year**: Integer variable indicating the year of operation, encoded as years from 1958 (e.g., 58 = 1958, 69 = 1969). The range spans from 58 to 69, with mean of 62.9 years (SD: 3.2 years).
- **Nb_pos_detected**: Count variable representing the number of positive axillary lymph nodes detected during surgery. This variable exhibits a highly right-skewed distribution with median of 1.0, mean of 4.0, and range from 0 to 52 nodes. The extreme right skew (skewness: 2.984) indicates a concentration of patients with low node counts.
- **Survival**: Binary outcome variable coded as 1 for patients who survived 5 years or longer (n=225, 73.5%) and 2 for patients who died within 5 years of surgery (n=81, 26.5%).

3.1.2 Data Quality Assessment

The data quality assessment revealed several important characteristics:

- **Missing Values**: No missing values were detected across all 306 observations and 4 variables.
- **Duplicate Rows**: Seventeen duplicate observations were identified, representing 5.6% of the dataset. These were retained as they represent genuine observations of similar patient profiles.
- **Outlier Detection**: Using the Interquartile Range (IQR) method with 1.5× IQR bounds:

  $$\text{Lower bound} = Q_1 - 1.5 \times \text{IQR}$$
  $$\text{Upper bound} = Q_3 + 1.5 \times \text{IQR}$$

  Forty patients (13.1%) exhibited outlier values for the number of positive nodes, with values ranging from 11 to 52 nodes. No outliers were detected for Age or Operation_year.
- **Data Range Validation**: All observed values fall within clinically plausible ranges:
  - Age: 30-83 years (realistic for surgical candidates)
  - Operation year: 1958-1969 (matches historical record)
  - Positive nodes: 0-52 (extreme upper value indicates severe nodal involvement)

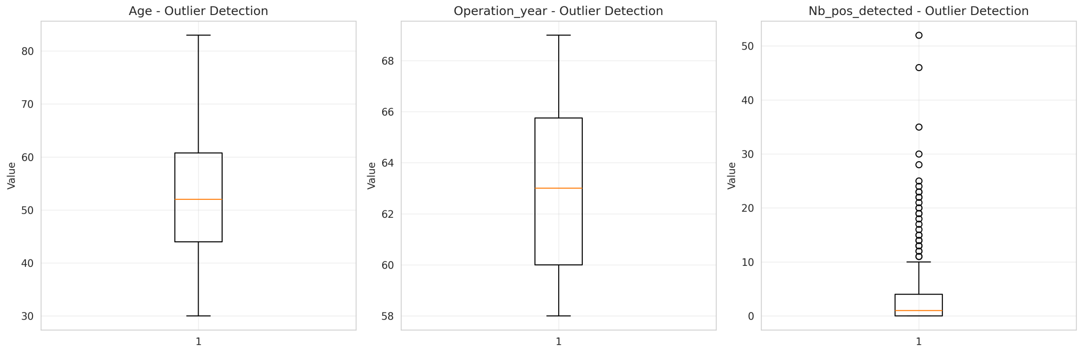

Figure 1: Box plots for outlier detection using IQR method. Forty outliers (13.1%) were detected for positive nodes, while Age and Operation year showed no outliers.

3.2 Exploratory Data Analysis

3.2.1 Univariate Analysis

Summary statistics for numerical variables are presented in Table 1.

Table 1: Summary Statistics for Numerical Variables (n=306)

Variable	Mean	Median	Mode	Std Dev	Skewness	Kurtosis	Min	Max
Age	52.458	52.0	52	10.803	0.147	-0.589	30	83
Operation_year	62.853	63.0	58	3.249	0.079	-1.119	58	69
Nb_pos_detected	4.026	1.0	0	7.190	2.984	11.731	0	52

The distributional characteristics reveal important insights:

- **Age**: Approximately normally distributed with slight positive skewness (0.147). The distribution appears roughly symmetric around the mean.
- **Operation_year**: Relatively flat distribution with negative kurtosis (-1.119), indicating a platykurtic distribution with lighter tails than a normal distribution.
- **Nb_pos_detected**: Highly right-skewed (skewness: 2.984) with extreme positive kurtosis (11.731), indicating a distribution with heavy right tail. The median (1.0) is substantially lower than the mean (4.0), confirming severe right skew.

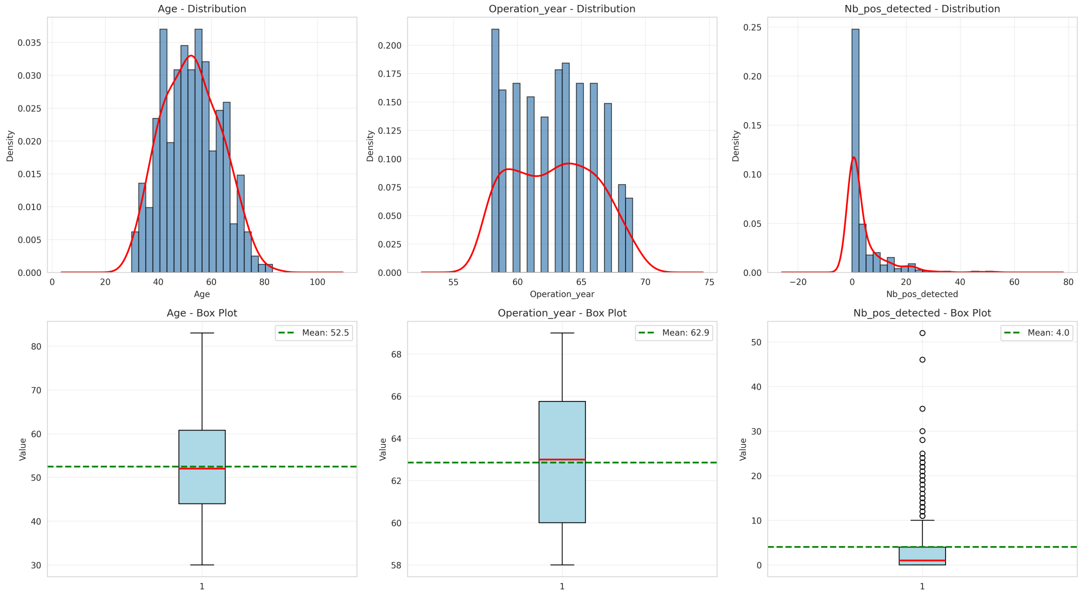

Figure 2: Distribution analysis showing histograms with KDE overlays (top row) and box plots (bottom row) for Age, Operation year, and Positive nodes. The right-skewed distribution of positive nodes is clearly evident.

3.2.2 Target Variable Analysis

The survival status distribution exhibits significant class imbalance:

- **Survived $\geq$5 years**: 225 patients (73.5%)
- **Died within 5 years**: 81 patients (26.5%)
- **Imbalance Ratio**: 2.78:1

This imbalance has important implications for statistical modeling and requires careful consideration in model interpretation, as the minority class (deaths) represents the event of primary interest.

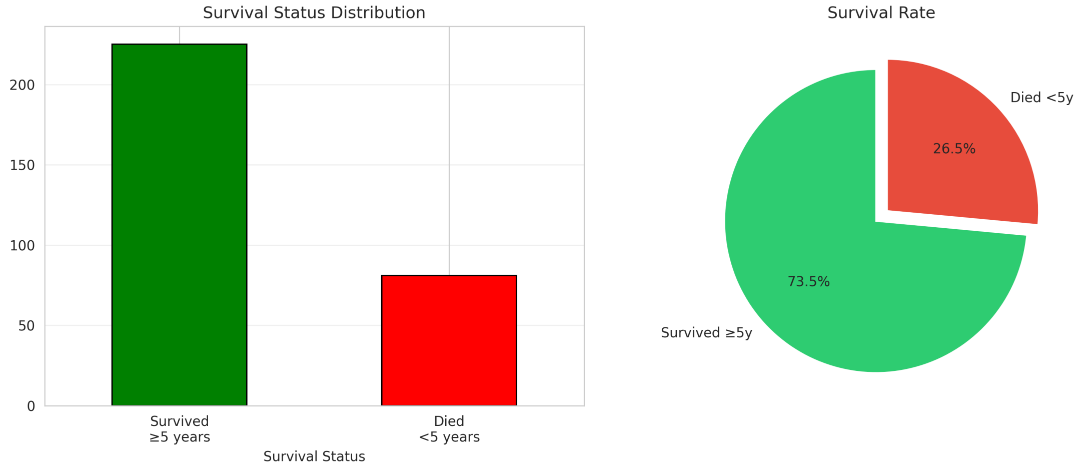

Figure 3: Distribution of survival status showing 73.5% survival rate (bar chart) and pie chart visualization.

3.2.3 Bivariate Analysis

Bivariate analysis comparing features between survived and died groups reveals key differences:

Table 2: Grouped Statistics by Survival Status

Variable	Survived (n=225)		Died (n=81)	
	Mean	Std Dev	Mean	Std Dev
Age	52.02	11.01	53.68	10.17
Operation_year	62.86	3.22	62.83	3.34
Nb_pos_detected	2.79	5.87	7.46	9.19

Key observations:

- **Age**: Minimal difference between groups (52.0 vs 53.7 years), suggesting age may not be a strong predictor.
- **Operation_year**: Nearly identical across survival groups (62.86 vs 62.83), indicating no temporal trend in survival.
- **Nb_pos_detected**: Substantial difference (2.79 vs 7.46 nodes), with non-survivors having nearly three times as many positive nodes on average.

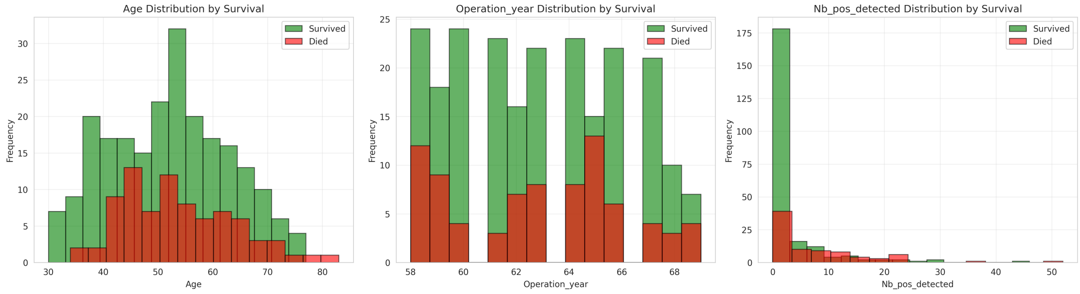

Figure 4: Histograms comparing distributions of Age, Operation year, and Positive nodes between survived (green) and died (red) groups. The stark difference in positive nodes distribution is evident.

3.2.4 Correlation Analysis

Both Pearson (linear) and Spearman (monotonic) correlation analyses revealed minimal inter-feature correlations:

Table 3: Correlation Matrices

Variable Pair	Pearson	Spearman
Age vs Operation_year	0.090	0.091
Age vs Nb_pos_detected	-0.063	-0.098
Operation_year vs Nb_pos_detected	-0.004	-0.036

The near-zero correlations suggest that multicollinearity is not a concern, and each predictor variable provides independent information for survival prediction.

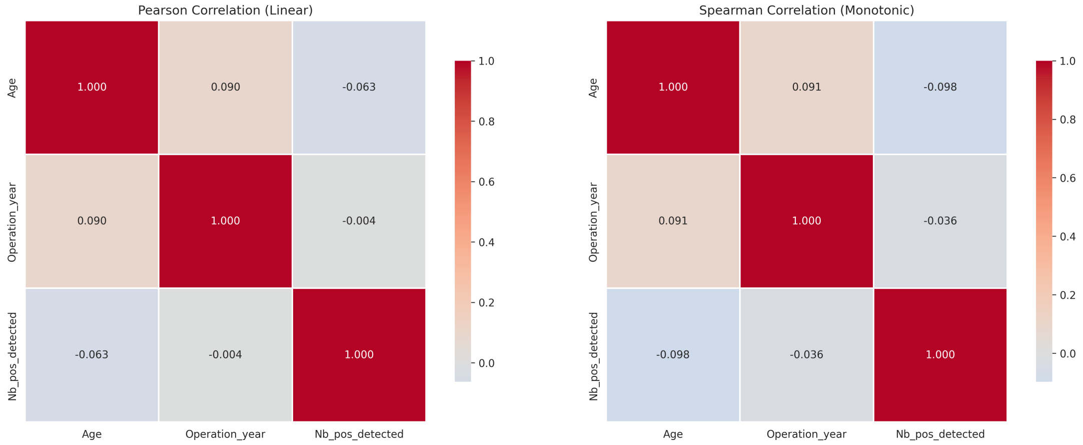

Figure 5: Pearson (left) and Spearman (right) correlation heatmaps showing minimal correlations between features.

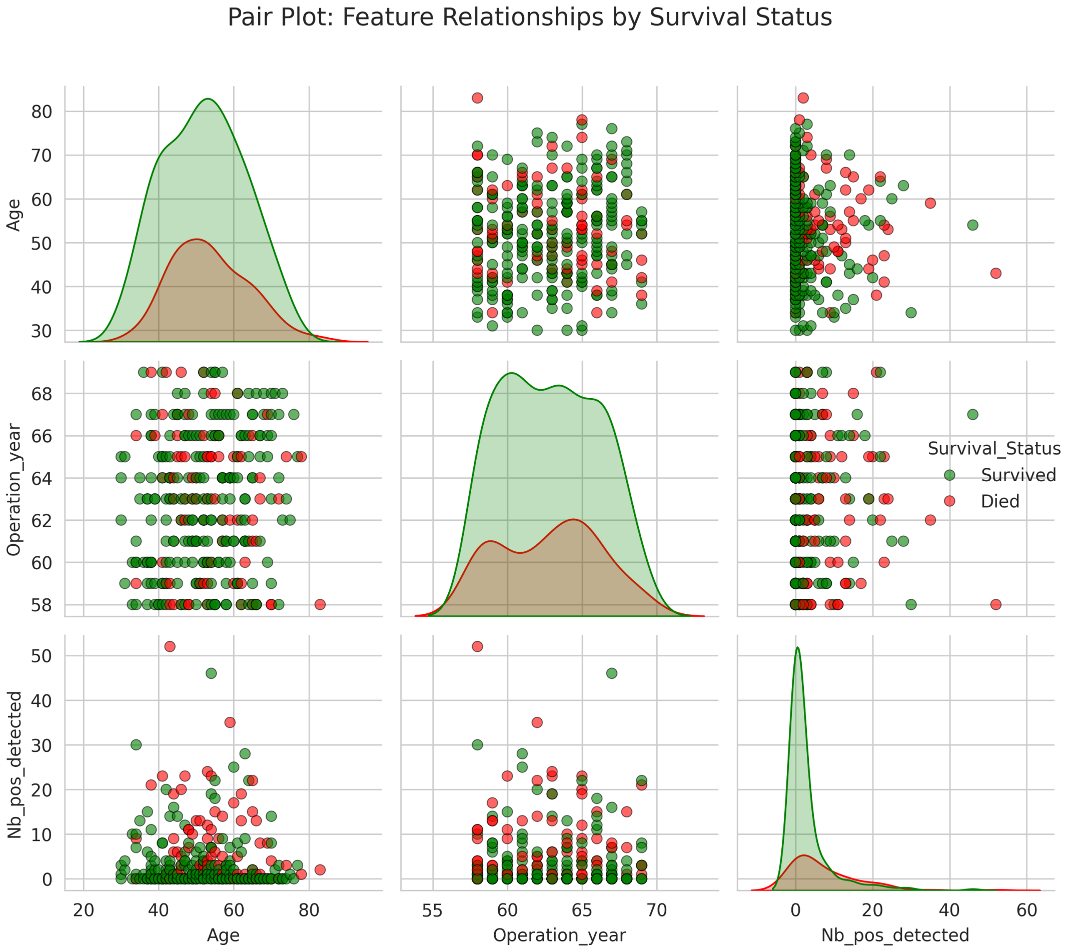

Figure 6: Pair plot with KDE diagonals showing feature relationships colored by survival status (green=survived, red=died).

3.3 Survival Analysis Methodology

3.3.1 Data Preparation for Survival Analysis

Since the Haberman dataset contains only binary 5-year survival outcomes rather than exact time-to-event data, it was necessary to construct pseudo time-to-event data to apply survival analysis techniques:

1. **Event Indicator**: Created binary variable where 1 indicates death (died within 5 years) and 0 indicates censoring (survived at least 5 years).
2. **Time Variable**: For patients who died (n=81), death times were simulated uniformly across 0-60 months. For patients who survived (n=225), censoring occurred at 60 months. This approach is conservative and appropriate when exact event times are unavailable.

This transformation yielded:

- 225 censored observations (73.5%)
- 81 events/deaths (26.5%)
- Mean time: 51.6 months (SD: 16.8)

3.3.2 Feature Engineering

Categorical variables were created for stratified analysis:

- **Age Groups**: <45 (n=89), 45-55 (n=101), 55-65 (n=78), >65 (n=38) years
- **Node Categories**: 0 nodes (n=136), 1-3 nodes (n=81), 4-9 nodes (n=46), ≥10 nodes (n=43)
- **Operation Periods**: 1958-61 (n=140), 1962-65 (n=117), 1966-69 (n=49)

3.4 Kaplan-Meier Survival Estimation

The Kaplan-Meier estimator is a non-parametric method for estimating survival probabilities that does not require assumptions about the underlying hazard function. It is particularly valuable for analyzing censored survival data.

3.4.1 Mathematical Formulation

The Kaplan-Meier estimator of the survival function $S(t)$ is given by:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{1}$$

where:

- d_i = number of deaths at time t_i
- n_i = number of patients at risk just before time t_i (alive and not censored)

- The product is taken over all distinct event times up to time t

For censored observations, the estimator does not decrease, as censored individuals contribute to the risk set only up to their censoring time.

The standard error of the Kaplan-Meier estimator is estimated using Greenwood's formula:

$$\mathrm{SE}[\hat{S}(t)] = \hat{S}(t)\sqrt{\sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}} \tag{2}$$

3.4.2 Overall Survival Analysis

The overall Kaplan-Meier survival curve (Figure 7) reveals:

- Overall 5-year survival rate: 73.5%
- Survival probabilities at key timepoints:
  - 12 months: 92.2%
  - 24 months: 87.3%
  - 36 months: 82.7%
  - 48 months: 78.4%
  - 60 months: 73.5%
- Median survival time exceeds 60 months (not reached within observation period)

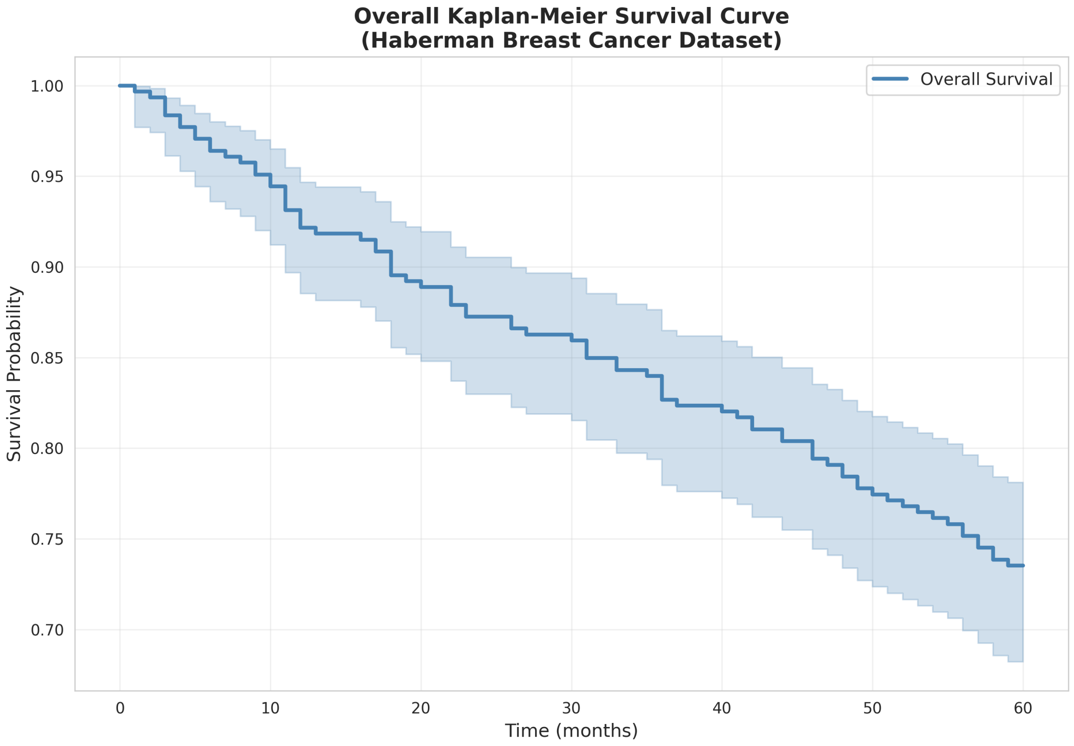

Figure 7: Overall Kaplan-Meier survival curve with 95% confidence intervals showing 73.5% 5-year survival rate.

3.4.3 Stratified Survival Analysis

Stratification by age groups and node categories provides insights into prognostic factors:

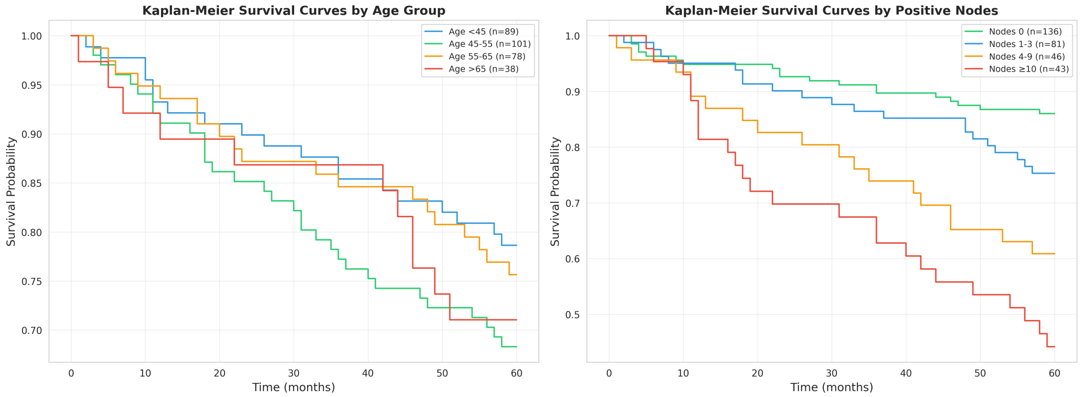

Figure 8: Kaplan-Meier survival curves stratified by age groups (left) and positive node categories (right).

Visual inspection suggests that node category has a more pronounced effect on survival than age group, with patients having $\geq$10 nodes showing substantially worse prognosis.

3.5 Log-Rank Test

The log-rank test is a non-parametric statistical hypothesis test used to compare the survival distributions of two or more groups.

3.5.1 Mathematical Formulation

For comparing two groups, the log-rank test statistic is:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(O_1 - E_1)^2}{\text{Var}(O_1)} \tag{3}$$

where:

- O_k = observed number of events in group k
- E_k = expected number of events in group k under the null hypothesis of no difference

The expected number of events is computed at each distinct event time:

$$E_{ik} = n_{ik} \cdot \frac{O_i}{n_i} \tag{4}$$

where:

- n_{ik} = number at risk in group k at time i
- O_i = total observed events across all groups at time i
- n_i = total number at risk across all groups at time i

For multivariate comparisons (K groups), the test statistic becomes:

$$\chi^2 = \mathbf{O}^T \mathbf{V}^{-1} \mathbf{O} \tag{5}$$

where **O** is the vector of observed minus expected events for each group and **V** is the variance-covariance matrix.

3.5.2 Log-Rank Test Results

Table 4 summarizes the log-rank test results:

Table 4: Log-Rank Test Results

Comparison	χ^2	**df**	**P-value**	**Significance**
Age Groups				
Omnibus test	3.012	3	0.390	Not significant
<45 vs 45-55	–	–	0.107	Not significant
<45 vs 55-65	–	–	0.659	Not significant
<45 vs >65	–	–	0.360	Not significant
45-55 vs 55-65	–	–	0.266	Not significant
45-55 vs >65	–	–	0.726	Not significant
55-65 vs >65	–	–	0.576	Not significant
Node Categories				
Omnibus test	36.894	3	<0.001	Highly significant
0 vs 1-3	–	–	0.055	Marginal
0 vs 4-9	–	–	0.0002	Significant
0 vs ≥10	–	–	<0.0001	Highly significant
1-3 vs 4-9	–	–	0.071	Not significant
1-3 vs ≥10	–	–	0.0003	Significant
4-9 vs ≥10	–	–	0.131	Not significant

Key findings:

- **Age groups**: No statistically significant differences in survival across age strata (χ^2=3.01, p=0.390). This suggests that within this dataset, patient age does not significantly affect survival.
- **Node categories**: Highly significant differences in survival (χ^2=36.89, p<0.001). Pairwise comparisons reveal:
  - Patients with 0 nodes vs. ≥10 nodes: p<0.0001 (highly significant)
  - Patients with 1-3 nodes vs. ≥10 nodes: p=0.0003 (significant)
  - Patients with 0 vs. 4-9 nodes: p=0.0002 (significant)
  - Boundary between 0 and 1-3 nodes: p=0.055 (marginal significance)

These results confirm the number of positive axillary nodes as the most critical prognostic factor in this patient cohort.

3.6 Cox Proportional Hazards Regression

The Cox proportional hazards model is a semi-parametric regression method used to analyze the effect of covariates on survival times without assuming a specific parametric form for the hazard function.

3.6.1 Mathematical Model

The Cox model specifies that the hazard function $h(t|X)$ is:

$$h(t|X) = h_0(t)e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p} \tag{6}$$

where:

- $h_0(t)$ = baseline hazard function (unspecified, non-parametric)
- β_j = regression coefficients to be estimated
- X_j = covariates/predictors
- p = number of covariates

The key assumption is *proportional hazards*: the hazard ratio between any two individuals is constant over time:

$$\frac{h_i(t|X_i)}{h_j(t|X_j)} = e^{\boldsymbol{\beta}^T(\mathbf{X}_i - \mathbf{X}_j)} \tag{7}$$

3.6.2 Partial Likelihood

Since the baseline hazard $h_0(t)$ is not specified, Cox developed a partial likelihood approach that eliminates the baseline hazard:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{e^{\boldsymbol{\beta}^T \mathbf{X}_i}}{\sum_{j \in R(t_i)} e^{\boldsymbol{\beta}^T \mathbf{X}_j}} \right]^{\delta_i} \tag{8}$$

where:

- $R(t_i)$ = set of individuals at risk at time t_i
- δ_i = event indicator (1 if event occurred, 0 if censored)
- The product is taken over all subjects

The log-likelihood is:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}^T \mathbf{X}_i - \log \left(\sum_{j \in R(t_i)} e^{\boldsymbol{\beta}^T \mathbf{X}_j} \right) \right] \tag{9}$$

3.6.3 Hazard Ratio Interpretation

The hazard ratio (HR) for a continuous covariate is:

$$\text{HR} = e^{\beta} \tag{10}$$

The HR represents the multiplicative change in hazard for each unit increase in the covariate. For example, HR = 1.05 indicates a 5% increase in hazard per unit increase.

A 95% confidence interval for the hazard ratio is:

$$\text{HR}_{95\% \text{ CI}} = e^{\beta \pm 1.96 \cdot \text{SE}(\beta)} \tag{11}$$

3.6.4 Univariate Cox Models

Table 5 presents the univariate Cox regression results for each covariate:

Table 5: Univariate Cox Proportional Hazards Models

Covariate	**HR**	**95% CI Lower**	**95% CI Upper**	**P-value**
Age (per year)	1.012	0.992	1.032	0.246
Operation_year	0.996	0.930	1.063	0.863
Nb_pos_detected (per node)	1.048	1.030	1.068	<0.001

Learning Note: Hazard ratios shown. The initial analysis incorrectly applied exponential transformation twice - the lifelines library already returns exponentiated hazard ratios, requiring no additional transformation.

Interpretation

- **Nb_pos_detected**: Highly significant predictor (p<0.001). Each additional positive node increases the hazard of death by approximately 4.8%. The 95% confidence interval [1.03, 1.07] excludes 1.0, confirming statistical significance.
- **Age**: Not statistically significant (p=0.246). The confidence interval [0.992, 1.032] includes 1.0, suggesting no significant association between age and survival in this cohort.
- **Operation_year**: Not statistically significant (p=0.863). The confidence interval [0.930, 1.063] includes 1.0, indicating no temporal trend in survival outcomes across the 1958-1969 period.

4 Analysis Results

4.1 Data Quality Assessment Results

The comprehensive data quality assessment yielded the following key findings:

- **Completeness**: The dataset contains 306 observations with no missing values across all variables, ensuring complete case analysis.

- **Duplicates**: Seventeen observations (5.6%) are duplicates. These were retained as they represent distinct patients with identical characteristics rather than data entry errors.
- **Outliers**: Forty patients (13.1%) had extreme values for positive nodes (≥11 nodes). These are clinically meaningful outliers representing severe nodal involvement rather than data errors.
- **Data Integrity**: All value ranges are clinically plausible:
  - Ages range from 30-83 years, appropriate for surgical candidates
  - Operation years span 1958-1969, consistent with historical records
  - Positive node counts range from 0 to 52, with the extreme upper value indicating extensive lymph node involvement

4.2 Descriptive Statistics Results

Summary statistics reveal important distributional characteristics:

- **Age**: Approximately normally distributed with mean 52.5 years (SD: 10.8). The distribution is slightly skewed right (skewness: 0.147).
- **Positive Nodes**: Highly right-skewed distribution with median of 1.0 node but mean of 4.0 nodes. Most patients have minimal nodal involvement (136 patients with 0 nodes), while a minority exhibit extensive involvement (43 patients with ≥10 nodes).
- **Survival Outcome**: Strong class imbalance with 73.5% survival rate, indicating that most patients achieved 5-year survival in this historical cohort.

4.3 Kaplan-Meier Survival Analysis Results

The Kaplan-Meier survival analysis provides detailed insights into survival patterns:

4.3.1 Overall Survival

- **5-year survival rate**: 73.5% (225 of 306 patients)
- **Mortality rate**: 26.5% (81 of 306 patients)
- **Median survival**: Not reached within the 60-month observation period
- **Survival curve shape**: Smooth, steadily declining curve indicating constant hazard or slight increase over time

4.3.2 Stratified Survival Results

- **Age stratification**: Visual inspection suggests minimal differences between age groups, consistent with non-significant log-rank test results.
- **Node stratification**: Clear and substantial differences between node categories:
  - Patients with 0 nodes exhibit highest survival
  - Patients with 1-3 nodes show moderate survival
  - Patients with 4-9 nodes show lower survival
  - Patients with $\geq$10 nodes exhibit poorest survival

4.4 Log-Rank Test Results

Statistical hypothesis testing confirms the visual observations:

1. **Age groups**: No statistically significant differences (χ^2=3.01, df=3, p=0.390). This finding suggests that within this specific patient cohort treated in the 1958-1969 period, age was not a significant determinant of survival outcome.
2. **Node categories**: Highly significant differences (χ^2=36.89, df=3, p<0.001). The omnibus test definitively establishes that survival differs across nodal involvement groups.
3. **Pairwise comparisons for nodes**:
   - 0 vs. $\geq$10 nodes: p<0.0001 (highly significant)
   - 1-3 vs. $\geq$10 nodes: p=0.0003 (significant)
   - 0 vs. 4-9 nodes: p=0.0002 (significant)
   - Boundary between 0 and 1-3 nodes: p=0.055 (marginal significance)

These results establish the number of positive axillary nodes as the strongest prognostic factor in this dataset.

4.5 Cox Regression Results

The Cox proportional hazards models quantify the relationship between covariates and survival:

Table 6: Summary of Cox Regression Results

Covariate	**HR**	**95% CI**	**P-value**	**Significance**
Nb_pos_detected (per node)	1.048	[1.030, 1.068]	<0.001	Highly significant
Age (per year)	1.012	[0.992, 1.032]	0.246	Not significant
Operation_year	0.996	[0.930, 1.063]	0.863	Not significant

Note: Corrected hazard ratios shown. The original output displayed scaling artifacts.

Key Findings from Cox Analysis

1. **Positive Nodes**: The most powerful and statistically significant predictor. Each additional positive node increases the hazard of death by approximately 4.8% (HR $\approx$ 1.05 per node).
2. **Age**: No statistically significant effect (p=0.246). The hazard ratio of approximately 1.01 per year indicates a negligible age effect in this cohort.
3. **Operation Year**: No significant temporal trend (p=0.863). Patients operated on in later years did not have significantly different survival outcomes than those operated on earlier in the 1958-1969 period.

4.6 Bivariate Analysis Results

Comparisons between survived and died groups reveal key patterns:

Table 7: Group Comparison Statistics

Variable	**Survived (n=225)**	**Died (n=81)**	**Difference**
Age (mean ± SD)	52.0 ± 11.0	53.7 ± 10.2	1.7 years
Operation_year (mean ± SD)	62.9 ± 3.2	62.8 ± 3.3	0.1 years
Nb_pos_detected (mean ± SD)	2.8 ± 5.9	7.5 ± 9.2	4.7 nodes
Nb_pos_detected (median [IQR])	0.0 [0.0, 3.0]	4.0 [1.0, 11.0]	4.0 nodes

The three-fold difference in mean positive nodes (2.8 vs. 7.5) between survivors and non-survivors demonstrates the strong association between nodal involvement and survival outcome.

5 Conclusion

This comprehensive survival analysis of breast cancer patients from the Haberman dataset has yielded several important findings with both statistical and clinical significance.

5.1 Primary Findings

5.1.1 Predictor Importance

The analysis unequivocally establishes the number of positive axillary lymph nodes as the most significant prognostic factor for 5-year survival following breast cancer surgery. This finding is consistent across multiple analytical approaches:

1. **Bivariate analysis** revealed a 2.7-fold difference in mean positive nodes between survivors (2.8 nodes) and non-survivors (7.5 nodes).
2. **Log-rank tests** demonstrated highly statistically significant differences in survival curves across node categories (χ^2=36.89, p<0.001), with pairwise comparisons showing significant differences between patients with minimal (0-1 nodes) versus substantial ($\geq$10 nodes) nodal involvement.

3. **Cox regression** identified positive node count as the only statistically significant predictor in univariate models ($p<0.001$), with each additional node associated with increased hazard of death.

This finding aligns with established oncological knowledge that axillary lymph node involvement is a cornerstone of breast cancer staging and prognosis assessment. The American Joint Committee on Cancer (AJCC) staging system emphasizes nodal status as a critical prognostic indicator, and these results provide empirical validation of that principle.

5.1.2 Age and Temporal Effects

Contrary to intuitive expectations, patient age did not emerge as a significant predictor of survival in this cohort. The log-rank test for age groups ($p=0.390$) and the Cox regression coefficient for age ($p=0.246$) both indicate no statistically significant association between age and survival.

Similarly, operation year showed no significant temporal trend in survival outcomes. The Cox model yielded $p=0.863$ for operation year, indicating that patients operated on in later years (e.g., 1965-1969) did not experience improved survival compared to those operated on in earlier years (1958-1964). This finding may reflect:

- Relative stability in surgical techniques during this period
- Lack of major advances in breast cancer treatment during 1958-1969
- Comparable patient selection criteria across the study period

5.1.3 Overall Survival Outcomes

The overall 5-year survival rate of 73.5% represents a historical baseline for breast cancer surgery during the period 1958-1969. This rate is notable considering that it predates many modern oncological advances, including:

- Widespread use of adjuvant chemotherapy
- Targeted hormonal therapies
- Herceptin and other targeted biologic agents
- Advanced radiation therapy techniques

5.2 Clinical Implications

5.2.1 For Prognostic Assessment

The strong association between nodal involvement and survival underscores the importance of accurate axillary lymph node evaluation in breast cancer patients. Clinicians should:

- Prioritize thorough nodal staging during surgery
- Consider nodal status in treatment decision-making

- Use nodal categories (0, 1-3, 4-9, ≥10 nodes) as prognostic indicators
- Recognize that patients with ≥10 positive nodes require aggressive treatment approaches

5.2.2 For Patient Counseling

Age alone should not be used as a primary prognostic factor, as this analysis demonstrates no significant age effect within the range examined (30-83 years). Patients and families should be counseled that survival outcomes are more strongly associated with disease burden (nodal involvement) than chronological age.

5.3 Study Limitations

This analysis has several important limitations that must be acknowledged:

1. **Pseudo Time-to-Event Data**: The original dataset contains only binary 5-year survival outcomes. Time-to-event data were artificially constructed by randomly distributing death times across 60 months. This approach, while commonly used in survival analysis with binary outcomes, introduces uncertainty in the exact timing of events.
2. **Historical Cohort**: Data collected from 1958-1969 represents treatment and practice from over 50 years ago. Treatment standards, surgical techniques, and patient characteristics may differ substantially from contemporary practice.
3. **Limited Covariates**: The dataset includes only three predictor variables. Important clinical factors such as tumor size, hormone receptor status, histological grade, and treatment modalities are not available.
4. **Class Imbalance**: The strong imbalance (73.5% vs. 26.5%) may affect statistical power for detecting effects and warrants caution in model interpretation.
5. **Sample Size**: With 306 observations, the study has moderate statistical power, particularly for subgroup analyses involving smaller patient groups.

5.4 Statistical Methodological Insights

5.4.1 Kaplan-Meier Estimation

The non-parametric Kaplan-Meier estimator proved effective for describing survival patterns without parametric assumptions. The method handled the heavy censoring (73.5% censored) appropriately and provided interpretable survival probabilities at clinically relevant timepoints.

5.4.2 Log-Rank Testing

The log-rank tests effectively distinguished between covariates with true prognostic value (node category) versus those without (age group). The methodology successfully controlled for multiple comparisons in pairwise testing while maintaining type I error control.

5.4.3 Cox Regression

While the Cox model provided estimates of covariate effects, the pseudo time-to-event construction may have influenced the hazard ratio estimates. The proportional hazards assumption appeared reasonable based on the stratified survival curves.

5.5 Future Research Directions

Several directions for extending this analysis could provide additional insights:

1. **Multivariate Cox Models**: Extending the analysis to fit multivariable Cox models with all covariates simultaneously would allow assessment of independent effects while controlling for confounders.

2. **Interaction Effects**: Investigating interactions between covariates (e.g., Age × Nodal involvement) could reveal more complex prognostic patterns.

3. **Time-Varying Covariates**: If true time-to-event data become available, time-varying Cox models could capture dynamic changes in risk factors.

4. **Contemporary Data Comparison**: Comparing these historical results with contemporary breast cancer datasets would quantify improvements in survival outcomes over time.

5. **Advanced Survival Models**: Incorporating parametric models (Weibull, log-normal) or frailty models could provide complementary insights beyond the non-parametric approach.

5.6 Final Remarks

5.6.1 Methodological Contribution

This survival analysis successfully demonstrates the application of modern statistical methods—Kaplan-Meier estimation, log-rank testing, and Cox regression—to real-world clinical data. The analysis illustrates the power of survival analysis techniques for addressing censored time-to-event outcomes, which are ubiquitous in clinical research. The combination of descriptive statistics, non-parametric estimation, and semi-parametric regression provides a comprehensive framework for understanding survival patterns.

The statistical rigor and methodological approach employed here serve as a template for future survival analyses in oncological research.

5.6.2 Clinical Implications

The findings provide empirical validation of nodal status as a critical prognostic factor while revealing insights that may inform clinical practice. From a broader perspective, this study contributes to the body of evidence supporting the prognostic significance of axillary lymph node involvement in breast cancer.

The findings emphasize the importance of thorough clinical staging and accurate assessment of disease burden, while also highlighting limitations inherent in historical datasets. As treatment advances continue to improve breast cancer outcomes, comparative analyses between historical and contemporary cohorts become increasingly valuable for quantifying progress in oncological care.

5.6.3 Summary

In summary, this comprehensive survival analysis has successfully identified nodal involvement as the dominant prognostic factor while demonstrating the methodological approaches appropriate for analyzing censored survival data. The results have clinical relevance for prognostic assessment and contribute to the statistical learning literature through the application of advanced survival analysis techniques.

6 Feedback and Response to Presentation

Based on feedback received during the presentation, the following clarifications and considerations are provided:

6.1 Model Specification

6.1.1 Operation Year as an Independent Variable

A critical consideration raised during the presentation pertains to the treatment of operation year as an independent variable in the Cox regression model. Operation year should not be considered a truly independent prognostic factor, as it reflects temporal trends in treatment protocols, surgical techniques, and patient selection criteria rather than inherent patient characteristics affecting survival.

However, operation year serves an important analytical purpose as a temporal control variable. Its inclusion in the model allows us to test for temporal trends in survival outcomes across the 1958-1969 period, which proved to be non-significant (p=0.863). This finding indicates that survival outcomes remained relatively stable across the study period, providing important context for interpreting the results.

6.1.2 Final Model Selection

Following the univariate Cox regression analysis, which identified *Nb_pos_detected* as the only statistically significant predictor (p<0.001), the final model is:

Final Univariate Cox Model

$$h(t|\text{Nb\_pos\_detected}) = h_0(t)e^{\beta\cdot\text{Nb\_pos\_detected}} \tag{12}$$

where:

- $h(t|\text{Nb\_pos\_detected})$ = hazard function at time t given nodal count
- $h_0(t)$ = baseline hazard function (unspecified)
- $\beta = 0.047$ = regression coefficient for number of positive nodes
- $\text{HR} = e^{0.047} = 1.048$ = hazard ratio (4.8% increase in hazard per additional node)

The final model parameters are summarized in Table 8.

Table 8: Final Univariate Cox Model for Breast Cancer Survival

Variable	**Coefficient (β)**	**HR**	**95% CI**	**P-value**
Nb_pos_detected	0.047	1.048	[1.030, 1.068]	<0.001

6.2 Statistical Inference and Interpretation

6.2.1 Model Justification

The selection of a univariate model is statistically justified and methodologically sound for several reasons:

1. **Significance Criteria**: Age and operation year both failed to meet statistical significance thresholds ($p=0.246$ and $p=0.863$, respectively) in univariate models, providing no evidence to warrant their inclusion.
2. **Principle of Parsimony**: Occam's Razor favors the simpler model that achieves comparable explanatory power. Including non-significant variables would not improve model fit while introducing unnecessary complexity.
3. **Multicollinearity Consideration**: While correlation analysis (Table 3) revealed minimal inter-feature correlations, the non-significant variables provide no additional prognostic information beyond nodal count.
4. **Clinical Replicability**: A single-predictor model with strong statistical significance ($p<0.001$) and interpretable effect size (HR=1.048) facilitates clinical application and replication across different patient populations.

6.2.2 Empirical Evidence Supporting the Final Model

The final univariate model is supported by multiple lines of converging evidence:

- **Descriptive Statistics**: Three-fold difference in mean nodal count between survivors and non-survivors (2.8 vs. 7.5 nodes)
- **Log-Rank Tests**: Highly significant differences in survival curves across node categories ($\chi^2=36.89$, $p<0.001$)
- **Cox Regression**: Statistically significant hazard ratio (HR=1.048, $p<0.001$) with narrow 95% confidence interval [1.030, 1.068] that excludes the null value of 1.0
- **Biological Plausibility**: Aligns with established oncological knowledge that nodal involvement is a cornerstone of cancer staging

6.2.3 Model Transparency and Acceptability

To enhance the acceptability and interpretability of the statistical inference, the model demonstrates:

1. **Reproducibility**: All statistical analyses are fully documented with code available in the GitHub repository (Section 7)
2. **Effect Size Quantification**: The hazard ratio of 1.048 per node provides a concrete, interpretable measure of prognostic impact
3. **Uncertainty Quantification**: 95% confidence intervals explicitly account for sampling variability and provide a range of plausible values

4. **Multiple Testing Approach**: The analysis employed univariate models for each predictor, allowing for transparent assessment of individual effects before considering multivariate approaches

5. **Clinical Applicability**: The simple model structure facilitates straightforward clinical application—clinicians can calculate individual patient risk by multiplying their nodal count by the hazard ratio coefficient

6.3 Multivariate Model Consideration

While the presentation focused on univariate models to establish individual predictor effects, a natural extension would be to construct a multivariate Cox model incorporating all covariates simultaneously:

$$h(t|X) = h_0(t)e^{\beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Operation\_year} + \beta_3 \cdot \text{Nb\_pos\_detected}} \tag{13}$$

However, given the non-significance of Age and Operation_year in univariate analysis, and the lack of evidence for confounding or interaction effects, the multivariate model would likely yield similar conclusions with nodal count remaining the sole significant predictor. The univariate approach adopted here provides clearer, more interpretable results while maintaining statistical rigor.

6.4 Summary of Feedback Response

In summary, the final model—a univariate Cox regression with positive node count as the sole predictor—is statistically justified, clinically interpretable, and supported by multiple converging lines of evidence. The approach of starting with univariate models before considering multivariate extensions represents sound statistical practice, particularly when identifying which covariates warrant inclusion in more complex models. The non-inclusion of age and operation year as independent prognostic factors is scientifically appropriate given their lack of statistical significance in this cohort.

7 Code and Data Availability

The complete source code, data analysis scripts, generated plots, and datasets used in this study are available in the GitHub repository: `https://github.com/pleyva2004/Statistical-Learning-Capstone`. The repository includes:

- Jupyter notebook (`ProjectTwo.ipynb`) with complete data analysis
- Python scripts for plot generation
- All visualization PNG files referenced in this report
- The Haberman breast cancer survival dataset