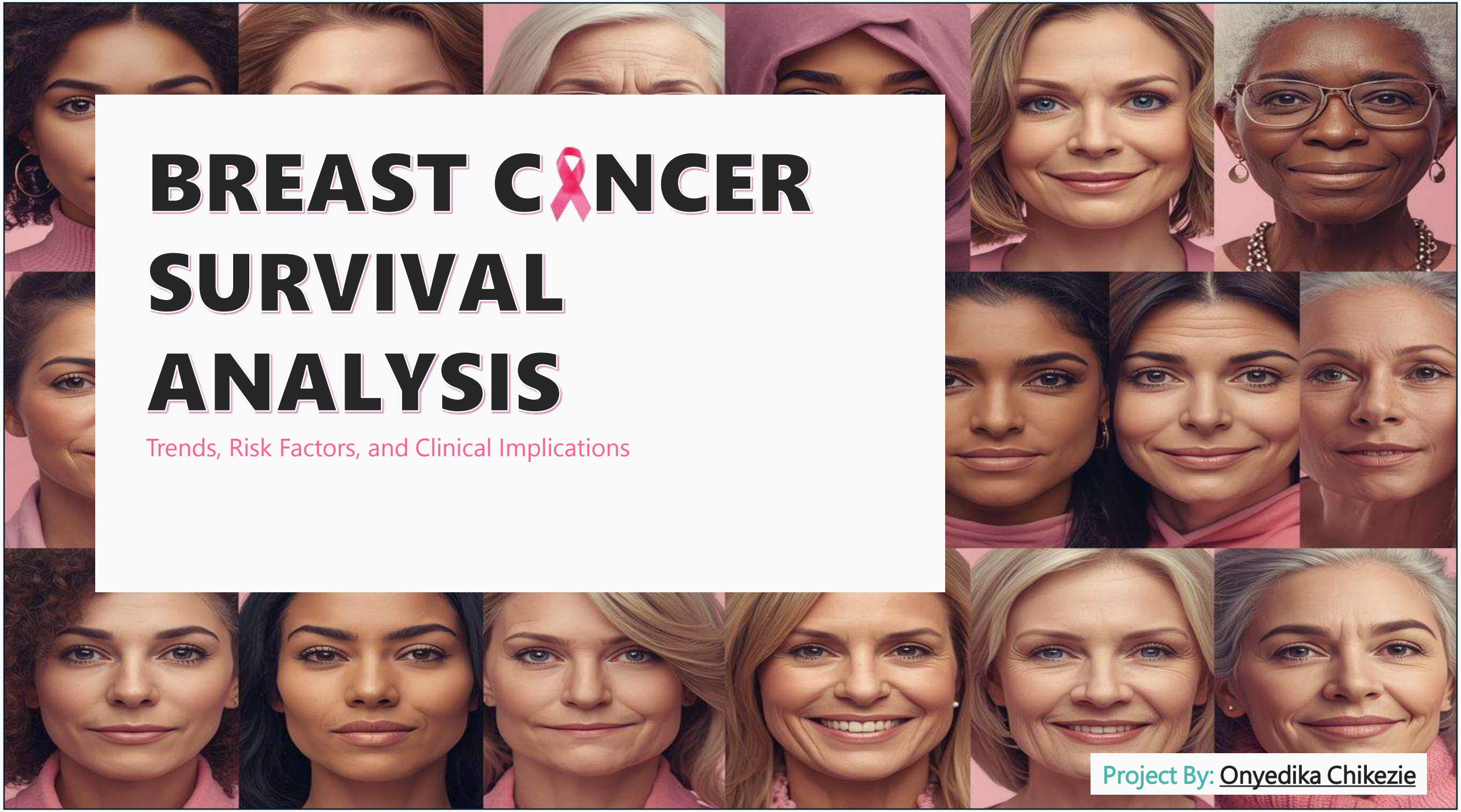


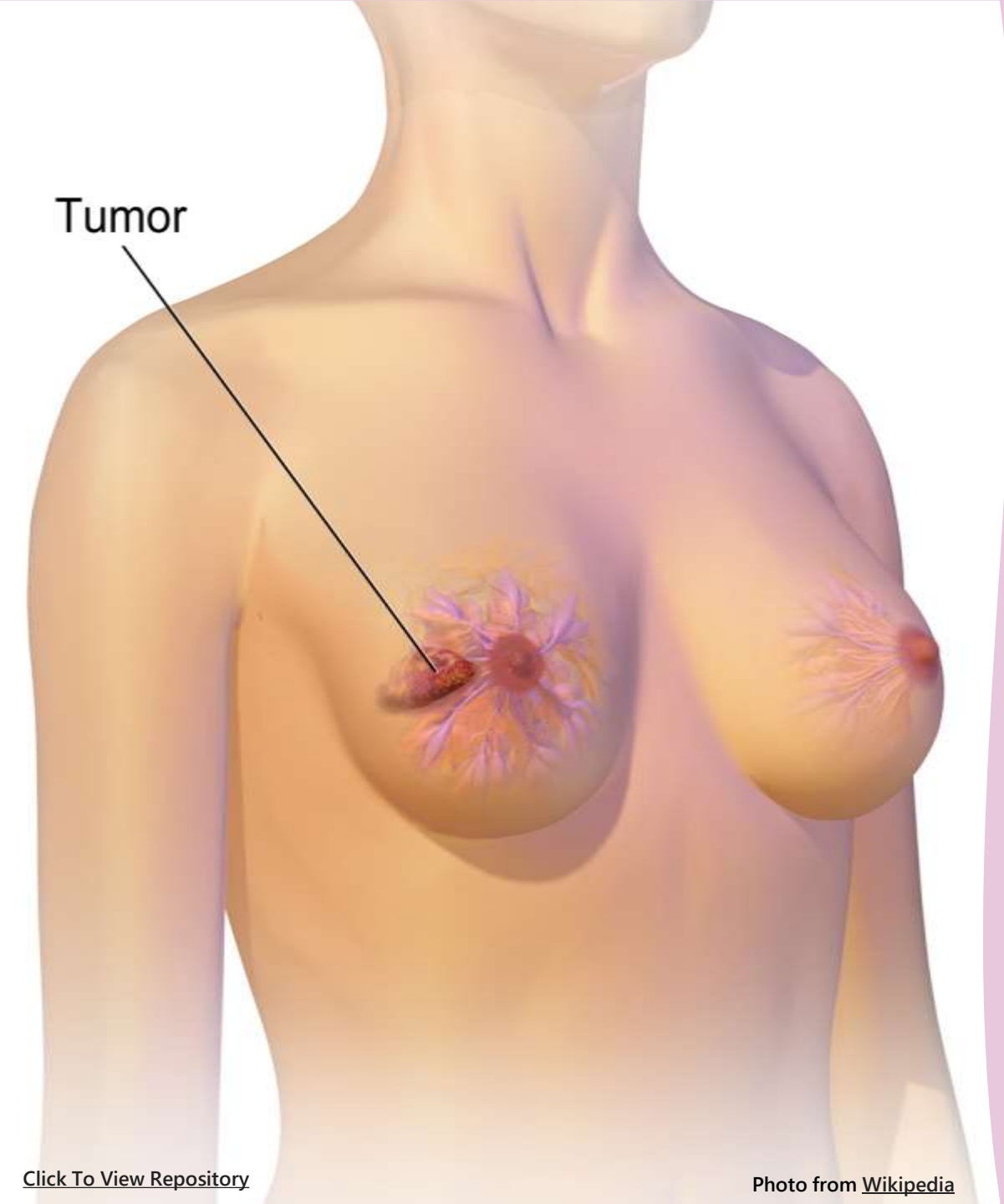
BREAST CANCER SURVIVAL ANALYSIS

Trends, Risk Factors, and Clinical Implications

Project By: Onyedika Chikezie

Contents

- Slide 1 - Introduction
- Slide 2 - Project Background
- Slide 2 - Aim of the Study
- Slide 3 - Project Questions
- Slide 5 - Why Do We Ask These Questions?
- Slide 6 - Data Source and Study Population
- Slide 7 - Analysis Methods
- Slide 8 - Dashboard 1
- Slide 9 - Dashboard 2
- Slide 10 - Insights
- Slide 14 - Discussion of Findings
- Slide 15 - Conclusion
- Slide 16 - Recommendation
- Slide 17 - Limitations of the Analysis
- Slide 18 - References



Click To View Repository

Photo from Wikipedia

Introduction

Breast cancer is the most commonly diagnosed cancer in women worldwide. In 2022 it was the leading cancer among women in 157 of 185 countries and caused an estimated ~670,000 deaths globally [3].

Incidence is rising globally, with Global Cancer Observatory (GLOBOCAN) / American Cancer Society (ACS) reports showing continued growth in case counts over the past decade. [1,2]

Early detection and accurate subtyping are crucial for improving outcomes because treatment options and prognosis depend heavily on stage and receptor status. [9]

Project Background

Breast cancer burden remains high globally; risk varies by stage and biology, and outcomes depend strongly on early detection and receptor status.

The dataset's "T stage," "N stage," and "6th stage" fields reflect the American Joint Committee on Cancer (AJCC) 6th edition tumour, node, metastasis cancer staging system (TNM system) used worldwide for anatomic extent and prognostication.

Estrogen and progesterone receptor (ER/PR) status are established prognostic and treatment-predictive biomarkers. Hormone receptor-positive (HR+) disease typically shows better survival with appropriate endocrine therapy.

For this project, a public breast cancer dataset is analysed, (4,024 rows; demographics, stage, tumour size, grade, ER/PR, survival months, status) curated on Kaggle with variable descriptions aligned to clinical concepts above.

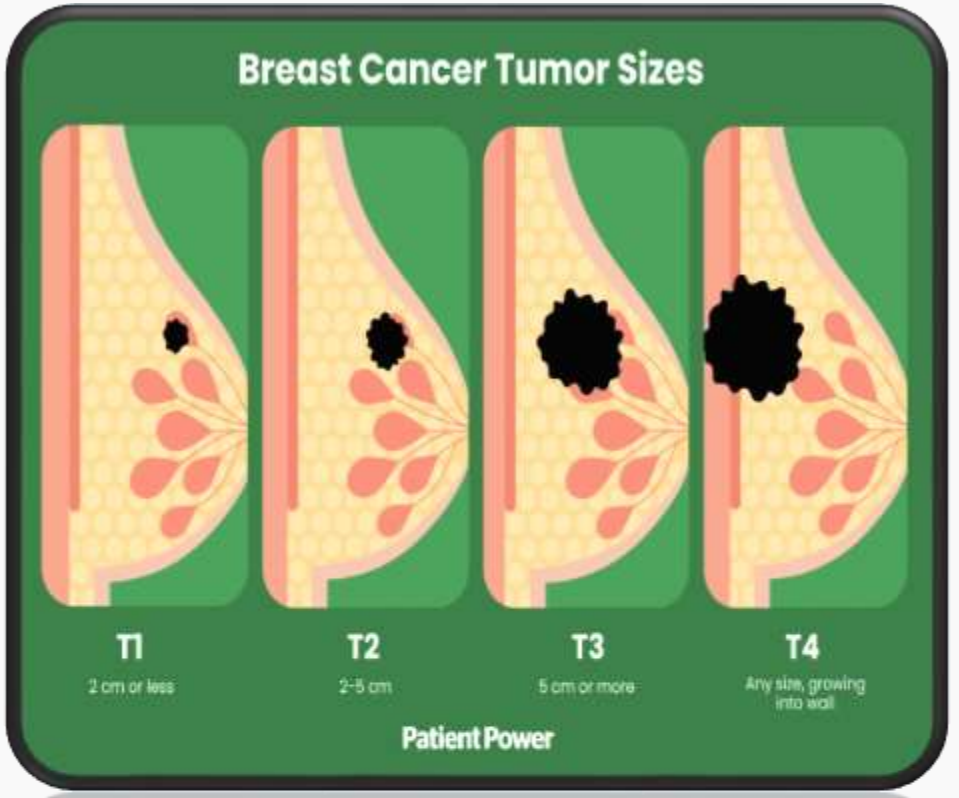


Photo from Patient Power

Aim of the Study

This study aims to identify demographic and clinical factors associated with survival outcomes in breast cancer patients, using real-life-style breast cancer data.

Click To View Repository

Project Questions

Question Type	Insight	Question
Descriptive Baseline	Alive vs Dead count	How many patients are alive compared to those who have died?
	Most common race	Which race occurs most frequently in the dataset?
	Most common T stage	Which T stage classification is most common among patients?
	Average Age	What is the average age of patients?
	Tumour size distribution	What is the average tumour size across all patients?
	Survival months range	What are the shortest and longest survival times recorded?
Trends & patterns by groups	Average age by status	What is the average age of patients in each survival group?
	Average tumour size by status	How does the average tumour size differ between alive and deceased patients?
	Grade distribution by status (most occurring grade by status)	What is the most common tumour grade for each survival status group?
	Count of tumours > 50 mm	How many patients have tumours larger than 50mm?
	Count with ≥1 positive lymph node	How many patients have at least one positive lymph node?

Click To View Repository

Project Questions (cont'd)

Question Type	Insight	Question
Deeper relationships	Race & survival rates	How does the proportion of survivors vary across races?
	Survival by estrogen receptor (ER) status	How does average survival time vary by estrogen receptor status?
	Hormone status (ER/PR) vs differentiation & grade	How do oestrogen and progesterone statuses relate to tumour differentiation and grade?
Possible risk factors	Age group most likely to survive	Which age group has the highest survival rate?
	Stages associated with highest mortality	Which cancer stages have the highest mortality rates?

Click To View Repository

Why Do We Ask These Questions?

Stage and tumour size are some of the strongest predictors of survival. Multiple studies show that increasing tumour size and nodal involvement are associated with lower long-term survival.

AJCC stage correlates closely to 5-year survival (stage IA to IIIA show different survival probabilities), reinforcing the clinical significance of staging in prognosis and treatment planning.

Hormone receptor status (ER/PR) is both prognostic and predictive. ER positive tumours generally have better long-term outcomes and respond to endocrine therapies; modern treatment guidelines base adjuvant therapy decisions on receptor status.

Screening reduces mortality at a population level (randomised trials and systematic reviews support mammography's mortality benefit), so documenting tumour size/ stage at diagnosis informs screening effectiveness.

Age, race, and social determinants influence outcomes. Existing literature documents age-related differences in tumour biology and survival, as well as persistent racial distinctions driven by structural, access, and biological factors. These motivate subgroup analyses by age and race.

Click To View Repository

Data Source and Study Population

Data Source: Kaggle

Dataset Variables

- Sample size: The dataset has 4,024 patient records.
- Demographic variables:
  - Age (30–69 years)
  - Race (White, Black, Other)
  - Marital status.
- Clinical staging:
  - T stage: Primary tumour size classification (T1–T3).
  - o N stage: Lymph node involvement (N0–N3).
  - o 6th AJCC stage: Composite anatomic staging (e.g., IIA, IIIA).
- Pathological features:
  - Tumour grade (1–4)
  - Differentiation (well differentiated→ poorly differentiated).
- Tumour burden:
  - Tumour size (1–140 mm)
  - Number of lymph nodes examined.
  - Number of lymph nodes positive.
- Biomarkers:
  - Estrogen receptor (ER) status (positive/negative)
  - Progesterone receptor (PR) status (positive/negative).
- Outcomes:
  - Survival time in months
  - Survival status (alive or dead).

Click To View Repository

Analysis Methods

- **Data Download & Storage:**

  The CSV was downloaded from Kaggle and imported into a table named `breast_cancer_raw` in MySQL.

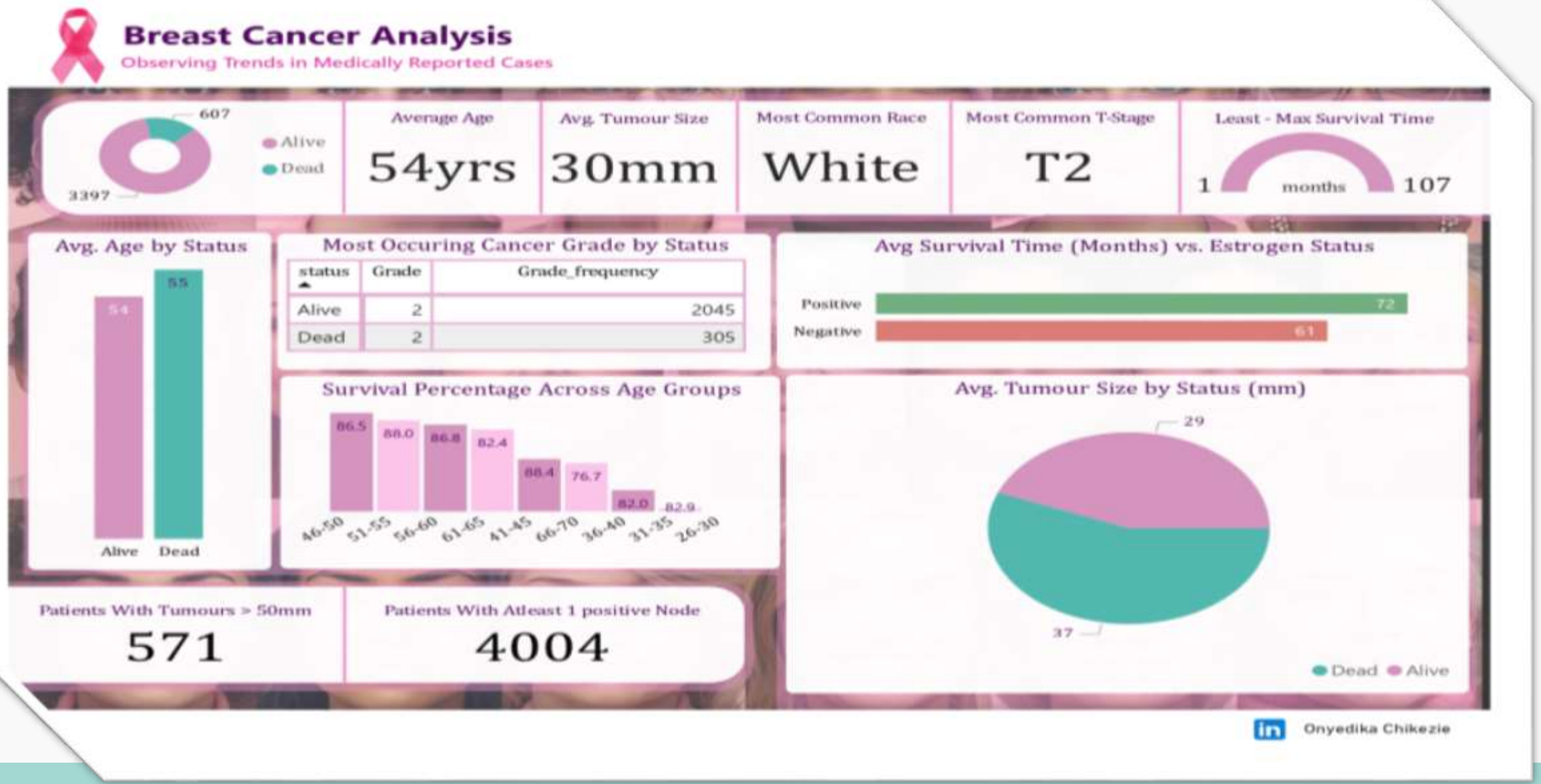
- **Data Cleaning:**
  - Checked data for missing values
  - Standardised column headers
  - Removed duplicate data entries
  - Saved cleaned data table as `breast_cancer_clean`

- **Analysis Approach:**
  - Descriptive aggregation,
  - Group comparisons (AVG, COUNT, proportions),
  - Cross-tabulations were performed using SQL queries.

- **Visualisation:**

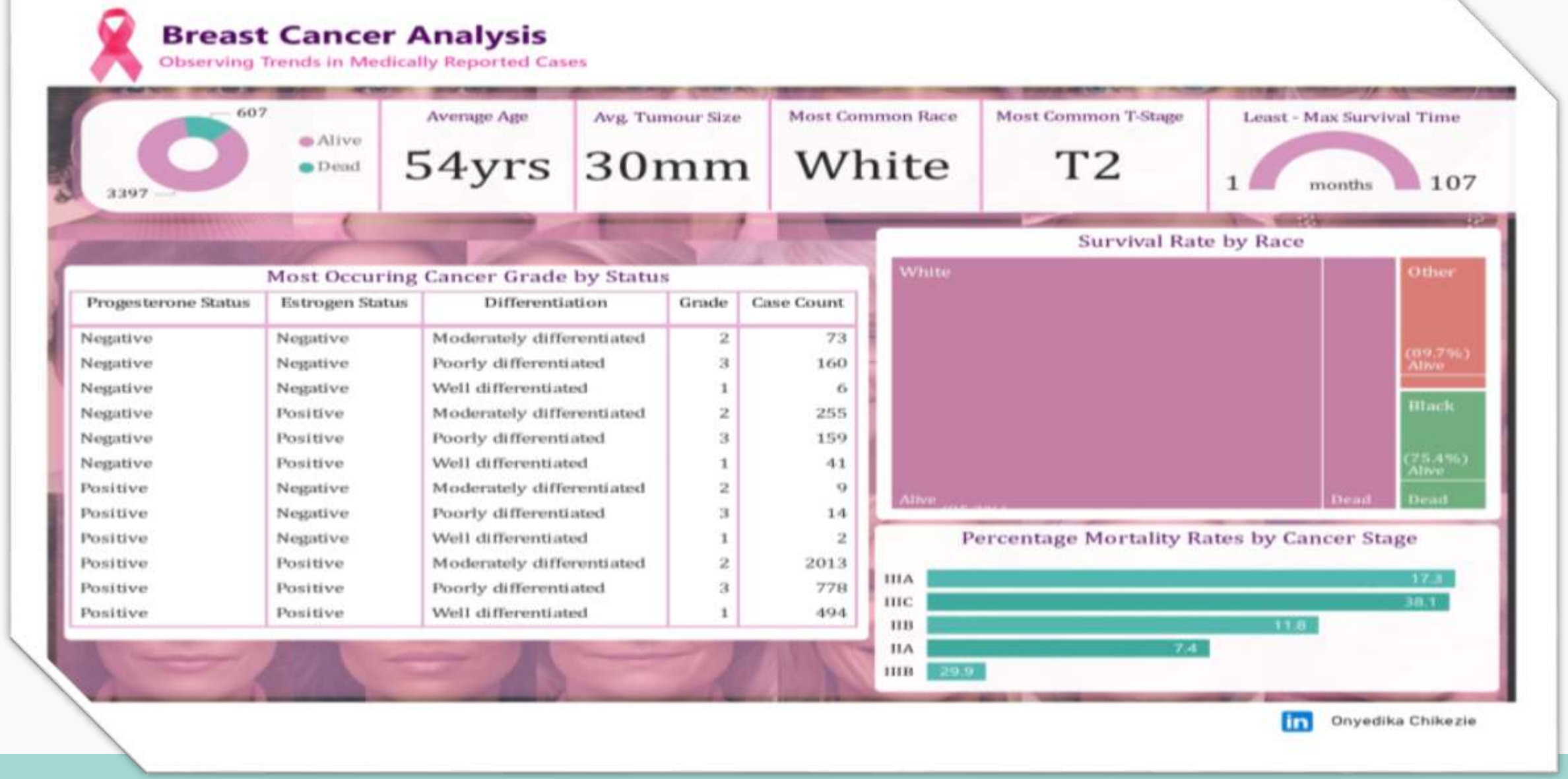
  Dashboard visuals were built in Power BI (static export). The `breast_cancer` database from MySQL was connected to Power BI using the MariaDB ODBC connector. This allowed pulling in queries from the SQL sheet for direct visualisation.

Click To View Repository



Dashboard 1

Click To View Repository



Dashboard 2

Click To View Repository

Insights

Descriptive Results

- **Alive vs Dead count**: 3397 patients are alive, 607 patients are dead.
- **Most common race**: White
- **Most common T stage**: T2
- **Average Age**: 54 years
- **Average Tumour Size**: 30mm
- **Survival Duration**: 1-107 months.

Clinical Importance

The average patient age in the dataset is 54 years. This is younger than the SEER average age '~62 years at diagnosis' [12]. This suggests the sample tilts toward slightly younger patients, though it is still within the known age range for breast cancer occurrence.

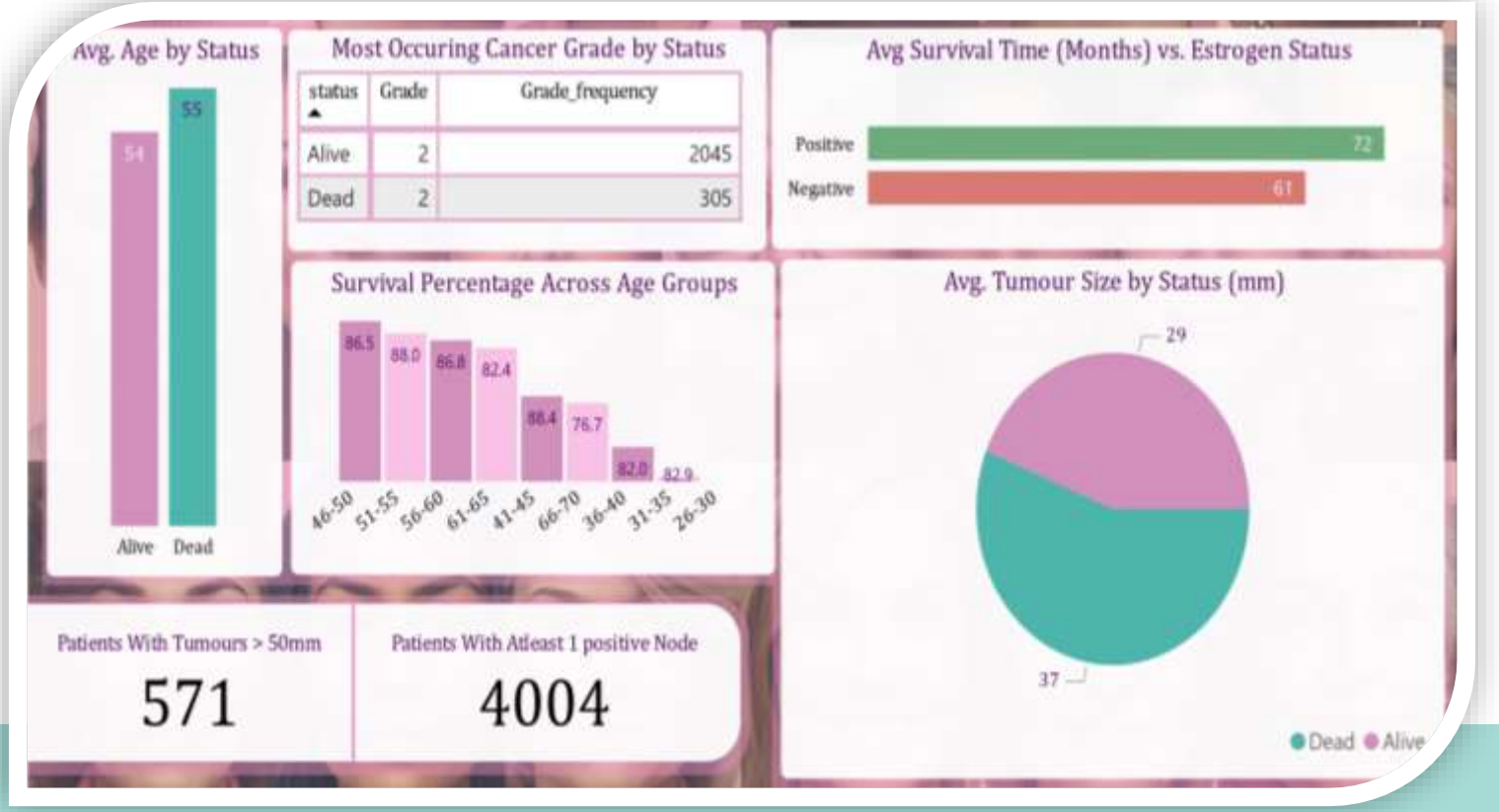
The average tumour size (30 mm) and predominance of living patients are consistent with reports that earlier detection contributes to smaller tumours and improved outcomes [13].

Click To View Repository

Insights

Trends & Patterns by Groups

- **Average age by status:** 54 for living patients, 55 for dead patients.
- **Average tumour size by status:** 29 mm for living patients,37 mm for dead patients.
- **Most occurring grade by status**: Grade 2 for both living and dead patients with a frequency of 2045 and 305 respectively.
- **Count of tumours > 50 mm:** 571 patients
- **Count with ≥1 positive lymph node:** 4,004 patients



Clinical Importance

Studies show that larger tumours (>50 mm) and higher grades are associated with poorer survival rates, pointing out the importance of early detection and accurate grading. Although nodal status could not be assessed properly from this dataset, tumour size and grade provide powerful measures for prognosis. Clinically, these markers directly influence treatment decisions, such as the use of chemotherapy or more aggressive treatment.

Click To View Repository

Insights

Most Occuring Cancer Grade by Status

Progesterone Status	Estrogen Status	Differentiation	Grade	Case Count
Negative	Negative	Moderately differentiated	2	73
Negative	Negative	Poorly differentiated	3	160
Negative	Negative	Well differentiated	1	6
Negative	Positive	Moderately differentiated	2	255
Negative	Positive	Poorly differentiated	3	159
Negative	Positive	Well differentiated	1	41
Positive	Negative	Moderately differentiated	2	9
Positive	Negative	Poorly differentiated	3	14
Positive	Negative	Well differentiated	1	2
Positive	Positive	Moderately differentiated	2	2013
Positive	Positive	Poorly differentiated	3	778
Positive	Positive	Well differentiated	1	494

Clinical Importance

These findings emphasise the importance of hormone receptor testing, as ER/PR positivity directly informs treatment decisions.

They also highlight the need to address racial disparities, which may stem from differences in tumour biology, access to screening, and treatment availability. Clinically, these insights stress both personalised treatment and broader equity in breast cancer care.

Deeper Relationships

- **Race & Survival Rates:** Survival rates are highest in patients of other races (89.7%), lower in White patients (85.2%) and lowest in Black patients (75.4%).
- **Survival by Estrogen Receptor (ER) Status:** Patients with ER-positive tumours survive an average of 72 months, compared to ER-negative tumour patients with an average of 61 months.
- **Hormone Status (ER/PR) vs Differentiation & Grade:** The largest subgroup is the group of patients with Estrogen Receptor + and Progesterone Receptor + tumours, especially the moderately differentiated tumours (2013 patients).

Click To View Repository

Insights

Clinical Importance

This dataset shows that survival is highest in women aged 41–45 (88.4%) but declines steadily in older groups, with the lowest survival in ages 66–70 (76.7%). This supports evidence that older age is associated with poorer prognosis due to biological aggressiveness, comorbidities, and treatment limitations [4]. We also observe mortality rising sharply with advancing AJCC stage (from 7.4% at stage IIA to 38.1% at stage IIIC), confirming stage at diagnosis as the strongest determinant of survival[11]. These results emphasise the importance of early detection programmes and age-appropriate treatment strategies to improve outcomes.

Possible Risk Factors

Age group most likely to survive: The highest survival rate is observed in the 41-45 bin (88.4%) and the 26-30 bin has the least people and percentage (60%). Overall, the age group survival range is between 60%-88% and data shows that mid-life groups fare better than older groups.

Stages associated with the highest mortality: The bar chart shows that mortality increases with stage.

Click To View Repository

Discussion of Findings

The findings from this study align with existing literature on breast cancer survival determinants. The observed higher survival rates in younger women (41–45 years, 88.4%) and declining survival in older groups support existing reports linking age with prognosis, comorbidities, and treatment tolerance [6,10].

The association between increasing tumour size, advanced AJCC stage, and mortality confirms established evidence that staging remains the strongest prognostic factor [7,8]. Similarly, ER/PR positivity is reaffirmed as predictive of improved survival with endocrine therapy [5].

The dataset's lower average age (54 years) compared to SEER data (~62 years) [10] suggests sample variation, yet survival patterns remain consistent with international trends. Racial differences observed also mirror known inequities in access to screening and treatment [3].

Click To View Repository

Conclusion

This report highlights the continued need for integrating clinical, demographic, and biological factors into breast cancer management. Beyond confirming already known survival predictors, these findings emphasise the urgency of translating data insights into practical health policies, ensuring that evidence informs screening strategies, equal access to therapies, and patient-centred care. Ultimately, improved outcomes will depend not only on clinical advances but also on addressing systemic barriers to timely diagnosis and effective treatment.

Click To View Repository

Recommendation

- Dataset used for this study was secondary, not collected in a controlled clinical setting.
- There are limited clinical details, i.e missing data on treatment regimens, comorbidities, Human Epidermal Growth Factor Receptor 2 (HER2) status, and lifestyle factors.
- Racial categories and social determinants are too simplified, limiting thorough analysis.
- Possible selection bias: average age (54 yrs) younger than SEER cohort (~62 yrs).

Click To View Repository

Limitations of the Analysis

The findings from this work align with existing literature on breast cancer survival determinants. The observed higher survival rates in younger women (41–45 years, 88.4%) and declining survival in older groups support existing reports linking age with prognosis, comorbidities, and treatment tolerance [6,10].

Nodal status could not be assessed properly in this dataset. The dashboard shows “Patients with ≥1 positive lymph node = 4,004”, which equals the total sample size. This may indicate an aggregation issue (every patient flagged as node-positive) and requires validation/cleaning before interpreting node positivity prevalence.

The association between increasing tumour size, advanced AJCC stage, and mortality confirms established evidence that staging remains the strongest prognostic factor [7,8]. Similarly, ER/PR positivity is reaffirmed as predictive of improved survival with endocrine therapy [5].

The dataset’s lower average age (54 years) compared to SEER data (~62 years) [10] suggests sample variation, yet survival patterns remain consistent with international trends. Racial differences observed also mirror known inequities in access to screening and treatment [3].

Click To View Repository

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2024 May;74(3):229-63.
2. Breast Cancer Research Foundation. Breast cancer statistics and resources [Internet]. New York: BCRF; [cited 2025 Sep 9]. Available from: https://www.bcrf.org/breast-cancer-statistics-and-resources/
3. Breast cancer: fact sheet. Geneva: WHO; 2025 [cited 2025 Sep 5]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
4. Chen RC, Royce TJ, Extermann M, Reeve BB. Impact of age and comorbidity on treatment and outcomes in elderly cancer patients. InSeminars in radiation oncology 2012 Oct 1 (Vol. 22, No. 4, pp. 265-271). WB Saunders.
5. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, Jemal A, Siegel RL. Breast cancer statistics, 2019. CA: a cancer journal for clinicians. 2019 Nov;69(6):438-51.
6. Dowsett M, Forbes JF, Bradley R, Ingle J, Aihara T, Bliss J, Boccardo F, Coates A, Coombes RC, Cuzick J, Dubsky P. Early Breast Cancer Trialists' Collaborative G. Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials. Lancet. 2015;386(10001):1341-52.
7. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Annals of surgical oncology. 2010 Jun;17(6):1471-4. Carter CL, Allen C, Henson DE. Relation of tumour size, lymph node status, and survival in 24,740 breast cancer cases. Cancer. 1989;63(1):181–7.

Click To View Repository

References

8. Hammond ME, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, Fitzgibbons PL, Francis G, Goldstein NS, Hayes M, Hicks DG. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. Journal of clinical oncology. 2010 Jun 1;28(16):2784-95.

9. Henderson JT, Webber EM, Weyrich MS, Miller M, Melnikow J. Screening for breast cancer: evidence report and systematic review for the US Preventive Services Task Force. Jama. 2024 Jun 11;331(22):1931-46.

10. Howlader N, Ward KC, Warren JL, Campbell DS, Coyle L, Mariotto AB. Assessment of oncology practice billing claims for supplementing chemotherapy: a pilot study in the Georgia SEER Cancer Registry. JNCI Monographs. 2020 May 1;2020(55):82-8.

11. Jedy-Agba E, McCormack V, Adebamowo C, dos-Santos-Silva I. Stage at diagnosis of breast cancer in sub-Saharan Africa: a systematic review and meta-analysis. The Lancet Global Health. 2016 Dec 1;4(12):e923-35.

12. National Cancer Institute. SEER Cancer Stat Facts: Female Breast Cancer [Internet]. Bethesda (MD): NCI; [cited 2025 Sep 11]. Available from: https://seer.cancer.gov/statfacts/html/breast.html

13. Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MM. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. Bmj. 2015 Oct 6;351.

Click To View Repository