

# Exploratory Survival Analysis of Breast Cancer Patients Using the METABRIC Dataset (Power BI)

## Introduction

This project presents an exploratory survival analysis of breast cancer patients using the publicly available METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset. The objective was to demonstrate the ability to preprocess clinical data, define survival endpoints, and perform time-to-event analysis using an interactive data analytics platform. Power BI was used as the primary environment to combine data cleaning, analytical logic, and visualization.

## Dataset

- Source: METABRIC breast cancer cohort (<https://www.kaggle.com/code/alexandervc/breast-cancer-metabric-data-survival-curves/input>)
- Sample size: ~2,500 patients
- Key clinical variables:
  - Overall Survival (months)
  - Relapse-Free Survival (months)
  - Survival status (Living / Deceased)
  - ER status
  - Molecular subtype (Pam50)
  - Treatment and pathological features

## Data Preparation

Data preprocessing was performed using Power Query, including:

- Conversion of time-to-event variables to numeric format
- Creation of binary event indicators:
  - OS\_Event (1 = death, 0 = censored)
  - RFS\_Event (1 = relapse, 0 = censored)
- Handling of missing values using clinically appropriate censoring logic
- Validation of data consistency prior to analysis

This ensured that the dataset was suitable for survival analysis.

## Survival Analysis

The following analyses were implemented:

### Median Survival Estimates

- Median Overall Survival (OS)
- Median Relapse-Free Survival (RFS)

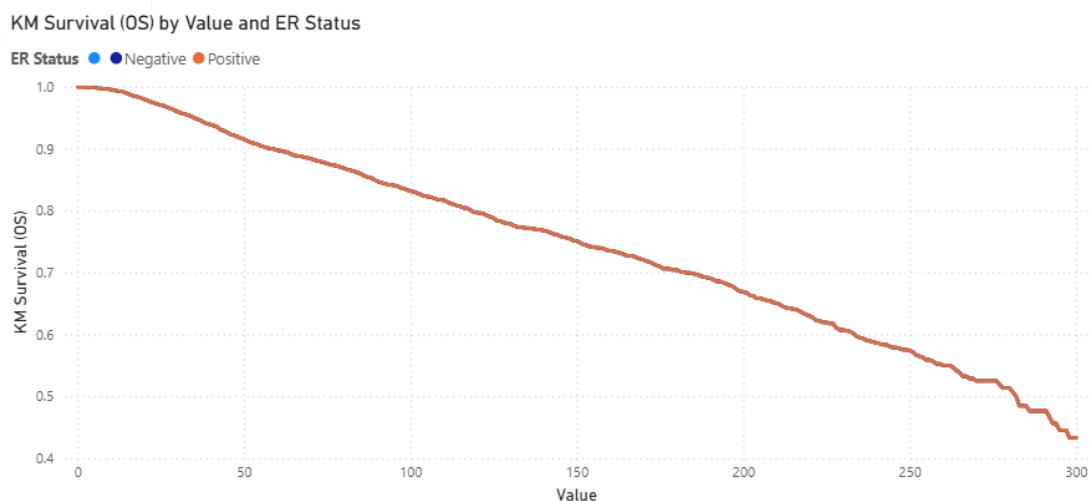
These were calculated dynamically and respond to cohort filters.

## Kaplan–Meier–Style Survival Curve

A Kaplan–Meier–style survival curve for Overall Survival was constructed using:

- Discrete monthly time bins
- Risk-set–based estimation logic
- Censoring-aware survival probability calculation

The survival curve demonstrates a clinically plausible decline in survival probability over time, consistent with published METABRIC analyses.



This Kaplan–Meier plot suggests no meaningful difference in overall survival between ER-positive and ER-negative patients in this dataset.

## **Stratified Survival Exploration**

Survival curves were explored interactively using clinical stratifications such as:

- Estrogen Receptor (ER) status
- Molecular subtype (Pam50)

This allowed visual comparison of survival trends across biologically meaningful groups.

## **Tools & Skills Demonstrated**

- Survival analysis concepts (censoring, time-to-event data)
- Clinical data preprocessing
- Power Query (ETL)
- DAX measures for analytical logic
- Interactive data visualization
- Translating biomedical questions into analytical workflows