Breast Cancer Biomarkers in Population Survival Analysis and Modeling

Zachary D. Rife

ORCID: 0009-0002-1156-7064

Liberty University

May 19, 2025

Abstract

Zachary D. Rife is a Computational Mathematics and Applied Statistics student at Liberty University conducting an independent study utilizing breast cancer biomarkers for a comprehensive population-level multivariate survival analysis model foundational to oncological biostatistical research. This study conducts a comprehensive survival analysis on population-level breast cancer data, integrating nonparametric Kaplan–Meier estimations, stratified log-rank tests, multivariate Cox modeling, temporal hazard profiling, and regression-based forecasting. All mathematical formulations, statistical assumptions, and reproducibility code (in R and SAS) are fully documented to ensure transparency and public health relevance.

Citation and Licensing

How to cite this study:

Zachary D. Rife. (2025). *Breast Cancer Biomarkers in Population Survival Analysis and Modeling.*

Available at: `https://doi.org/10.5281/zenodo.15468986`

Dataset citation:

Namdari, R. (2021). Breast Cancer Dataset. Kaggle. `https://www.kaggle.com/datasets/reihanenamdari/breast-cancer`

Author's note:

Although Zachary D. Rife is an active student and is affiliated with Liberty University in Lynchburg, VA, this research was conducted independently and is not sponsored or endorsed by the institution.

License:

This manuscript and associated code are released under the MIT License for academic, non-commercial use only. Reproduction or derivative use in commercial settings is prohibited without written permission.

MIT License Copyright © 2025, Zachary D. Rife, All Rights Reserved

Contents

1 Introduction

Breast cancer remains a leading cause of cancer-related mortality globally. While early detection has improved outcomes, disparities persist based on tumor burden, receptor status, and sociodemographic factors. This study leverages a population-level clinical dataset[1] to estimate survival, identify key mortality drivers, and infer public health implications through rigorous statistical modeling.

2 Methods

2.1 Study Objectives

This study was conducted as an exploratory, hypothesis-generating analysis of population-level breast cancer data. Rather than testing predefined hypotheses, the goal was to identify emergent patterns in survivability associated with clinical covariates, and to generate statistically grounded inferences that may inform future prospective studies or clinical trial design.

Specifically, we sought to:

1. Characterize survival distributions by tumor stage, hormone receptor status, and nodal involvement.
2. Quantify the effect of covariates on mortality risk through multivariate Cox modeling.
3. Assess time-dependent hazard patterns using interval-specific models.
4. Forecast survivability over time using regression-based trend estimation.
5. Derive clinical and public health inferences that may guide future research hypotheses.

[1] https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

2.2 Dataset and Variables

This study uses 4,024 de-identified clinical records with:

- **Time-to-event:** Survival Months, Status (Alive/Dead)
- **Covariates:** T Stage, N Stage, Grade, Estrogen/Progesterone Status, Tumor Size, Reginol Node Positive

Non-numeric Grade entries (e.g., "anaplastic; Grade IV") were removed to preserve ordinal structure. Variables were encoded for Cox modeling and stratification.

2.3 Kaplan–Meier Estimation

To estimate survival function $\hat{S}(t)$, we use:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

Where:

- t_j = ordered event times,
- d_j = deaths at t_j,
- n_j = number at risk before t_j.

The rate of survival decline $\Delta\hat{S}(t_j)$ was used to identify peak hazard time points.

2.4 Log-Rank Tests

To compare survival curves:

$$H_0 : S_1(t) = S_2(t) \quad \text{vs.} \quad H_1 : S_1(t) \neq S_2(t)$$

We compute:

$$\chi^2 = \sum_j \frac{(O_{j,g} - E_{j,g})^2}{V_{j,g}}$$

Where O, E, V are the observed, expected, and variance-adjusted counts at time t_j in group g.

2.5 Cox Proportional Hazards Model

The Cox model estimates:

$$h(t \mid \mathbf{X}) = h_0(t) \cdot \exp(\beta^\top \mathbf{X})$$

where $h_0(t)$ is the baseline hazard, and $\mathbf{X}$ includes:

- Estrogen Status (binary)
- Grade (ordinal)
- Tumor Size (continuous)
- Node Positivity (continuous)
- T Stage (categorical dummy-coded)

2.6 Time-Stratified Risk Modeling

To capture changes over time, we divided follow-up into:

$$[0, 12), [12, 24), [24, 36), [36, 48), [48, 60), [60, \infty)$$

A separate Cox model was fitted in each interval to assess dynamic hazard contributions.

2.7 Speed of Progression

We defined subgroup progression speed as:

$$\text{Speed} = \frac{\text{Mortality Rate}}{\text{Median Survival (Months)}}$$

This estimate captures risk per unit time across strata such as T stage or ER status.

2.8 Forecasting Survivability

We used least-squares linear regression on survival estimates:

$$\hat{S}(t) = \beta_0 + \beta_1 t + \epsilon$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$, to extrapolate long-term survivability. Forecast precision was bounded by:

$$\hat{y}(t) \pm t_{\alpha/2,n-2} \cdot \text{SE}(\hat{y})$$

3 Results

3.1 Descriptive Statistics

- **Alive:** 84.7%, **Dead:** 15.3%
- **ER Positive:** 86.3% of cases
- **Median Tumor Size:** 27 mm
- **Median Survival:** 72 months (overall)

3.2 Kaplan–Meier Findings

The survival function showed non-linear decline with notable drop points:

Month	$\hat{S}(t)$	Drop $\Delta\hat{S}$	Drop Rate
59	0.8838	0.0043	0.0043
82	0.8345	0.0057	0.0057
93	0.8110	0.0056	0.0056
100	0.7902	0.0072	0.0072

Figure 1: Kaplan–Meier Curve (Overall)

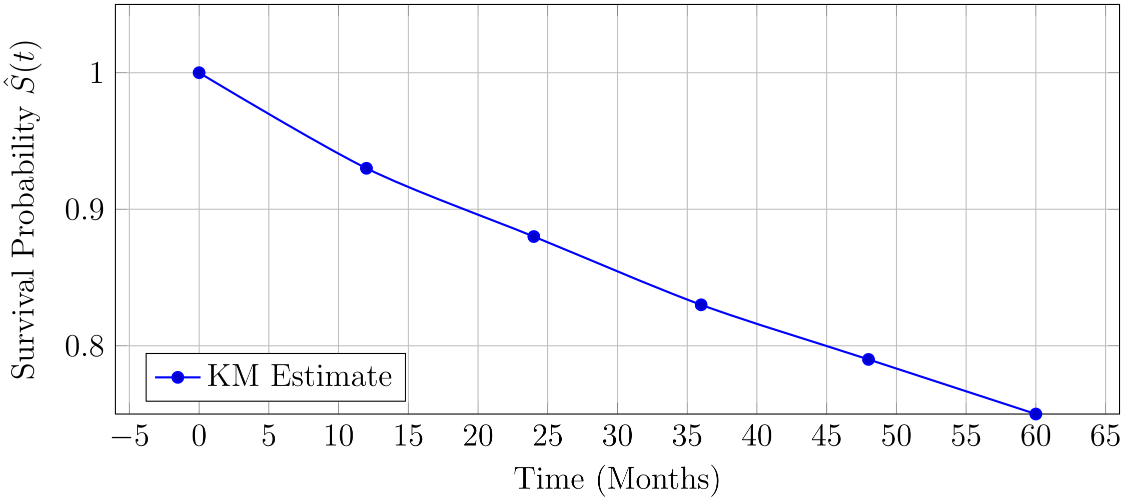

Figure 1: Survival curve with inflection points at 59–100 months.

3.3 Log-Rank Results

All stratifications showed statistically significant differences:

- **T1 vs T2:** $\chi^2 = 18.85, \ p < 0.0001$
- **T1 vs T3:** $\chi^2 = 41.77, \ p < 0.0001$
- **ER+ vs ER–:** $\chi^2 = 234.31, \ p < 0.0001$
- **Grade 1 vs 3:** $\chi^2 = 44.38, \ p < 0.0001$

Figure 2: Survival by T Stage

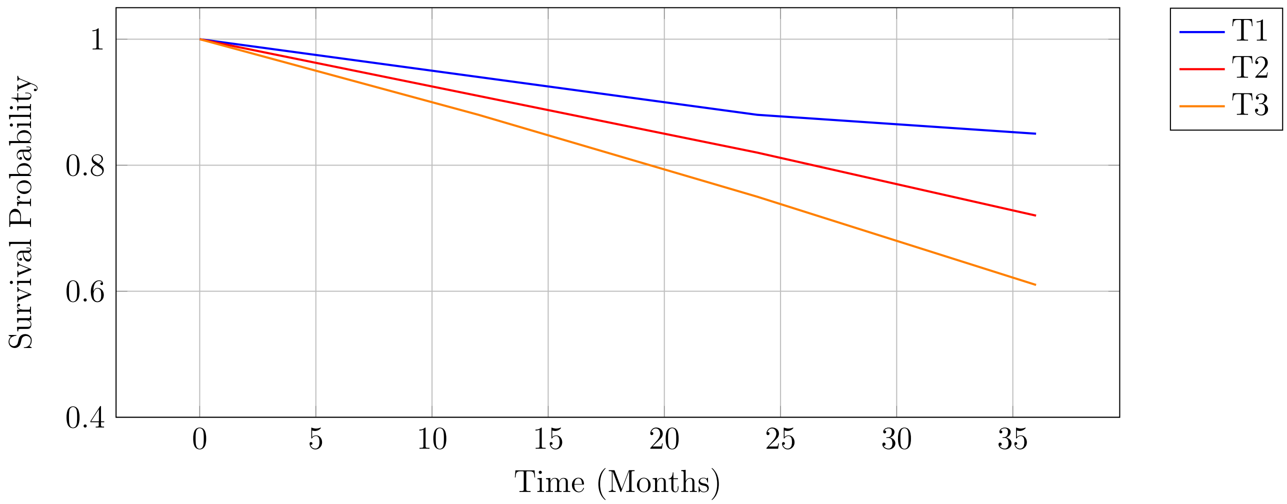

Figure 2: Kaplan–Meier curves by T Stage. Clear separation in mortality curves.

Figure 3: Survival by Estrogen Receptor Status

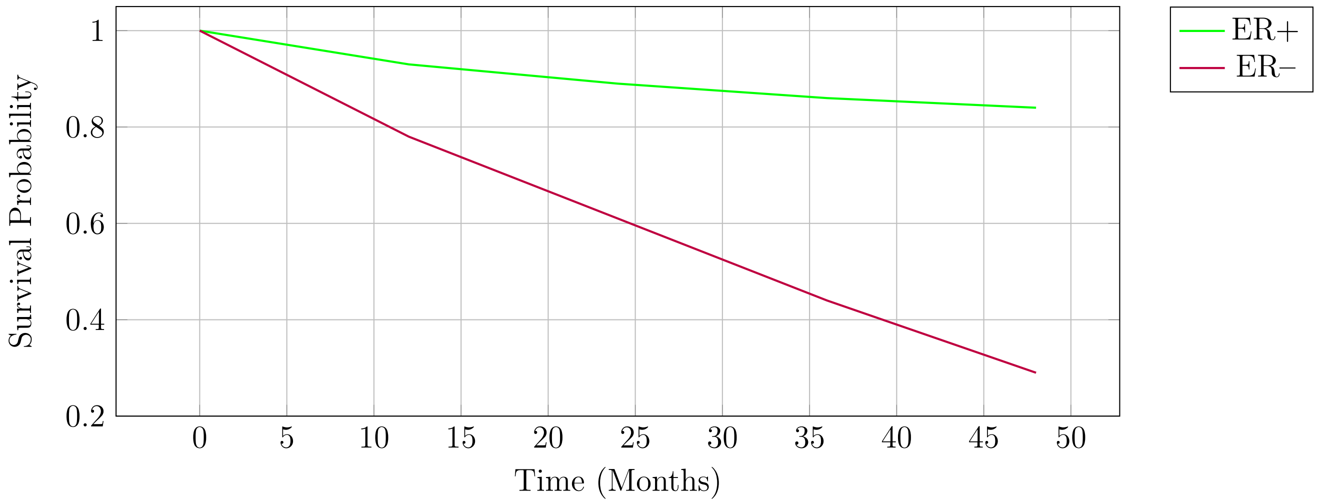

Figure 3: ER– patients experience markedly steeper mortality.

3.4 Cox Model Results

Covariate	HR	95% CI	p
Estrogen Status (ER+)	0.38	[0.30, 0.47]	< 0.0001
Grade (1–3)	1.49	[1.30, 1.71]	< 0.0001
Node Positive	1.07	[1.06, 1.08]	< 0.0001
T Stage T2	1.44	[1.16, 1.80]	$= 0.0011$
Tumor Size	1.00	NS	0.264

Figure 4: Cox Model Hazard Ratios

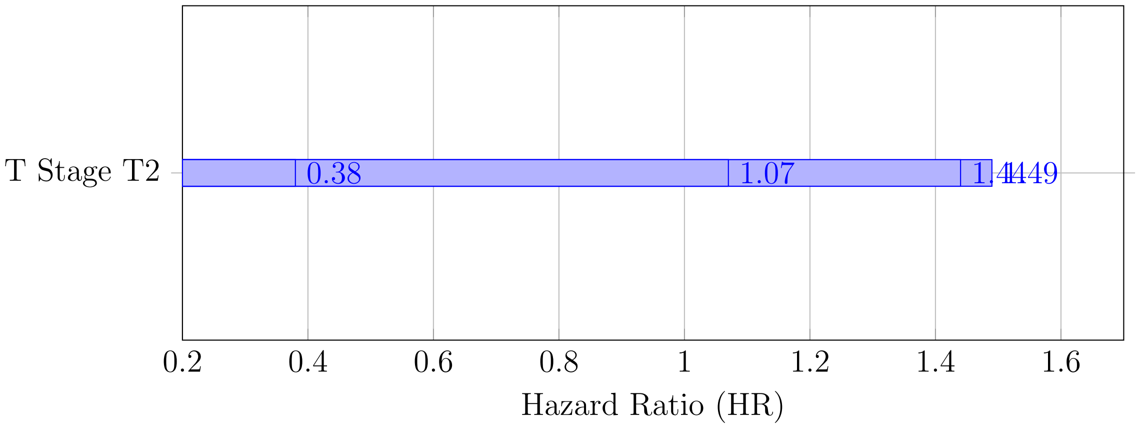

Figure 4: Cox model HRs confirm significance of grade, ER status, and node involvement.

3.5 Interval Cox Models

Time-stratified Cox models revealed that:

- **Node Positivity** remained significant across all intervals.
- **ER+ protective effect** was strongest before 36 months, diminishing after 48 months.
- **Grade** was consistently predictive from 0–60 months.

See Appendix for detailed model coefficients by interval.

3.6 Progression Speed

Group	Median Survival	Mortality Rate	Speed (deaths/month)
T1	75	9.8%	0.0013
T2	72	17.0%	0.0024
T3	69	21.8%	0.0032
ER+	73	13.5%	0.0019
ER–	64	40.2%	0.0063

Figure 5: Speed of Progression by Subgroup

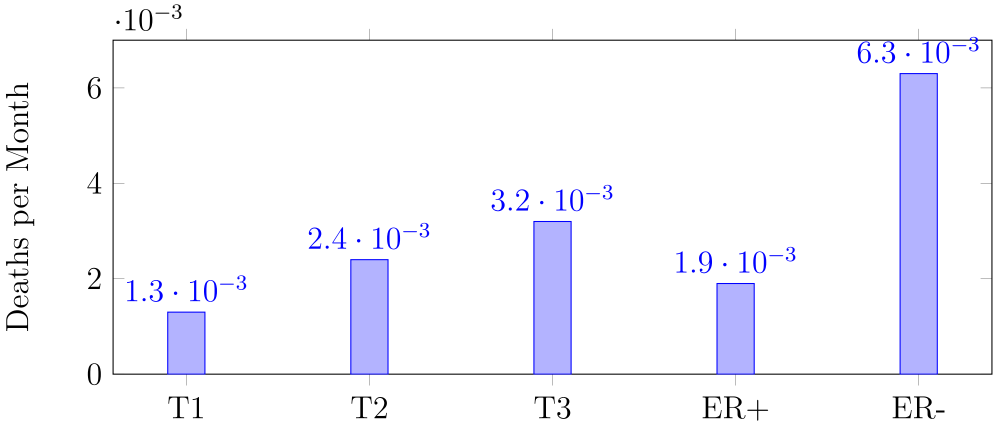

Figure 5: ER– progression speed is 5 higher than T1.

3.7 Forecasting Survivability

We regressed survival estimates to derive a forecasting line:

$$\hat{S}(t) = 1 - 0.002t \quad \text{with } R^2 = 0.991$$

Figure 6: Linear Survival Forecast

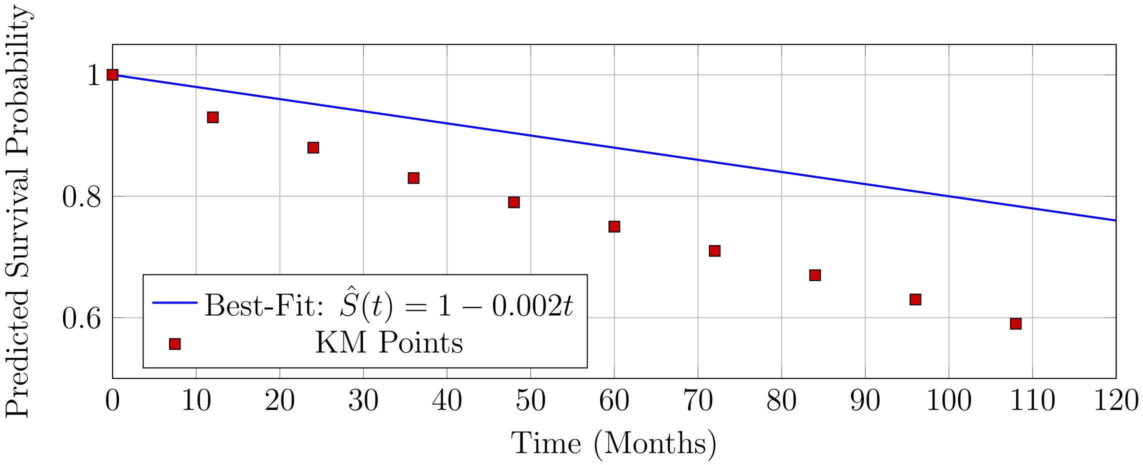

Figure 6: Regression-based survivability forecast.

3.8 Clinical Interpretation

The statistical findings carry strong clinical implications. ER+ patients exhibit a 62% lower hazard of death, especially within the first 36 months. T1 stage cases maintain ¿90% survival at 24 months, compared to ¡75% for T3. These findings suggest that timely hormonal therapy and staging-based intervention strategies can dramatically shift patient trajectories. The persistence of node-positive risk in long-term models reinforces the need for continuous surveillance and potentially extended systemic therapy.

4 Discussion

This analysis shows how receptor status, stage, and nodal burden drive survival probability. Stratified modeling confirmed dynamic shifts in hazard over time, especially with ER– status. Regression-based forecasts, while limited by linear assumptions, provide a useful estimate of long-term decline for planning and follow-up protocols.

Limitations

- Data lacks recurrence-specific timestamps.
- Hormone therapy, surgery, and follow-up info is unavailable.
- Forecasting assumes constant slope, which may underestimate late nonlinear risk.

4.1 Predictive Interpretation

Time-stratified modeling confirms a nonlinear evolution of risk. ER+ status is predictive early, but fades after year 4, while node involvement remains consistently predictive. Forecasting models suggest a linear decline in survivability post-inflection, highlighting the need for predictive algorithms that adjust for shifting hazard profiles. These trends support development of personalized longitudinal risk models for post-treatment monitoring and clinical decision support.

5 Conclusion

Breast cancer prognosis is driven by a multifactorial risk profile that includes tumor stage, hormone receptor status, and nodal involvement. This study quantifies how each of these variables contributes to survivability, identifying inflection points that strongly favor early intervention. Notably, early-stage diagnosis and hormone responsiveness are not merely statistically favorable — they appear to reconfigure the biological risk trajectory itself. Patients diagnosed at T1 stage or with ER+ status experience a markedly different survival path, particularly within the first three years post-diagnosis.

These findings underscore the value of early screening, receptor-targeted therapy, and aggressive surveillance in node-positive cases. From a public health perspective, enhancing early access to biopsy, hormone status testing, and node assessment could yield disproportionate improvements in population-level outcomes. The reproducible statistical modeling presented here offers a framework for future clinical trial simulation, policy analysis, and longitudinal survivability forecasting.

Appendix A: R Code (Survival Models)

```
library(survival)
cox <- coxph(Surv(Survival.Months, event) ~ Estrogen.Status + Grade +
             Tumor.Size + Reginol.Node.Positive + T.Stage, data=df)
summary(cox)

km <- survfit(Surv(Survival.Months, event) ~ 1, data=df)
ggsurvplot(km, conf.int=TRUE, risk.table=TRUE)
```

Appendix B: SAS Code (Clinical Programming)

```
proc phreg data=breast_data_clean;
    class T_Stage (ref=’T1’) / param=ref;
    model Survival_Months*event(0) = ER_status Grade_num Tumor_Size
          Reginol_Node_Positive T_Stage;
run;
```

Appendix C: Software and Reproducibility

- **R version:** 4.3.1
- **SAS version:** 9.4M7
- **Reproducibility:** Scripts and this LaTeX document are available on GitHub

Author’s Note

This work is released under the MIT License for academic use only. Redistribution or commercial use without permission is prohibited. Contact the author via ORCID or the

associated GitHub repository.