METABRIC Breast Cancer Final Project

Survival Analysis and Predictive Modeling

Presenter: Miro Everaert and Dylan Holt
Date: 08/10/2025

Outline

1. Brief Refresher on the METABRIC Dataset & Its Variables
2. Feature Reduction with PCA & Relationships to Survival Outcomes
3. Survival Models comparing Traditional and Modern ML Methods

Brief Refresher on the METABRIC Dataset

Key Takeaways

Dataset Link: [Breast Cancer Gene Expression Profiles (METABRIC)]

Dataset Size: 1,904 patients with comprehensive clinical and genomic data

Clinical Relevance: Largest integrated breast cancer genomics study to date

Survival Endpoints: Overall Survival (OS) and Disease-Specific Survival (DSS)

Modeling Goal: Predict survival time using clinical and molecular features

Expected Outcome: Identify key prognostic factors for personalized treatment

The METABRIC Study

(Molecular Taxonomy of Breast Cancer International Consortium)

Study Overview:

- **Study Period:** 2006-2012 (Nature 2012, Nature Communications 2016)
- **Sample Size:** 1,904 primary breast cancer patients
- **Geographic Scope:** UK and Canada
- **Follow-up:** Long-term clinical outcomes (up to 20+ years)

Why this Dataset matters:

- Enables precision medicine approaches
- Informs treatment stratification
- Supports drug development priorities
- Validates biomarker signatures

Dataset Structure & Variables

31 Clinical Attributes

- Demographics: Age at diagnosis, menopausal status
- Tumor characteristics: Size, grade, stage, lymph node status
- Treatment history: Chemotherapy, hormone therapy, radiotherapy
- Pathology: ER/PR/HER2 status, histological type

506 Genomic Variables

- m-RNA levels z-score for 331 genes
- mutation in 175 genes

Survival Outcomes

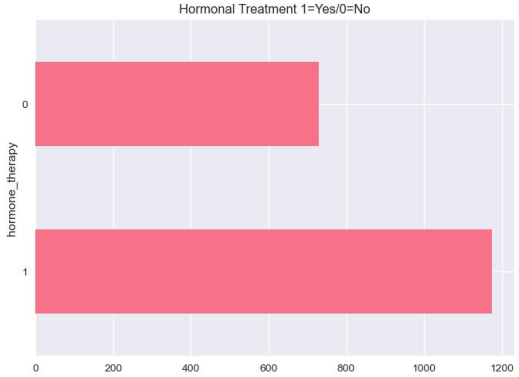
- Overall Survival (OS):
  Time to death from any cause
- Disease-Specific Survival (DSS):
  Time to breast cancer death
- Relapse-Free Survival (RFS):
  Time to disease recurrence

Exploratory Data Analysis

Clinical Indicators

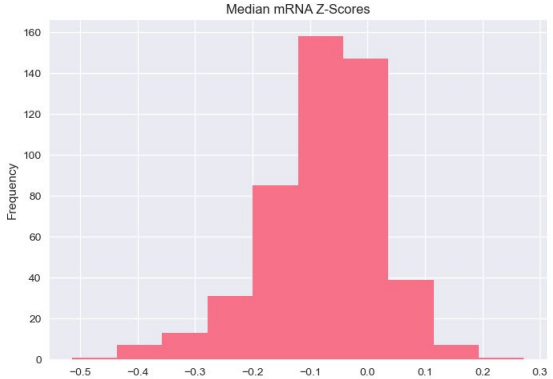
The mean age of 61 years aligns perfectly with **established epidemiological patterns**, while the range suggests early-onset cases are also detected.

Treatment Exposure



The standard of care for breast cancer is hormonal therapy, especially when breast cancers are **ER/PR positive.** This is more so ideal for the average patient in this dataset who is **61 years old and postmenopausal.**

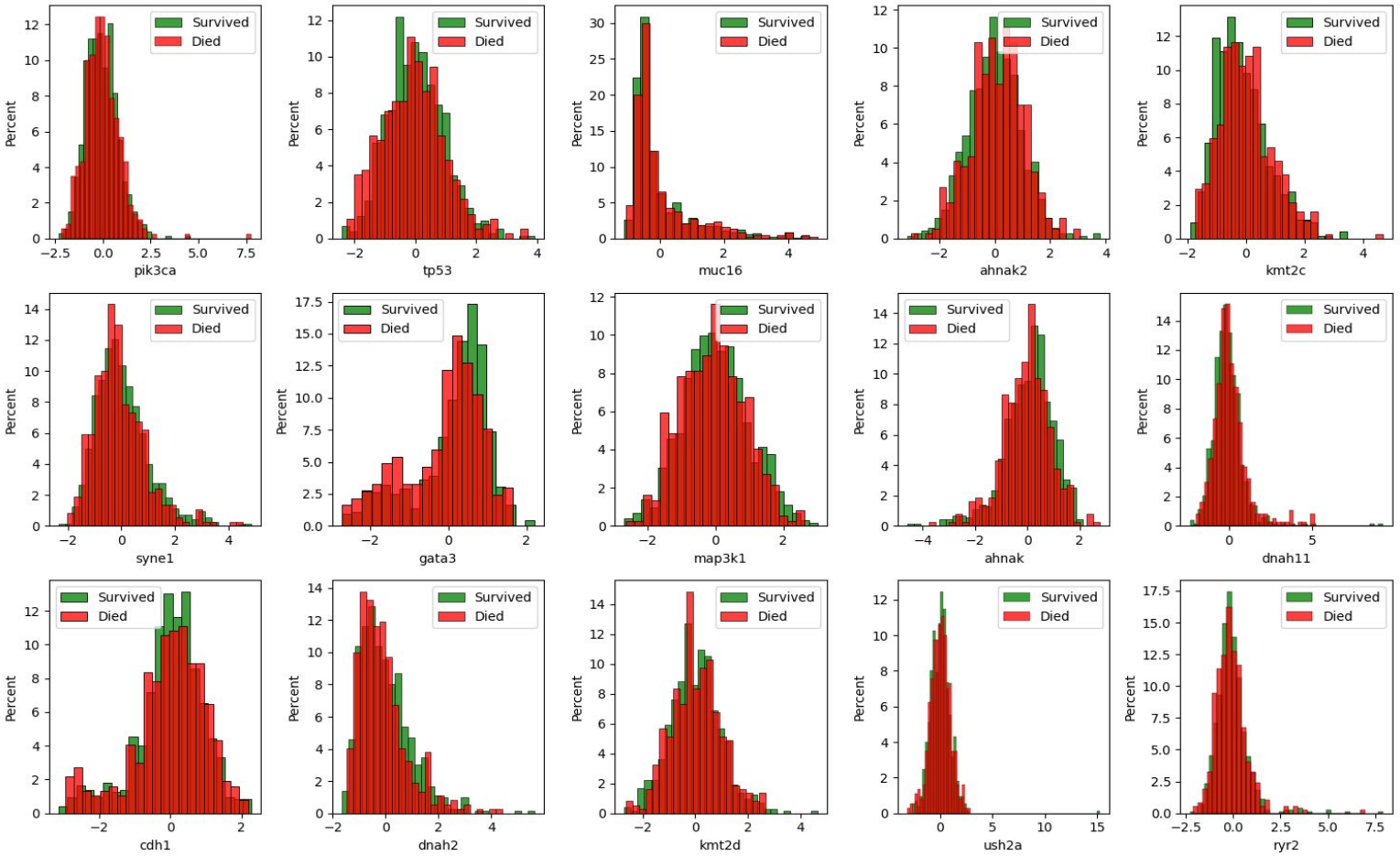
Genomic Variables



The slight negative skew indicates that **tumor suppressor genes are more commonly silenced than oncogenes are activated**. This suggests that **loss of tumor suppression** is the predominant molecular mechanism driving breast cancer development in this cohort.

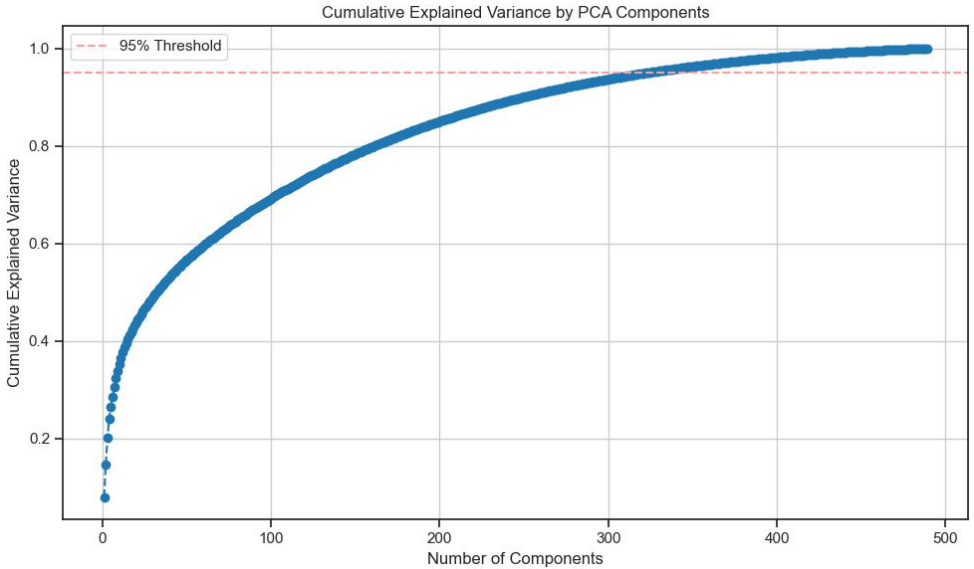
Feature Reduction with PCA & Relationships to Survival Outcomes

Survival of patients with some of gene mutations.



PCA on Gene Mutation Data

- We can use PCA to reduce the dimensions of the 489 gene mutation features for inclusion in simple models
  - Multiple linear regression to predict **overall_survival_months**
  - Logistic regression to predict **death_from_cancer**



Model Specification

- Target:
  - **death_from_cancer** using logistic regression
  - **Overall_survival_time** using linear regression
- Clinical features: `age_at_diagnosis, mutation_count, neoplasm_histologic_grade, nottingham_prognostic_index, tumor_size, tumor_stage, radio_therapy, chemotherapy, hormone_therapy, lymph_nodes_examined_positive`
- First 25 PCs of gene mutation data (~40% explained variance)
- Build and test two versions of each model (clinical only vs clinical and gene mutation) to observe the change in model accuracy by including gene mutation data.

Multiple Linear Regression for overall_survival_time

- Without gene mutation data: MSE of 5,472 and R2 of 0.15
- With gene mutation data: MSE of 4,777 and R2 of 0.25

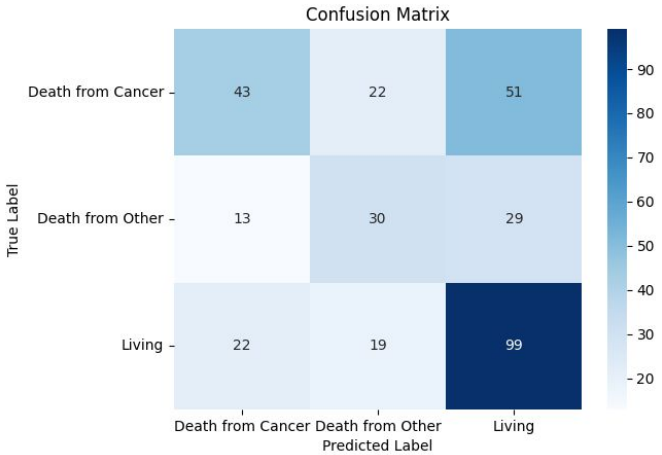
OLS Regression Results

Dep. Variable:	overall_survival_months	R-squared:	0.156
Model:	OLS	Adj. R-squared:	0.136
Method:	Least Squares	F-statistic:	7.890
Date:	Sun, 10 Aug 2025	Prob (F-statistic):	5.71e-26
Time:	15:52:18	Log-Likelihood:	-6201.6
No. Observations:	1092	AIC:	1.246e+04
Df Residuals:	1066	BIC:	1.259e+04
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	126.5455	2.169	58.337	0.000	122.289	130.802
PC1	-1.5255	0.369	-4.134	0.000	-2.249	-0.802
PC2	0.2106	0.382	0.552	0.581	-0.538	0.959
PC3	0.8983	0.424	2.120	0.034	0.067	1.730
PC4	3.0771	0.529	5.822	0.000	2.040	4.114
PC5	-2.2806	0.616	-3.701	0.000	-3.490	-1.072
PC6	-0.0194	0.671	-0.029	0.977	-1.336	1.297
PC7	-2.9857	0.685	-4.357	0.000	-4.330	-1.641
PC8	-2.1602	0.746	-2.895	0.004	-3.624	-0.696
PC9	-2.6527	0.762	-3.480	0.001	-4.149	-1.157
PC10	-2.1072	0.817	-2.578	0.010	-3.711	-0.503
...						

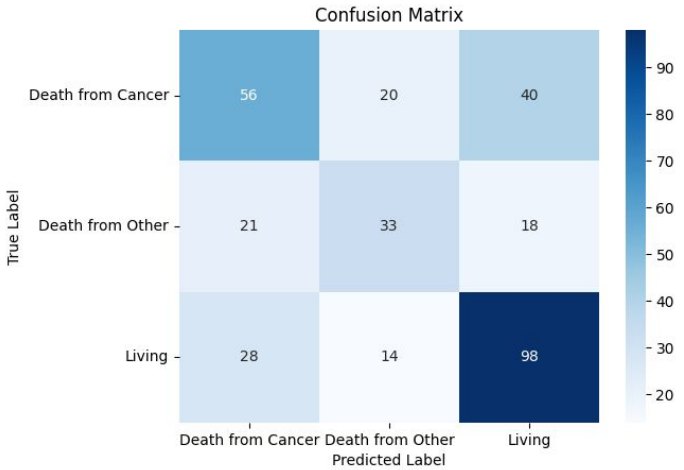
Logistic Regression for death_from_cancer

Without Gene Data



	Precision	Recall
Death from Cancer	55%	37%
Death from Other	42%	42%
Living	55%	71%

With Gene Data First 25 PCs



	Precision	Recall
Death from Cancer	53%	48%
Death from Other	49%	46%
Living	63%	70%

Survival Models comparing Traditional and Modern ML Methods

Survival Analysis Framework

Survival Analysis Primer

Survival Analysis studies time-to-event data where the primary interest is the time until an event occurs (death, disease recurrence, etc.).

Censoring is used to handle patients who have not experienced the target event by the end of the study period (e.g., 5-year survival).

Evaluation metrics included Concordance Index (C-Index) and Integrated Brier Score (IBS).

Framework from Astley et al. (2024)

Astley, J. R., Reilly, J. M., Robinson, Wild, J. M., Hatton, M. Q., & S., Tahir, B. A. (2024). Explainable deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy. Radiotherapy and Oncology, 193, 110084.

Noted Survival Analysis is typically limited to Cox Proportional Hazards (CPH) models.

Expanded survival analysis in NSCLC to Random Survival Forest and Deep Learning models to capture non-linear relationships in survival

Study Parameters & Model Designs

Feature Sets

11 Clinical Features	['age_at_diagnosis', 'chemotherapy', 'er_status', 'her2_status', 'hormone_therapy', 'inferred_menopausal_state', 'lymph_nodes_examined_positive', 'nottingham_prognostic_index', 'pr_status', 'radio_therapy', 'tumor_stage']
All 489 mRNA Scores Features	['brca1', 'brca2', 'palb2, 'pten', 'tp53', 'atm', 'cdh1', 'chek2','nbn', 'nf1', 'stk11', 'bard1, 'mlh1', 'msh2', 'msh6', 'pms2', 'epcam', 'rad51c', 'rad51d', 'rad50', 'rb1', 'rbl1', 'rbl2', 'ccna1', 'ccnb1',..., 'tnk2', 'tulp4', 'ugt2b15', 'ugt2b17', 'ugt2b7']

Models

Cox Proportional Hazards (CPH) Regression	• Linear regression model on hazard function • Parameters: alpha=0.01
Random Survival Forest (RSF)	• Ensemble Model with series of decision trees adapted for censored data • Parameters: n_estimators=200, max_depth=5, min_samples_leaf=20

Data Context

Full Dataset No Features with Missing Values (Events: 801/1904)	**Reduced Dataset** Only Records with Tumor Stage (Events: 611/1403)

Evaluation Metrics

Hazard Ratios What is the instantaneous risk of he event at any time between groups?	**Concordance Index** Can the model distinguish high-risk from low-risk patients?

Traditional Survival Analysis

TABLE 1: Cox Regression Analysis with Clinical Variables Only (Reduced Dataset to include Tumor Stage)

Variable	Univariate HR (95% CI)	Uni p-value	Multivariate HR (95% CI)	Multi p-value
Age (years)	**0.880* (0.813, 0.952)**	**0.002***	**0.882* (0.815, 0.955)**	**0.002***
Tumor size (cm)	1.011 (0.934, 1.095)	0.785	N/A	N/A
Tumor stage	1.114* (1.029, 1.206)	0.008*	1.064 (0.983, 1.152)	0.127
Positive lymph nodes	1.192* (1.102, 1.291)	<0.001*	1.044 (0.964, 1.130)	0.291
Nottingham Prog. Index	1.066 (0.985, 1.154)	0.113	0.962 (0.889, 1.042)	0.341
ER positive	0.871* (0.805, 0.943)	<0.001*	1.015 (0.937, 1.098)	0.719
PR positive	0.912* (0.842, 0.987)	0.023*	0.946 (0.874, 1.024)	0.171
HER2 positive	**1.342* (1.240, 1.453)**	**<0.001***	**1.264* (1.167, 1.368)**	**<0.001***

* indicates p < 0.05

Key Metric

Hazard Ratio compares the instantaneous risk of experiencing the event at any given time point between different groups

Main Takeaway: 2 Significant multivariate predictors

- HER2 positive: HR=1.264, p=0.000 (risk factor)
- Age (years): HR=0.882, p=0.002 (protective)
- Several other features are significant in univariate tests

Traditional Survival Analysis

TABLE 2: Cox Regression Analysis with Clinical Variables and mRNA Score Variables (Reduced Dataset to include Tumor Stage)

Variable	Univariate HR (95% CI)	Uni p-value	Multivariate HR (95% CI)	Multi p-value
Age (years)	0.880* (0.813, 0.952)	0.002*	0.902* (0.833, 0.976)	0.011*
Tumor size (cm)	1.011 (0.934, 1.095)	0.785	N/A	N/A
Tumor stage	1.114* (1.029, 1.206)	0.008*	1.009 (0.932, 1.093)	0.819
Positive lymph nodes	1.192* (1.102, 1.291)	<0.001*	0.986 (0.910, 1.067)	0.720
Nottingham Prog.Index	1.066 (0.985, 1.154)	0.113	1.132* (1.046, 1.225)	0.002*
ER positive	0.871* (0.805, 0.943)	<0.001*	1.142* (1.055, 1.237)	0.001*
PR positive	0.912* (0.842, 0.987)	0.023*	0.906* (0.837, 0.981)	0.015*
HER2 positive	1.342* (1.240, 1.453)	<0.001*	2.587* (2.390, 2.800)	<0.001*
brca1	0.868* (0.802, 0.940)	<0.001*	1.213* (1.120, 1.313)	<0.001*
…	…	…	…	…
jak1	1.529* (1.413, 1.655)	<0.001*	1.176* (1.087, 1.273)	<0.001*
mtor	0.854* (0.789, 0.924)	<0.001*	0.832* (0.768, 0.900)	<0.001*
col12a1	1.178* (1.088, 1.275)	<0.001*	1.758* (1.624, 1.903)	<0.001*
hsd3b7	0.711* (0.657, 0.770)	<0.001*	0.958 (0.885, 1.037)	0.289
serpini1	1.129* (1.043, 1.222)	0.003*	1.276* (1.179, 1.381)	<0.001*
tnk2	0.772* (0.713, 0.835)	<0.001*	0.878* (0.811, 0.951)	0.001*

* indicates p < 0.05

Model Evaluation Outcomes

Model Design	Clinical Features Only	Plus mRNA Score Features
CPH (Full Dataset)	0.608 ± 0.020	0.606 ± 0.029
RSF (Full Dataset)	0.630 ± 0.025	**0.701 ± 0.018**
CPH (Reduced Dataset)	0.600 ± 0.028	0.578 ± 0.025
RSF (Reduced Dataset)	0.626 ± 0.020	**0.718 ± 0.016**

Key Metrics

C-Index: Discrimination (Higher is better)
Can the model distinguish high-risk from low-risk patients?

Main Takeaway: RSF outperforms CPH in each scenario

- RSF and CPH do not differ significantly with Clinical Features Only
- RSF gains significant advantage when mRNA scores are included
- CPH drops in performance when mRNA scores are included

Survival Models in Conclusion

1. **Clinical Impact:** Modern ML methods (RSF) are essential for leveraging genomic data in survival prediction, while traditional Cox models remain adequate for clinical-only assessments.

2. **Methodological Insight:** Non-linear relationships between genomic features and survival outcomes require sophisticated modeling approaches that can capture complex gene-gene and gene-clinical interactions.

3. **Precision Medicine Advancement:** The substantial performance gains with genomic data (C-index improvement from 0.626 to 0.718) demonstrate the potential for personalized survival prediction, moving beyond traditional clinical staging toward molecular-based prognostication.