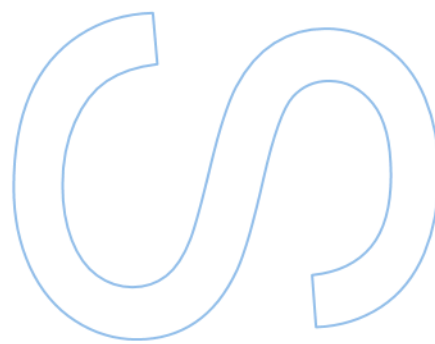
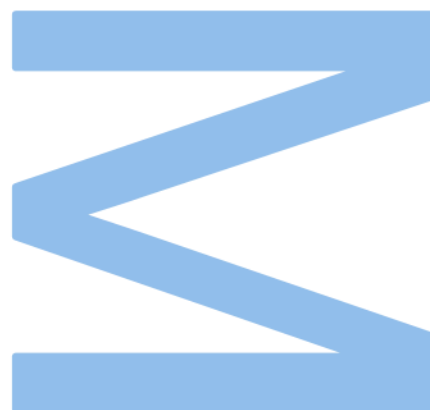


# **Determinants of Breast Cancer Survival: A Registry-Based Analysis with Classical and Visual Analytical Approaches**



**Leonor Vieira Sottomayor**

Master in Bioinformatics and Computational Biology  
Department of Biology and Department of Computer Science  
Faculty of Sciences of the University of Porto  
2025





# **Determinants of Breast Cancer Survival: A Registry-Based Analysis with Classical and Visual Analytical Approaches**

**Leonor Vieira Sottomayor**

Dissertation carried out as part of the Master in  
Bioinformatics and Computational Biology  
Department of Biology and Department of Computer Science  
2025

**Supervisor**

PhD Professor Álvaro Pedro de Barros Borges Reis Figueira,  
Department of Computer Science, Faculty of Sciences of the  
University of Porto

**Co-Supervisor**

PhD Professor Luís Filipe Costa de Castro, Department of Biology,  
Faculty of Sciences of the University of Porto

# Acknowledgements

I would like to begin by expressing my deepest gratitude to my supervisor, Professor Álvaro Figueira, for his dedication, patience, and guidance from the very start of this project. I am especially grateful for the enthusiasm with which you embraced this work and for the invaluable advice and encouragement you have offered throughout the year. I also extend my sincere thanks to Professor Luís Castro for his availability and willingness to provide essential guidance for the development of this thesis. To both of you, thank you for welcoming my ideas and helping me bring them to life.

To my parents, sisters, and extended family, thank you for your unconditional support from day one. Whatever the future holds for me, I owe it to your love and encouragement.

Inês, while I cannot say this would have been impossible without you, it would certainly have been much harder. Thank you for always being there, for keeping me grounded, and for encouraging me always to take the leap of faith.

Finally, to all my friends, thank you for making life richer and for being the ones I turn to when I need to take my mind off work. This accomplishment is also thanks to you.

# Resumo

O cancro da mama continua a neoplasia maligna mais comum entre as mulheres em todo o mundo e compreender os seus determinantes prognósticos é crucial para melhorar os resultados. Esta dissertação analisou dados do programa Surveillance, Epidemiology, and End Results (SEER) entre 2000 e 2021 para avaliar os determinantes da sobrevivência e desenvolver uma nova aplicação interativa. Estatísticas descritivas, estimativa de Kaplan-Meier e regressão multivariada de Cox foram empregadas para quantificar o impacto das variáveis demográficas, clínicas e socioeconómicas na sobrevivência em cinco anos. A análise destacou vários fatores prognósticos importantes, incluindo o estágio no momento do diagnóstico, o tamanho do tumor, a idade, o subtipo molecular e as disparidades socioeconómicas, que influenciaram consistentemente os resultados de sobrevivência, com a doença distante conferindo um aumento de mais de dez vezes no risco de morte em comparação com a doença localizada. Para complementar essas estatísticas descritivas, foram desenvolvidos modelos supervisionados de aprendizagem de máquina para melhorar a previsão ao nível individual. Entre os algoritmos testados, o XGBoost alcançou o melhor desempenho (ROC AUC = 0,870 e PR AUC = 0,978), superando os classificadores de regressão logística e floresta aleatória. Os padrões de importância das variáveis foram consistentes com os resultados da regressão de Cox, aumentando a confiança na sua relevância clínica.

Uma das contribuições deste trabalho é o desenvolvimento do Breast Cancer Data Explorer, uma aplicação Shiny que integra análise descritiva, modelação de sobrevivência e previsões de aprendizagem de máquina numa única plataforma fácil de usar. Ao permitir visualização interativa, estimativas prognósticas personalizadas e resultados exportáveis, a aplicação preenche uma lacuna na literatura: a falta de ferramentas acessíveis que tornem os dados de sobrevivência ao cancro baseados em registos utilizáveis por médicos, investigadores e decisores políticos sem conhecimentos avançados de estatística ou programação.

Os resultados mostram que combinar a análise de sobrevivência tradicional com a modelagem preditiva moderna numa plataforma interativa pode melhorar tanto a compreensão científica quanto a acessibilidade prática. O Breast Cancer Data Explorer serve como uma prova de conceito, mostrando que é possível equilibrar o rigor metodológico com a usabilidade, com potencial para evoluir para um valioso recurso de acesso aberto que apoia a investigação, a educação e as políticas relacionadas ao cancro.

**Palavras-chave:** Cancro da mama; análise de sobrevivência; regressão de Cox; Kaplan-Meier; aprendizagem de máquina; XGBoost; SEER; modelagem prognóstica; aplicação Shiny; explorador interativo de dados

# Abstract

Breast cancer remains the most commonly diagnosed malignancy among women worldwide, and understanding its prognostic determinants is crucial for improving outcomes. This thesis analyzed data from the Surveillance, Epidemiology, and End Results (SEER) program between 2000 and 2021 to identify survival determinants and develop a new interactive application. Descriptive statistics, Kaplan-Meier estimation, and multivariable Cox regression were employed to quantify the impact of demographic, clinical, and socioeconomic variables on five-year survival. The analysis highlighted several key prognostic factors, including stage at diagnosis, tumor size, age, molecular subtype, and socioeconomic disparities, which consistently influenced survival outcomes, with distant disease conferring more than a ten-fold increase in hazard of death compared to localized disease. To complement these descriptive statistics, supervised machine-learning models were developed to improve individual-level prediction. Among the tested algorithms, XGBoost achieved the highest performance (ROC AUC = 0.870 and PR AUC = 0.978), outperforming logistic regression and random forest classifiers. Variable importance patterns were consistent with Cox regression findings, enhancing confidence in their clinical relevance.

One of the contributions of this work is the development of the Breast Cancer Data Explorer, a Shiny application that integrates descriptive analysis, survival modeling, and machine-learning predictions into a single, user-friendly platform. By enabling interactive visualization, personalized prognostic estimates, and exportable outputs, the application addresses a gap in the literature: the lack of accessible tools that make registry-based cancer survival data usable by clinicians, researchers, and policymakers without advanced statistical or programming expertise.

The results show that combining traditional survival analysis with modern predictive modeling in an interactive platform can improve both scientific understanding and practical accessibility. The Breast Cancer Data Explorer serves as a proof of concept, showing that it is possible to balance methodological rigor with usability, with potential to evolve into a valuable open-access resource supporting cancer research, education, and policy.

**Keywords:** Breast cancer; survival analysis; Cox regression; Kaplan–Meier; machine learning; XGBoost; SEER; prognostic modeling; Shiny application; interactive data

# Table of Contents

|                                                                |      |
|----------------------------------------------------------------|------|
| List of Tables .....                                           | viii |
| List of Figures .....                                          | ix   |
| 1. Introduction .....                                          | 1    |
| 1.1. Breast Cancer as a Global Health Challenge .....          | 1    |
| 1.2. Importance of survival analysis in breast cancer .....    | 2    |
| 1.3. The SEER dataset as a research resource .....             | 3    |
| 1.4. Predictive modeling in oncology .....                     | 3    |
| 1.5. Statistics vs. Visual Analytics .....                     | 4    |
| 1.6. Research gap and motivation .....                         | 5    |
| 1.7. Objectives of the study .....                             | 6    |
| 1.8. Thesis Organization .....                                 | 6    |
| 2. State Of The Art .....                                      | 8    |
| 2.1. Prognostic factors in breast cancer survival .....        | 8    |
| 2.1.1. Tumor-Intrinsic Factors .....                           | 8    |
| 2.1.2. Host Factors .....                                      | 9    |
| 2.1.3. Treatment-Related Factors .....                         | 10   |
| 2.1.4. Structural and Sociodemographic Determinants .....      | 11   |
| 2.2. Survival analysis methodologies in oncology .....         | 12   |
| 2.3. Machine learning approaches for survival prediction ..... | 14   |
| 2.4. Visualization Systems' Strengths and Weaknesses .....     | 18   |
| 2.4.1. The Need for Interactive and Accessible Tools .....     | 20   |
| 2.5. A Visual Analytics Approach Using R and Shiny .....       | 22   |
| 2.6. Sum up .....                                              | 22   |
| 3. Materials and Methods .....                                 | 24   |
| 3.1. SEER Dataset .....                                        | 24   |
| 3.1.1. Variables .....                                         | 24   |
| 3.2. Data Preprocessing .....                                  | 27   |
| 3.2.1. Variable Missingness .....                              | 27   |
| 3.2.2. Data Cleaning .....                                     | 29   |
| 3.2.3. Outcome Variables .....                                 | 32   |
| 3.2.4. Data Imputation .....                                   | 33   |
| 3.2.5. Derived Variables .....                                 | 33   |
| 3.3. Descriptive Analysis .....                                | 36   |

|                                                                    |    |
|--------------------------------------------------------------------|----|
| 3.4. Traditional survival/statistical analysis.....                | 37 |
| 3.4.1. Contingency Tables .....                                    | 37 |
| 3.4.2. Chi-square Test of Independence .....                       | 37 |
| 3.4.3. Kaplan-Meier Survival Analysis .....                        | 37 |
| 3.4.4. Log-rank Test .....                                         | 38 |
| 3.4.5. Cox Proportional Hazards Models .....                       | 38 |
| 3.5. Machine Learning Models .....                                 | 40 |
| 3.5.1. Data Preparation .....                                      | 40 |
| 3.5.2. Train/Test Split .....                                      | 40 |
| 3.5.3. Model Selection.....                                        | 40 |
| 3.5.4. Handling Class Imbalance .....                              | 41 |
| 3.5.5. Evaluation Metrics.....                                     | 41 |
| 3.6. Shiny App.....                                                | 42 |
| 3.7. Implementation Details .....                                  | 42 |
| 4. Results.....                                                    | 43 |
| 4.1. Univariate Analysis .....                                     | 43 |
| 4.1.1. Age .....                                                   | 43 |
| 4.1.2. Breast Subtype.....                                         | 44 |
| 4.1.3. Chemotherapy.....                                           | 45 |
| 4.1.4. COD to Site .....                                           | 46 |
| 4.1.5. Combined Summary Stage.....                                 | 47 |
| 4.1.6. Derived HER2 .....                                          | 49 |
| 4.1.7. ER Status .....                                             | 50 |
| 4.1.8. First Malignant Primary Indicator .....                     | 51 |
| 4.1.9. Marital Status at Diagnosis .....                           | 52 |
| 4.1.10. Median Household Income (Inflation-Adjusted to 2022) ..... | 53 |
| 4.1.11. PR Status .....                                            | 54 |
| 4.1.12. Primary Site.....                                          | 55 |
| 4.1.13. Race and Origin .....                                      | 56 |
| 4.1.14. Radiation .....                                            | 58 |
| 4.1.15. Rural-Urban Continuum Code .....                           | 59 |
| 4.1.16. SEER Cause-Specific Death Classification .....             | 61 |
| 4.1.17. Sex .....                                                  | 62 |
| 4.1.18. Survival Months.....                                       | 63 |
| 4.1.19. Time from Diagnosis to Treatment .....                     | 64 |
| 4.1.20. Total Number of Benign/Borderline Tumors .....             | 66 |
| 4.1.21. Total Number of In Situ/Malignant Tumors.....              | 67 |
| 4.1.22. Tumor Size Over Time.....                                  | 68 |

|                                                                     |     |
|---------------------------------------------------------------------|-----|
| 4.1.23. Tumor Size Summary .....                                    | 70  |
| 4.1.24. Year of Diagnosis.....                                      | 71  |
| 4.2. Contingency Tables and Chi-square Tests .....                  | 73  |
| 4.3. Survival Analysis: Kaplan-Meier Curves and Log-rank Test ..... | 75  |
| 4.4. Cox Proportional Hazards Model Results .....                   | 77  |
| 4.4.1. 2000-2021 Diagnoses.....                                     | 77  |
| 4.4.2. 2010-2021 Diagnoses.....                                     | 79  |
| 4.4.3. Proportional Hazards Assumptions .....                       | 81  |
| 4.5. Machine Learning Models .....                                  | 82  |
| 4.6. Breast Cancer Data Explorer.....                               | 86  |
| 4.7. Main Results Summary.....                                      | 95  |
| 5. Conclusion.....                                                  | 97  |
| Bibliography .....                                                  | 100 |

# List of Tables

|                                                                                                               |     |
|---------------------------------------------------------------------------------------------------------------|-----|
| 3.1. Variables in the dataset with descriptions and corresponding data types.....                             | 24  |
| 3.2. Missingness Distribution. ....                                                                           | 29  |
| 3.3. Summary of variable transformations during data cleaning. ....                                           | 30  |
| 1. Five-Year Survival Counts and Percentages by Age Group .....                                               | 117 |
| 2. Five-Year Survival Counts and Percentages by Breast Cancer Subtype .....                                   | 118 |
| 3. Five-Year Survival Counts and Percentages by Chemotherapy Status.....                                      | 118 |
| 4. Five-Year Survival Counts and Percentages by Combined Summary Stage .....                                  | 118 |
| 5. Five-Year Survival Counts and Percentages by Derived HER2 Status .....                                     | 119 |
| 6. Five-Year Survival Counts and Percentages by ER Status .....                                               | 119 |
| 7. Five-Year Survival Counts and Percentages by First Malignant Primary Indicator .....                       | 119 |
| 8. Five-Year Survival Counts and Percentages by Marital Status at Diagnosis.....                              | 120 |
| 9. Five-Year Survival Counts and Percentages by Median Household Income (Inflation<br>Adjusted to 2022) ..... | 121 |
| 10. Five-Year Survival Counts and Percentages by PR Status .....                                              | 121 |
| 11. Five-Year Survival Counts and Percentages by Primary Site .....                                           | 122 |
| 12. Five-Year Survival Counts and Percentages by Race/Ethnicity.....                                          | 122 |
| 13. Five-Year Survival Counts and Percentages by Radiation Therapy Type .....                                 | 123 |
| 14. Five-Year Survival Counts and Percentages by Rural-Urban Continuum Code.....                              | 123 |
| 15. Five-Year Survival Counts and Percentages by Sex.....                                                     | 123 |
| 16. Five-Year Survival Counts and Percentages by Time From Diagnosis to Treatment..                           | 124 |
| 17. Five-Year Survival Counts and Percentages by Total Number of Benign/Borderline<br>Tumors .....            | 124 |
| 18. Five-Year Survival Counts and Percentages by Total Number of In Situ Malignant<br>Tumors .....            | 124 |
| 19. Five-Year Survival Counts and Percentages by Tumor Size .....                                             | 125 |
| 20. Five-Year Survival Counts and Percentages by Year of Diagnosis .....                                      | 125 |

# List of Figures

|                                                                                                                                                                                                                                                                                                                                                                                                       |    |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1. Percentage of Missing Data by Variable.....                                                                                                                                                                                                                                                                                                                                                      | 28 |
| 4.1. Distribution of Age at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). ....                                                                                                                                                                                                                                                                                                                  | 43 |
| 4.2. Distribution of Age at Diagnosis as a categorical variable with 5-year bins in Breast Cancer Cases (SEER, 2000-2021).....                                                                                                                                                                                                                                                                        | 44 |
| 4.3. Distribution of Breast Cancer Subtypes (SEER, 2000-2021). The bar plot displays the frequencies of breast cancer cases across molecular subtypes defined by hormone receptor (estrogen and progesterone) status and HER2 (human epidermal growth factor receptor 2) expression. ....                                                                                                             | 45 |
| 4.4. Distribution of Chemotherapy Status in Breast Cancer Cases (SEER, 2000-2021). The pie chart illustrates the proportion of breast cancer patients who received chemotherapy versus those who did not or for whom the status is unknown, as recorded in the SEER database. ....                                                                                                                    | 46 |
| 4.5. Wordcloud showing the distribution of site-specific cause of death in Breast Cancer Cases (SEER, 2000-2021).....                                                                                                                                                                                                                                                                                 | 47 |
| 4.6. Distribution of Stage at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the number of breast cancer cases by stage at diagnosis, as recorded in the SEER database. ....                                                                                                                                                                                            | 49 |
| 4.7. Distribution of Derived HER2 at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the status of HER2 at diagnosis, as recorded in the SEER database. ....                                                                                                                                                                                                             | 50 |
| 4.8. Distribution of ER Status at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the status of ER at diagnosis, as recorded in the SEER database. ....                                                                                                                                                                                                                  | 51 |
| 4.9. Distribution of First Malignant Primary Tumor Status in Breast Cancer Cases (SEER, 2000-2021). The pie chart illustrates whether the recorded breast tumor was the patient's first malignant primary tumor, as documented in the SEER database. This indicator helps estimate cancer incidence by identifying initial malignancies, even if prior cancers are not recorded in the registry. .... | 52 |
| 4.10 Distribution of Marital Status at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the number of breast cancer cases by marital status at the time of diagnosis, as recorded in the SEER database. ....                                                                                                                                                              | 53 |
| 4.11 Distribution of Median Household Income Among Breast Cancer Patients (SEER, 2000-2021). The bar plot illustrates the distribution of breast cancer cases by median household income, with values inflation-adjusted to 2022 dollars.....                                                                                                                                                         | 54 |
| 4.12 Distribution of PR Status at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the status of PR at diagnosis, as recorded in the SEER database. ....                                                                                                                                                                                                                  | 55 |
| 4.13 Distribution of primary tumor sites in breast cancer cases (SEER, 2000-2021). The bar plot illustrates the frequency of tumor locations coded according to SEER's anatomical site classification. ....                                                                                                                                                                                           | 56 |
| 4.14 Distribution of Race and Origin in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the frequency of breast cancer cases across different racial and ethnic groups as recorded in the SEER database. ....                                                                                                                                                                         | 57 |
| 4.15 Distribution of Radiation Types in Breast Cancer Cases (SEER, 2000-2021). The bar plot displays the frequencies of different radiation therapies received by patients, as recorded in the SEER database. ....                                                                                                                                                                                    | 59 |

|      |                                                                                                                                                                                                                                                                                                                                                           |    |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4.16 | Distribution of Rural/Urban Continuum Codes in Breast Cancer Cases (SEER, 2000-2021). The bar plot displays the frequencies of breast cancer cases across different geographic classifications based on the Rural/Urban Continuum Codes (RUCC), which reflect the degree of urbanization and proximity to metropolitan areas of patients' residences..... | 60 |
| 4.17 | Distribution of Cause-Specific Death Classification in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the number of patients by cause-specific death classification, as recorded in the SEER database. This classification distinguishes between deaths due to breast cancer and those alive, or dead due to other causes..              | 62 |
| 4.18 | Distribution of Patients' Sex in Breast Cancer Cases (SEER, 2000-2021). The pie chart illustrates the proportion of breast cancer cases by gender, as recorded in the SEER database. ....                                                                                                                                                                 | 63 |
| 4.19 | Distribution of Survival Months in Breast Cancer Cases (SEER, 2000-2021). The density plot illustrates the distribution of the number of months that breast cancer patients survived, as recorded in the SEER database. ....                                                                                                                              | 64 |
| 4.20 | Distribution of Time from Diagnosis to Treatment in Breast Cancer Cases (SEER, 2000-2021). The density plot illustrates the distribution of breast cancer cases based on the number of days between diagnosis and the initiation of treatment, as recorded in the SEER database. ....                                                                     | 65 |
| 4.21 | Distribution of Time from Diagnosis to Treatment in Breast Cancer Cases (SEER, 2000-2021) by category. ....                                                                                                                                                                                                                                               | 66 |
| 4.22 | Distribution of Total Number of Benign/Borderline Tumors in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the distribution of breast cancer cases based on the number of benign or borderline tumors reported per patient, as recorded in the SEER database. ....                                                                       | 67 |
| 4.23 | Distribution of Total Number of In Situ/Malignant Tumors in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the distribution of breast cancer cases based on the number of in situ or malignant tumors reported per patient, as recorded in the SEER database. ....                                                                       | 68 |
| 4.24 | Distribution of Tumor Size in Breast Cancer Cases (SEER, 2000-2021). The density plot illustrates the distribution of tumor sizes, measured in millimetres, among breast cancer cases as recorded in the SEER database. ....                                                                                                                              | 69 |
| 4.25 | Distribution of Tumor Size in Breast Cancer Cases (SEER, 2000-2021) by Category. The bar plot illustrates the distribution of tumor sizes, measured in millimetres, among breast cancer cases as recorded in the SEER database. ....                                                                                                                      | 70 |
| 4.26 | Distribution of Tumor Size in Breast Cancer Cases (SEER, 2016-2021). The density plot illustrates the distribution of tumor sizes, measured in millimetres, among breast cancer cases as recorded in the SEER database. ....                                                                                                                              | 71 |
| 4.27 | Distribution of Tumor Size in Breast Cancer Cases (SEER, 2016-2021) by Category. The bar plot illustrates the distribution of tumor sizes, measured in millimetres, among breast cancer cases as recorded in the SEER database. ....                                                                                                                      | 71 |
| 4.28 | Distribution of Year of Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the distribution of when the tumor was diagnosed, based on year, among breast cancer cases as recorded in the SEER database. ....                                                                                                                    | 73 |
| 4.29 | Distribution of Year of Diagnosis in Breast Cancer Cases (SEER, 2000-2021) by Category. The bar plot illustrates the distribution of when the tumor was diagnosed, based on year, among breast cancer cases as recorded in the SEER database.....                                                                                                         | 73 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |    |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4.30 | <b>Home page of the Breast Cancer Data Explorer.</b> The interface introduces the SEER dataset, summarizes key characteristics, and provides interactive components such as the registry coverage map and diagnosis trends over time. It also presents the analysis pipeline, usage instructions, and acknowledgments, offering users a clear overview of the tool's scope and functionality.....                                                                                                                                                                                                                     | 87 |
| 4.31 | <b>Home page of the Breast Cancer Data Explorer (2).</b> The interface introduces the SEER dataset, summarizes key characteristics, and provides interactive components such as the registry coverage map and diagnosis trends over time. It also presents the analysis pipeline, usage instructions, and acknowledgments, offering users a clear overview of the tool's scope and functionality.....                                                                                                                                                                                                                 | 88 |
| 4.32 | <b>Sidebar navigation menu of the Breast Cancer Data Explorer.</b> The sidebar provides access to all major modules of the application, including the Home page, Dataset Description, Exploratory Analysis, Survival Analysis, Prognostic Factor Analysis, Prediction, Data, and About/Help sections. Its collapsible design ensures intuitive navigation while maintaining a clean interface. ....                                                                                                                                                                                                                   | 89 |
| 4.33 | <b>Dataset Description tab of the Breast Cancer Data Explorer.</b> This section summarizes the dataset before preprocessing, displaying the number of entries, variables, and missing values. Users can search or sort variables, and each variable card provides type, missingness, unique values, and descriptive statistics, with options to collapse or expand for detailed inspection.....                                                                                                                                                                                                                       | 90 |
| 4.34 | <b>Dataset Description tab of the Breast Cancer Data Explorer.</b> This section summarizes the dataset before preprocessing, displaying the number of entries, variables, and missing values. Users can search or sort variables, and each variable card provides type, missingness, unique values, and descriptive statistics, with options to collapse or expand for detailed inspection.....                                                                                                                                                                                                                       | 90 |
| 4.35 | <b>Exploratory Analysis tab of the Breast Cancer Data Explorer.</b> The tab displays each variable of the preprocessed dataset in interactive, sortable boxes. Users can visualize distributions and trends through bar charts, line plots, pie charts, histograms, and word clouds, filter by gender, or toggle between numeric and categorical representations for numeric variables. All plots are interactive, allowing users to hover over elements to view exact counts or values. Boxes can be collapsed or expanded for detailed inspection, enabling both high-level overviews and in-depth exploration..... | 92 |
| 4.36 | <b>Survival Analysis tab of the Breast Cancer Data Explorer.</b> Users can explore Kaplan-Meier survival curves and associated risk tables for overall or cause-specific survival. Survival can be stratified by any categorical variable in the dataset. Plots are interactive, allowing users to hover over time points to view group, survival probability, number at risk, and number of events. Log-rank p-values are displayed to assess differences between groups. ....                                                                                                                                       | 93 |
| 4.37 | <b>Prognostic Factor Analysis tab of the Breast Cancer Data Explorer.</b> Users can select variables to include in a Cox Proportional Hazards model to examine their impact on survival. The tab provides a summary table of hazard ratios with confidence intervals and p-values, a forest plot visualizing hazard ratios, and Cox-adjusted survival curves with risk tables. Plots are interactive, allowing users to hover to view exact values and facilitating detailed exploration of the influence of multiple covariates on patient survival. ....                                                            | 94 |
| 4.38 | <b>Data tab of the Breast Cancer Data Explorer.</b> The tab presents the SEER dataset in an interactive table format, allowing users to browse, filter, and sort patient-level records. Pagination and horizontal scrolling support navigation across a large number of observations, while column visibility controls enable customization of the displayed variables. Together, these features ensure transparency of the dataset and provide a flexible, user-friendly environment for exploring the variables. ....                                                                                               | 95 |

|                                                                                               |     |
|-----------------------------------------------------------------------------------------------|-----|
| 1. Kaplan-Meier Survival Curve for Age Categorical. ....                                      | 126 |
| 2. Kaplan-Meier Survival Curve for Breast Cancer Subtype. ....                                | 127 |
| 3. Kaplan-Meier Survival Curve for Chemotherapy Status. ....                                  | 128 |
| 4. Kaplan-Meier Survival Curve for Combined Summary Stage. ....                               | 129 |
| 5. Kaplan-Meier Survival Curve for Derived HER2 Status. ....                                  | 130 |
| 6. Kaplan-Meier Survival Curve for ER Status. ....                                            | 131 |
| 7. Kaplan-Meier Survival Curve for First Malignant Primary Indicator. ....                    | 132 |
| 8. Kaplan-Meier Survival Curve for Marital Status at Diagnosis. ....                          | 133 |
| 9. Kaplan-Meier Survival Curve for Median Household Income (Inflation Adjusted to 2022). .... | 134 |
| 10. Kaplan-Meier Survival Curve for PR Status. ....                                           | 135 |
| 11. Kaplan-Meier Survival Curve for Primary Site. ....                                        | 136 |
| 12. Kaplan-Meier Survival Curve for Race/Ethnicity. ....                                      | 137 |
| 13. Kaplan-Meier Survival Curve for Radiation Therapy Recode. ....                            | 138 |
| 14. Kaplan-Meier Survival Curve for PR Status. ....                                           | 139 |
| 15. Kaplan-Meier Survival Curve for Patients' Sex. ....                                       | 140 |
| 16. Kaplan-Meier Survival Curve for Time From Diagnosis to Treatment. ....                    | 141 |
| 17. Kaplan-Meier Survival Curve for Total Number of Benign/Borderline Tumors. ....            | 142 |
| 18. Kaplan-Meier Survival Curve for Total Number of In Situ Malignant Tumors. ....            | 143 |
| 19. Kaplan-Meier Survival Curve for Tumor Size. ....                                          | 144 |
| 20. Kaplan-Meier Survival Curve for Year of Diagnosis. ....                                   | 145 |



## List of Abbreviations

**AUC** Area Under the Curve.

**BC** Breast Cancer.

**BCSS** Breast Cancer-Specific Survival.

**CDSS** Clinical Decision Support Systems.

**CI** Confidence Interval.

**Cox** Cox Proportional Hazards.

**DTI** Diagnosis-to-Treatment Interval.

**ENN** ensembles of neural networks.

**ER** Estrogen Receptor.

**HER2** Human Epidermal Growth Factor Receptor 2.

**HR** Hazard Ratio.

**IARC** International Agency for Research on Cancer.

**IHC** Immunohistochemistry.

**KM** Kaplan-Meier.

**Log-rank** Log-rank Test.

**MAR** missing at random.

**MCAR** missing completely at random.

**MICE** Multiple Imputation by Chained Equations.

**MNAR** missing not at random.

**NIR** No Information Rate.

**NPV** Negative Predictive Value.

**OS** Overall Survival.

**PMM** Predictive Mean Matching.

**PPV** Positive Predictive Value.

**PR** Progesterone Receptor.

**PR** Precision-Recall.

**PR AUC** Precision-Recall Area Under the Curve.

**RF** Random Forest.

**ROC** Receiver Operating Characteristic.

**ROC AUC** Receiver Operating Characteristic Area Under the Curve.

**RUCC** Rural-Urban Continuum Code.

**SEER** Surveillance, Epidemiology, and End Results.

**SES** Socioeconomic Status.

**SHAP** SHapley Additive exPlanations.

**SMOTE** Synthetic Minority Oversampling Technique.

**TNM** Tumor, Node, Metastasis (Staging System).

**USA** United States of America.

**WHO** World Health Organization.

**XGBoost** Extreme Gradient Boosting.

# 1. Introduction

This chapter introduces breast cancer as a major global health challenge, highlighting its incidence, mortality, and disparities in outcomes across regions. It then discusses the importance of survival analysis in understanding prognosis and identifying high-risk patients, including distinctions between overall and cause-specific survival and the rationale for focusing on 5-year outcomes. The chapter presents the SEER database as a key resource for population-based survival studies, followed by an overview of predictive modeling in oncology and the integration of statistical and visual analytics to interpret complex clinical and epidemiological data. Finally, it outlines the research gap, motivation, and objectives of this study, providing the context for the development of an interactive Shiny application to support personalized risk assessment and decision-making.

## 1.1. Breast Cancer as a Global Health Challenge

Breast cancer (BC) poses a persistent global health challenge, not only because of its rising incidence and mortality but also because outcomes differ substantially across regions in terms of detection, treatment, and survival [1, 2]. According to the International Agency for Research on Cancer (IARC), a specialized cancer agency of the World Health Organization (WHO), in its most recent studies [3], BC is the second most diagnosed cancer globally and the leading cause of cancer-related mortality among women. In 2022, approximately 2.3 million women received a breast cancer diagnosis, resulting in approximately 670,000 deaths worldwide, and by 2050, projections estimate more than 3 million new cases and more than 1 million deaths annually, largely driven by population aging and growth.

Despite progress in early detection and treatment, global survival rates remain uneven due to geographic, socioeconomic, and biological factors [4]. The burden of breast cancer increases most rapidly in regions with lower sociodemographic indices and countries undergoing economic transition [5]. Survival rates vary greatly between nations, with significantly lower outcomes observed in low- and middle-income countries (LMICs), where patients are more likely to be diagnosed in advanced stages due to limited access to early detection, effective treatment, and robust healthcare infrastructure. In contrast, high-income countries benefit from widespread screening programs and better treatment

modalities, contributing to earlier detection and improved prognosis. Differences in socioeconomic status, public awareness, and healthcare infrastructure contribute heavily to the uneven survival patterns observed worldwide [4].

Additionally, BC is a heterogeneous disease characterized by various tumor subtypes, biological behaviors, and patient responses, making standardized treatment and management a persistent challenge [6]. In this context, early diagnosis, accurate risk stratification, and timely intervention are critical to improve survival and quality of life in patients with breast cancer. However, since the complexity of clinical and epidemiological data often hinders timely and evidence-based decision-making, tackling the global burden of breast cancer demands not only investment in prevention, screening, and treatment, but also strengthening health systems and the development of innovative analytical tools [7]. Given its magnitude and weight in public health, understanding prognostic factors and developing predictive tools are crucial for clinical support.

## 1.2. Importance of survival analysis in breast cancer

Survival studies model the time from diagnosis or treatment to specific outcomes, like death or recurrence, enabling researchers and clinicians to estimate survival rates and identify high-risk groups. These analyses help in visualizing patient prognosis across different subgroups, such as by stage, tumor size, age, or treatment modality [8]. These analyses are essential for understanding breast cancer prognosis because they allow clinicians and researchers to track outcomes over time and to pinpoint which factors most strongly shape survival after diagnosis and treatment [9].

The key difference between overall survival and cause-specific survival lies in what is counted as an endpoint (event) in survival analyses. Overall survival measures the length of time from diagnosis (or treatment) to death from any cause, including deaths due to cancer, other diseases, accidents, or any other reason, and reflects the total mortality experience of the patient group instead of distinguishing the cause of death [10]. Cause-specific survival measures the length of time from diagnosis (or treatment) to death, specifically from the disease of interest [11]. Deaths from all other causes are treated as censored, not as events in the analysis, and only disease-specific deaths count toward the endpoint, providing a clearer picture of mortality risk directly attributable to the cancer itself [10].

In cancer research, five-year survival is often used as a benchmark because it offers a practical window for comparing outcomes across populations while capturing the period in

which most recurrences and cancer-related deaths occur, reflecting both immediate and medium-term effects of diagnosis and treatment [12]. The choice of a 5-year analysis is based on the fact that many recurrences and cancer-related deaths occur within the first 5 years after diagnosis, making this period critical for monitoring and evaluating treatment success. This period, reported as closely correlated to long-term survival in many cancers, including breast cancer [13], is long enough to capture major outcomes, yet short enough for reliable and timely reporting [14].

Clinical and demographic prognostic factors are consistently set in epidemiological studies as having a significant influence on patient survival across a wide range of diseases, including cancers [15]. These factors include age, gender, race, ethnicity, and disease stage, among others.

### 1.3. The SEER dataset as a research resource

The Surveillance, Epidemiology, and End Results (SEER) program is a consortium of cancer registries that covers approximately 50% of all cancer patients in the United States of America (USA). It collects detailed demographic, tumor, and treatment data alongside longitudinal survival follow-up, making it highly representative and well-suited for population-based survival analysis [16]. SEER data include variables such as age, sex, race, tumor site, stage at diagnosis, and vital status information, which allow for comprehensive survival analyses and understanding of temporal survival trends.

SEER's longitudinal coverage, with data spanning decades (from 1970s to recent years) and annual updates, provides rich, detailed survival information over time. This enables researchers to examine time-varying associations of patient and tumor characteristics with survival outcomes and allows modeling that accounts for competing risks and changing hazards over post-diagnosis time [17].

### 1.4. Predictive modeling in oncology

In oncology, predictive models are increasingly used to identify high-risk patients at an earlier stage, with the long-term aim of integrating these tools into Clinical Decision Support Systems (CDSS) [18]. These models analyze clinical, demographic, genetic, and lifestyle data to stratify patients by risk, enabling timely interventions aimed at preventing adverse events, improving patient outcomes, and optimizing resource allocation [19].

Predictive models are extensively used to identify patients at high risk of treatment-related toxicities, disease progression, or emergency events during cancer therapy. This early

risk stratification allows healthcare providers to tailor surveillance intensity and intervene earlier, potentially reducing hospitalizations and improving quality of life [20–22]. For example, programs pairing predictive risk algorithms with targeted interventions like remote symptom monitoring have shown promise in reducing preventable acute care visits among cancer patients by identifying those most likely to benefit from such care [22].

The integration of predictive models into CDSS offers clinicians actionable insights by combining risk stratification with personalized treatment planning. Predictive analytics can guide decisions on therapeutic intensity, monitoring frequency, and resource prioritization tailored to individual risk profiles. These systems can also support dynamic clinical workflows by incorporating updated patient data, enhancing treatment efficacy while minimizing unnecessary interventions [23]. Despite challenges like data quality, ethical considerations, and clinical adoption, such integration holds promise for transforming oncology care delivery [24].

With this being said, predictive modeling in oncology facilitates early identification of high-risk patients and, when integrated into clinical decision support systems, offers the possibility of making cancer care more personalized and better targeted to individual patient needs.

## 1.5. Statistics vs. Visual Analytics

Combining traditional statistical methods with visual analytics helps reveal complex patterns in breast cancer survival data that would be harder to interpret with numbers alone. Techniques such as Kaplan-Meier survival estimates and Cox proportional hazards regression provide robust quantitative assessments of the relationship between clinical factors and patient outcomes, yielding statistically significant insights into prognostic variables. However, interpreting these results can be challenging, particularly in large, multi-dimensional datasets.

Visual analytics (VA) addresses this challenge by transforming statistical outputs into intuitive, interactive graphical representations, enhancing comprehension and facilitating hypothesis generation. By enabling users to explore survival trends across subgroups, detect time-dependent changes, and identify anomalies, VA supports iterative exploration and real-time testing of hypotheses [25].

As an interdisciplinary approach, VA combines human intuition with computational techniques, leveraging interactive visualizations alongside advanced analytical models. This

combination allows users to manipulate data dynamically, by adjusting filters, parameters, and perspectives, and observe immediate effects on visual outputs. Interactive exploration supports tasks such as spotting anomalies, comparing subgroups, or developing new hypotheses, capabilities that static tables and charts generally lack [26].

In breast cancer research, where clinical datasets encompass heterogeneous variables such as tumor size, histological subtype, genetic markers, treatment history, and survival outcomes, VA might be especially valuable. These tools facilitate the integration and exploration of diverse data types, supporting patient stratification, identification of subpopulations with distinct outcomes, and simulation of treatment strategies. By enabling interactive and targeted analysis, VA ultimately contributes to personalized care and more effective public health interventions.

## 1.6. Research gap and motivation

Clinical stage is a well-established prognostic factor in breast cancer, with five-year survival rates exceeding 99% in localized cases and declining to approximately 32% in metastatic disease. Nevertheless, numerous other variables, including patient age, histological grade, and hormonal receptor status, also play critical roles in determining prognosis. The integration of these factors using large-scale population-based datasets, such as the Surveillance, Epidemiology, and End Results (SEER) program, enables a comprehensive evaluation of their relative prognostic importance and the identification of complex interactions that may not be evident in isolated analyses. Recent studies have demonstrated that machine learning models can outperform traditional statistical approaches in predicting clinical outcomes. However, their adoption in clinical practice remains limited, primarily due to challenges related to model interpretability and the lack of interactive tools to support the exploration and application of these models in real-world settings.

This research aims to make contributions on both a scientific and a practical level. Scientifically, it will provide a robust, large-scale quantification of prognostic clinical and demographic factors influencing breast cancer survival, filling gaps in the current literature by leveraging a comprehensive dataset spanning over two decades. It will also provide empirical evidence on whether machine learning models can improve predictive accuracy compared to traditional survival analysis, potentially influencing future prognostic modeling approaches in oncology.

Practically, the development of a personalized decision-support tool will facilitate more precise risk stratification in clinical settings, enabling healthcare providers to identify high-risk patient subgroups that may benefit from intensified monitoring or targeted therapies. This aligns with precision medicine goals by tailoring interventions based on individualized risk profiles.

Furthermore, our Shiny application represents a technological contribution by democratizing access to complex analytical results and enabling interactive exploration by clinicians, researchers, and policy-makers. Finally, by demonstrating the value of the SEER database as an open, large-scale resource, this study encourages similar research efforts and promotes open science principles within cancer epidemiology.

## 1.7. Objectives of the study

The main research question for this study is:

Which clinical and demographic factors are most strongly associated with 5-year survival among breast cancer patients diagnosed between 2000 and 2021 in the SEER database, and to what extent can machine learning models accurately predict individual survival outcomes?

Consequently, we derive the specific objectives of this study:

- Characterize the population (age distributions, stage, molecular subtype, etc.) and assess data quality (outliers, missingness) for the dataset.
- Identify prognostic variables using statistical analyses (Kaplan-Meier, Cox).
- Develop machine learning models to predict 5-year survival and compare the results with those of traditional statistical approaches.
- Develop an interactive Shiny application that integrates visualizations, data exploration, and individualized forecasting.

## 1.8. Thesis Organization

The thesis is organized into five chapters. Chapter 2 provides the state of the art, presenting a literature review on survival analysis in oncological datasets, discussing the main prognostic factors in breast cancer, common survival analysis methods, and relevant machine learning approaches in this domain. The chapter also examines existing

visualization systems, their limitations, and concludes with an overview of visual analytics strategies using R and Shiny. Chapter 3 introduces the methods and results related to the SEER dataset, detailing the preprocessing of variables and exploratory data analysis through both traditional statistical techniques and machine learning models. This chapter also describes the development of the Shiny application and the technical aspects of its implementation. Chapter 4 focuses on the results and discussion, presenting the outcomes of univariate analyses, Chi-square tests, Kaplan-Meier survival curves, the Cox proportional hazards model, and machine learning approaches. A comparative assessment of the models' performance is also included. Finally, Chapter 5 summarizes the main conclusions of the study and outlines potential directions for future research.

## 2. State Of The Art

The literature reviewed in this work was selected through a narrative, semi-structured process aimed at identifying relevant contributions in the areas of breast cancer research, clinical and epidemiological data analysis, and visual analytics. The initial search strategy was guided by exploratory use of the Consensus tool and complemented with database searches in Web of Science, PubMed, and Google Scholar. Keywords and queries combined terms such as “breast cancer”, “visual analytics”, “clinical data”, “prognostic factors”, “SEER dataset”, “predictive modeling”, and “machine learning survival analysis”.

The inclusion criteria focused on studies that were published in peer-reviewed journals or presented at international conferences, that directly addressed clinical or epidemiological aspects of breast cancer, proposed or evaluated visualization or computational approaches for health data, and were published within the last two decades to ensure contemporary relevance. Exclusion criteria included works not published in English, purely theoretical contributions without empirical application, and publications in venues with limited academic impact. It should be noted that this review does not follow the strict methodology of a systematic review. Instead, it adopts a narrative approach, designed to capture a representative and thematically relevant body of literature that situates the thesis within the broader research landscape.

This section synthesizes current knowledge on key prognostic factors influencing survival in breast cancer patients, with particular emphasis on studies employing the Surveillance, Epidemiology, and End Results (SEER) dataset and machine learning methodologies for survival prediction in oncology. The integration of clinical and epidemiological data through advanced computational tools underscores the context in which this thesis is situated, while highlighting its novel contributions.

### 2.1. Prognostic factors in breast cancer survival

#### 2.1.1 Tumor-Intrinsic Factors

##### **Stage at Diagnosis and Tumor Burden**

One of the most consistently reported determinants of breast cancer survival is the stage at diagnosis. A comprehensive meta-analysis encompassing over 1.3 million patients from 30 countries found that advanced stages were associated with markedly increased

disease-specific mortality. In more detail, compared with early-stage cases, those diagnosed at Stage III were more than three times as likely to die, and those at Stage IV were about twelve times as likely. Tumor size reinforces this gradient as mortality increases from 6.9%, for tumors smaller than 10 mm, to more than 60%, for tumors greater than 90 mm [27]. Lymph node involvement refines risk classification as a positive nodal status increases the risk of recurrence and decreases survival. The importance of early detection is therefore confirmed not only by stage but also by the prognostic significance of tumor aggressiveness and anatomical spread [28, 29].

### **Molecular Subtype and Receptor Status**

Molecular profiling has allowed for more refined prognostication by distinguishing subtypes based on ER, PR, and HER2 receptor status. Luminal A tumors (ER+/PR+/HER2-) demonstrate the most favorable outcomes, while triple-negative breast cancer (TNBC), lacking all three receptors, displays the poorest prognosis [30, 31]. Hormone receptor positivity (ER+ and/or PR+) confers a more favorable prognosis, both short- and long-term, with increased responsiveness to endocrine therapies and lower risk of aggressive recurrence [32, 33]. HER2-positive cancers, once considered aggressive, now benefit from targeted therapies, having significantly improved survival [34]. However, reliance on tissue biopsies for molecular characterization introduces spatial and temporal limitations, particularly in metastatic settings, where biomarker discordance may lead to suboptimal treatment.

#### **2.1.2 Host Factors**

##### **Age and Sex**

Age remains a paradoxical prognostic factor as younger women (diagnosed before 40) are more likely to present with aggressive tumor characteristics, like high-grade, lymph node involvement, and negative receptor status, resulting in poorer outcomes [35]. However, older patients often contend with comorbidities and receive less aggressive treatment, especially for hormone receptor-positive tumors, potentially diminishing survival despite biologically less aggressive disease [36]. Multivariate analyses confirm that both age extremes independently predict poorer outcomes, though via distinct pathways. Sex hormones (estrogens and androgens) play a crucial role in breast cell growth and division as their levels and fluctuations can influence breast cancer risk and prognosis. Age and menopausal status affect hormone levels and can influence breast cancer risk, with

postmenopausal women having a higher risk due to longer exposure to estrogen [37]. In breast cancer, sex is a complex prognostic factor with both direct and indirect influences on disease development and progression. Male sex, while rare in breast cancer cases, is associated with worse overall survival, largely attributable to delayed diagnosis and older age at presentation [38]. The molecular applicability of female-derived subtype classifications to male breast cancers remains uncertain, further complicating prognostic and treatment standardization in this group [39].

### 2.1.3 Treatment-Related Factors

#### **Diagnosis-to-Treatment Interval (DTI)**

The timeliness of care significantly influences outcome and, while the ideal DTI varies depending on the specific treatment, generally, delays of more than 90 days for surgery and 120 days for chemotherapy are associated with poorer outcomes. For radiotherapy, when chemotherapy is administered, a DTI of less than 365 days is recommended [40]. A DTI  $\geq 60$  days in early-stage cancer is especially detrimental [41]. Prolonged intervals can allow disease progression, potentially altering the stage and treatment options [42].

#### **Radiotherapy and Chemotherapy**

Radiotherapy, while crucial for local tumor control in breast cancer, can also be a prognostic factor, particularly when considering its impact on local recurrence-free survival and overall survival. Specifically, factors like the presence of cancer-related fatigue after radiotherapy, certain pathological features, and the extent of lymph node involvement can influence prognosis. Adjuvant radiotherapy, particularly after breast-conserving surgery, significantly reduces local recurrence rates. However, factors like positive surgical margins and late start of radiotherapy can increase the risk of local recurrence [43]. Estrogen receptor (ER) positivity is generally associated with better outcomes in patients receiving radiotherapy after neoadjuvant chemotherapy and radical mastectomy [44]. Chemotherapy remains pivotal, particularly when a pathological complete response (pCR) is achieved following neoadjuvant treatment, a strong predictor of long-term survival. Yet its benefit is highly subtype-specific [45]. The degree to which a tumor responds to chemotherapy (either a complete or partial response) can be a key factor in predicting overall survival and disease-free survival [46].

## 2.1.4 Structural and Sociodemographic Determinants

### **Socioeconomic Status (SES) and Income**

The relationship between income and breast cancer is complex and influenced by various factors like lifestyle, access to healthcare, and tumor biology, with other socioeconomic factors like education level and access to health insurance also playing a crucial role in breast cancer outcomes [47, 48]. Studies show that women from higher income households tend to have higher breast cancer incidence rates but also tend to have better survival rates after diagnosis. Conversely, women from lower income households may experience lower breast cancer incidence but are more likely to have poorer survival outcomes [47]. Women in higher-income areas are more likely to undergo regular mammograms, which can lead to earlier detection of breast cancer, possibly explaining the higher overall incidence rate in these areas [49].

### **Race, Ethnicity, and Geography**

Ethnicity and race remain prominent epidemiological predictors and, although incidence rates among non-Hispanic black and non-Hispanic white women are comparable, Black women are more often diagnosed at younger ages with aggressive tumor types and face poorer survival. Mortality among Black women is about 42% higher than in White women, and there is evidence to think that these differences are linked to later-stage diagnoses, limited healthcare access, biological and genetic tumor characteristics, and higher prevalence of risk factors. Additional contributors include reduced access to screening, lack of insurance, and unequal availability of recent treatment advances [50]. Hispanic women are more likely than non-Hispanic Whites to present with aggressive subtypes, particularly triple-negative breast cancer (TNBC), which contributes to worse outcomes because TNBC responds poorly to targeted therapies [51]. At the same time, Hispanic breast cancer patients are about 2.5 times more likely than Whites to experience treatment delays over three months, which adversely impacts prognosis [52]. The Rural-Urban Continuum Code (RUCC) can be a prognostic factor in breast cancer, with rural residence often associated with a higher likelihood of late-stage diagnosis. This means individuals living in rural areas may be diagnosed with breast cancer at a more advanced stage, potentially impacting treatment options and outcomes [53]. Rural patients are also more likely to receive less guideline-concordant treatment [54] and face logistical barriers such as travel distances to treatment centers [55].

## Marital Status

Marital status at diagnosis can be a prognostic factor in breast cancer, with unmarried women including those who are single, divorced, or widowed, often exhibiting poorer survival outcomes compared to their married counterparts. This difference in survival can be attributed to factors like later stage at diagnosis, lower treatment uptake, reduced psychosocial support and potential socioeconomic disparities [56, 57]. Interestingly, a study conducted by Martínez et al. [58] showed that widowed women were more likely to be older and to be non-Hispanic white (NHW), and less likely to be uninsured than married or single patients. Furthermore, married women were less likely to reside in low SES neighborhoods and more likely to have private insurance than unmarried women. Lastly, widowed patients had a lower proportion of HER2+ and triple negative tumors than other women, while also being less likely to undergo surgery and to have chemotherapy or radiation therapy.

## 2.2. Survival analysis methodologies in oncology

The landscape of breast cancer prognosis is shaped by a complex interplay of biological, demographic, clinical, and social variables. Recent studies leveraging the SEER database [59–62] and other population-based datasets have employed various statistical and modeling techniques to dissect these dimensions.

A common statistical foundation among several studies is the use of Cox proportional hazards modeling for estimating hazard ratios (HRs) related to survival. However, Cai et al. [59] distinguish themselves by incorporating multivariable fractional polynomial modeling to assess the non-linear relationship between age and breast cancer-specific mortality (BCSM). This technique allows for greater flexibility in capturing complex, non-linear effects, contrasting with the more standard linear treatment of covariates seen in the other studies. Additionally, they introduced interaction terms between age and immunohistochemical (IHC) subtype within their Cox model, enabling subtype-specific hazard estimation, a refinement that most directly addresses biological heterogeneity.

In a study conducted by Moore et al. [60] expanded traditional modeling by applying multilevel logistic regression (i.e., Generalized Linear Mixed Models) to account for the nested structure of patients within counties, thus adjusting for contextual variables such as county-level socioeconomic status (SES) and healthcare access. This hierarchical

modeling approach addresses between-county variability and allows for the disentangling of individual versus system-level influences on treatment access and late-stage diagnosis. For survival outcomes, they reverted to a standard Cox model but with stepwise adjustment schemes, sequentially incorporating covariates such as stage and treatment to quantify the contribution of each variable set. This layered modeling strategy allowed for more nuanced interpretation of disparities in BCSM.

In contrast, Wei et al. [61] used more conventional statistical techniques but applied them to tumor location, an underexplored variable. The primary methods included Kaplan-Meier survival analysis, log-rank tests, and multivariable Cox regression. Notably, their study emphasized subgroup analyses stratified by T and N stages and age, allowing them to examine how the prognostic effect of tumor site varied across clinical strata. Although their modeling remained within a standard framework, they advanced predictive capabilities by developing nomograms for both overall survival (OS) and breast cancer-specific survival (BCSS). These nomograms integrated multiple covariates and were evaluated using C-index and calibration plots, providing a tool for individualized risk prediction, one of the few steps toward clinical translation seen in the compared studies.

Liu et al. [62] also employed standard Cox modeling, adjusting for demographic and clinical covariates to isolate the prognostic impact of marital status in inflammatory breast cancer (IBC). Their focus on subgroup analyses by AJCC stage helped confirm the consistency of marital status as an independent predictor across clinical severity. However, their methodological scope was narrower, lacking the interaction modeling or predictive tools seen in other studies.

Despite differences in research focus, the statistical pipelines share a reliance on Kaplan-Meier survival estimation, log-rank testing, and multivariable Cox regression as core techniques. Where the studies diverge, and where analytical sophistication increases, is in the use of flexible modeling frameworks (e.g., fractional polynomials), hierarchical models, interaction terms, and predictive tools such as nomograms. Cai et al.'s implementation of non-linear modeling and interaction effects, and Moore et al.'s multilevel regression strategy, are particularly notable for increasing model interpretability and specificity, especially in heterogeneous populations. Another key distinction is in the handling of confounders and covariate adjustment. While all studies controlled for standard clinical variables (e.g., age, stage, treatment), Moore et al.'s tiered model construction allowed for disentangling the influence of structural inequities. In contrast, studies like Liu et al. applied a single,

fully adjusted model, limiting interpretive flexibility. From a methodological perspective, the integration of stratified analyses, interaction terms, and predictive modeling represents a clear evolution in survival analysis, moving from general associations toward personalized risk assessment. These approaches not only improve the granularity of insights but also enhance the translational potential of prognostic models.

In summary, while the studies vary in focus, they collectively illustrate how methodological choices, ranging from model type (standard vs. multilevel), variable treatment (linear vs. non-linear), and modeling depth (associative vs. predictive), critically influence the power and applicability of findings. As breast cancer research advances, the integration of advanced statistical frameworks, especially those accounting for heterogeneity, structural context, and patient-level prediction, will be essential for developing robust, equitable, and clinically actionable prognostic tools.

### 2.3. Machine learning approaches for survival prediction

In recent years, data science has profoundly reshaped oncological research and clinical decision-making by applying machine learning (ML), ensemble strategies, deep learning (DL), and hybrid models to diverse cancer datasets. With the rise of the technological revolution, AI, particularly machine learning, has begun to play a transformative role in many domains, especially in the diagnosis and treatment of diseases. Advances in ML have enabled early and more accurate detection of complex conditions, such as cancer, by leveraging large-scale datasets [63]. Analyzing and comparing recent ML studies in different oncological fields, focusing on methodological choices, model performance, data preprocessing strategies, interpretability, and generalizability, it is possible to identify commonalities and distinctions that influence clinical applicability and predictive reliability.

#### **Model Performance Across Studies**

XGBoost consistently emerged as a top performer across different cancer types [64–66], likely due to its robustness with tabular data and hyperparameter optimization.

In Song et al.'s study on 5-year breast cancer survival, XGBoost achieved the highest Area Under the Curve (AUC)(0.813), outperforming traditional models such as logistic regression, support vector machines, and k-nearest neighbors. Similarly, in a study predicting lymph node metastasis in osteosarcoma, Li et al. reported that XGBoost achieved an AUC of 0.874 with an accuracy of 0.851, outperforming five other classification models

including random forests and naïve Bayes. However, in datasets characterized by severe class imbalance, ensemble methods did not always dominate in every metric. Kim et al., who studied lung metastasis in hepatocellular carcinoma using SEER data, found that while XGBoost achieved the highest overall accuracy (0.87), it was a neural network trained with a recall-optimized loss function that achieved the best sensitivity (0.66), which was later improved to 0.77 via an ensemble approach, underscoring the trade-off between accuracy and recall in clinical tasks where detecting rare outcomes is critical.

Reshan et al. [67], in contrast, achieved nearly perfect classification performance in a binary tumor classification task using the WDBC dataset. By employing a stacking ensemble with a neural network as the meta-classifier, they reached an AUC of 1.00 and an accuracy of 99.89%. However, this result should be interpreted with caution, as the dataset is relatively small and clean, making it prone to overfitting, especially when using high-capacity models like deep learning.

### **Data Preprocessing and Feature Engineering**

Preprocessing strategies varied considerably across studies, affecting model performance and reproducibility. Song et al. (2024) implemented a comprehensive pipeline that included encoding categorical variables, multiple imputation using MICE, and normalization of continuous features. This level of preprocessing helped ensure that each model received appropriately formatted and complete data, minimizing the impact of missingness and scale differences.

In contrast, Kim et al. [65] placed greater emphasis on class imbalance by applying SMOTE and customizing their neural network's loss function to optimize for recall. Their preprocessing approach was tailored to the task of metastasis detection, where the positive class was significantly underrepresented. This strategy directly influenced the model's ability to detect rare outcomes, demonstrating how domain-specific preprocessing choices can affect the choice of optimal algorithms.

Reshan et al.'s study [67] used recursive feature elimination with cross-validation (RFECV) to reduce dimensionality, which likely contributed to their high performance. However, their report did not include details on handling missing values or encoding strategies, limiting the reproducibility of their pipeline. Similarly, Li et al. [66] did not specify imputation methods but incorporated clinical features from the SEER database along with external institutional data, expanding the scope of their feature set. This decision likely improved

model generalizability despite the absence of preprocessing transparency.

### **Evaluation Metrics and Validation Techniques**

All reviewed studies applied internal validation techniques, typically using 80:20 train-test splits combined with cross-validation. Song et al. [64] explicitly tested multiple data splits and selected the 80:20 configuration based on cross-validation performance. Reshan et al. [67] applied cross-validation not only during training but also in feature selection, strengthening the model's robustness against overfitting.

Among the reviewed studies, only Li et al. [66] implemented external validation, evaluating their model using data from four geographically distinct hospitals. This approach provided stronger evidence for generalizability and real-world applicability, which is often missing in ML studies confined to single-institution datasets.

Across studies, evaluation typically relied on common metrics such as AUC, accuracy, recall (sensitivity), specificity, precision, and F1-score. However, the prioritization of these metrics varied. Kim et al. [65] emphasized recall due to the clinical importance of identifying lung metastases, which are rare but critical. Their use of a recall-optimized loss function is a noteworthy methodological innovation that other studies did not adopt. Song et al. [64], by contrast, balanced evaluation between sensitivity and specificity, aiming for well-rounded performance rather than optimizing a single metric.

### **Interpretability and Feature Importance**

A major barrier to clinical adoption of ML models is their limited interpretability, yet only a subset of the reviewed studies explicitly addressed it. Song et al. [64] distinguished themselves by integrating SHapley Additive exPlanations (SHAP) to explain the predictions of their XGBoost model, identifying metastatic lymph node involvement, tumor size, and endocrine therapy as the most influential predictors of 5-year survival. These results aligned with clinical expectations, providing reassurance that the model captured medically relevant relationships. Karatza et al. [68] also prioritized interpretability in their application of random forests, neural networks, and ensembles of neural networks (ENN) to the WDBC dataset. They demonstrated that ENN achieved the best predictive performance, with an accuracy of 96.6% and an AUC of 0.96. Importantly, they applied Individual Conditional Expectation (ICE) plots to visualize how changes in individual features influenced the

model's predictions. The consistency of these explanations with current medical understanding supported the clinical plausibility of the model and indicated its potential for uncovering novel pathophysiological insights into breast cancer. Furthermore, Karatza et al. employed feature selection based on interpretability tools to enhance model performance. Feature importance derived from a Global Surrogate model improved the accuracy of the random forest from 96.49% to 97.18% and the AUC from 0.96 to 0.97. Similarly, feature selection using Shapley values improved the neural network's accuracy from 94.6% to 95.53%, with a corresponding AUC increase from 0.94 to 0.95. These findings emphasize that interpretability is not merely a transparency mechanism but can also serve as a tool for optimizing model architecture and generalization. In contrast, several other studies neglected explainability despite employing complex models such as deep learning and ensemble stacks. Reshan et al. [67] and Kim et al. [65] did not report any interpretability methods, even though their models achieved high accuracy on metastasis prediction tasks. Li et al. [66], while validating their XGBoost model on external clinical cohorts, also did not include any interpretability techniques. The absence of explanation frameworks in these studies limits the transparency and clinical trustworthiness of otherwise promising models.

### **Generalizability and Dataset Limitations**

The generalizability of ML models is often constrained by data source and cohort diversity. Song et al. [64] and Kim et al. [65] relied on retrospective datasets from single institutions or national registries like SEER, which may introduce selection biases and limit applicability to broader populations. Reshan et al. [67] used the WDBC dataset from the UCI repository, a clean and well-curated dataset that is useful for benchmarking but lacks clinical complexity and external variability.

In contrast, Li et al. [66] made a significant contribution by including external validation using clinical data from four different hospitals. This step strengthened the model's robustness and demonstrated its applicability across varying patient demographics and clinical practices. Despite the lack of interpretability tools, their methodological design offers a useful template for future ML research in oncology.

### **Comparative Summary and Key Insights**

Despite differences in cancer type, outcome, and data source, several themes emerged.

XGBoost repeatedly delivered strong performance across tasks and datasets, confirming its status as a reliable model for structured clinical data. However, in settings characterized by class imbalance and rare outcomes, such as metastasis prediction, neural networks with customized loss functions provided superior recall, illustrating the value of tailoring model design to clinical objectives. Interpretability tools like SHAP should become standard practice, especially in studies aiming for clinical deployment, as models that achieve high accuracy without explaining their decisions may face skepticism from healthcare providers. Furthermore, external validation remains a major gap in most studies and should be prioritized to ensure real-world utility.

## 2.4. Visualization Systems' Strengths and Weaknesses

Static visualizations and simple plots offer clear, interpretable, and familiar means to represent and communicate complex cancer research data, supporting quick understanding and reliable comparison across cohorts and outcomes. Simple visualizations, such as bar charts, line graphs, and Kaplan-Meier curves, enable rapid interpretation of data by both researchers and clinicians, as these forms of representation reduce cognitive load and make patterns, trends, and outliers immediately apparent [69]. The combination of quantitative charts with concise explanatory text further enhances the educational value of visualizations, supporting the efficient communication of information in minimal space [70]. Clinicians and researchers often favor traditional visualizations, such as line and bar graphs, due to their long-standing use and the ease with which they can be interpreted, making familiarity and reliability additional strengths of static plots. This familiarity fosters trust in the results and minimizes the risk of confusion, which can arise when unfamiliar or overly complex graphics obscure the intended message or require extensive explanation [69, 71]. Their simplicity also facilitates sharing in publications, presentations, and digital platforms, ensuring consistent dissemination of findings to diverse audiences, including scientists, clinicians, and policymakers. Such visualizations are particularly valuable in collaborative contexts and during peer review, where transparency, reproducibility, and clarity are essential [70, 71]. Finally, static charts contribute to statistical rigor and comparability in clinical research as widely used plots, including Kaplan-Meier curves, underpin core analyses in oncology, enabling direct comparisons of survival rates, treatment effects, and other key outcomes across studies. By supporting accurate reporting and reducing potential bias in the communication of risks or harms, these visualizations not only strengthen the integrity of research findings but also enhance understanding and inform behavioral intentions based on study results [72, 73].

However, static visualizations and simple plots can oversimplify complex cancer datasets, struggle with multidimensional or high-volume data, and may lead to misunderstanding if details or interactive exploration are needed. Static charts, such as bar and line graphs, cannot capture multiple variables, complex interactions, or spatial relationships, which are increasingly crucial in advanced cancer research, tissue imaging, and multi-omics contexts. At the same time, important nuances, heterogeneity, or trends may be hidden or lost in overly simplistic visual formats or summary-level plots [74]. While widely preferred, simple plots are not always the easiest to interpret, as axis labeling, legend design, and color choices can introduce confusion or misinterpretation. With familiarity not always correlating with understanding, what clinicians find intuitive may still mask errors or limit recognition of patterns in the data [69]. Another common limitation is the fact that static graphics do not allow users to drill down, filter, or interactively explore the data, restricting deeper hypothesis generation or verification of underlying patterns. The inability to zoom, pan, or dynamically interrogate subsets is also a downside of static plots, as it prevents revealing correlations, regional variations, or subgroup effects in large-scale datasets. [? ].

Despite the availability of analytical tools for oncology, many platforms still rely heavily on static outputs and lack support for real-time interaction or user feedback [75, 76]. Moreover, they limit the ability of clinicians and researchers to explore data dynamically, drill down into patient details, or tailor visualizations to specific clinical questions in real time, and, at the same time, static presentations tend to be more rigid and less adaptable for different users or evolving data sets, which reduces flexibility and responsiveness in evolving clinical and translational research environments.[77] Usability remains a critical barrier in clinical and public health contexts, as interfaces that demand advanced technical skills or programming knowledge often discourage adoption by clinicians, public health practitioners, and multidisciplinary research teams. For clinicians, such tools may disrupt established workflows and increase cognitive or time burdens, while public health professionals frequently face additional challenges stemming from limited programming familiarity or infrastructure [78]. These factors align with findings from technology acceptance studies, which emphasize that ease of use and minimal disruption are decisive predictors of adoption [79]. In addition to these challenges, relatively few platforms currently provide seamless integration of clinical and epidemiological variables within a single exploratory environment. While some initiatives highlight the potential of integrated tools [80–82], most require significant data wrangling or are optimized for acute or narrow-use

cases. As a result, many research teams must still rely on fragmented workflows, combining multiple platforms or programming-intensive environments to achieve comprehensive analysis. Addressing these usability and integration gaps is essential to expanding accessibility, fostering collaboration, and enabling the generation of actionable insights in breast cancer research.

#### 2.4.1 The Need for Interactive and Accessible Tools

In the era of digital health and personalized medicine, the demand for tools that effectively support clinical and epidemiological research in breast cancer is rapidly increasing. These tools must not only be computationally powerful but also interactive and accessible, enabling researchers and clinicians to navigate complex datasets and extract meaningful insights with ease. As the volume and complexity of health data continue to grow, the ability to translate raw information into clinically relevant insights becomes essential. Clinicians, for example, are often confronted with large amounts of patient data that are difficult to interpret quickly. Visual and interactive tools that can summarize outputs, such as risk scores, survival estimates, or predictions from machine learning models, can alleviate cognitive burden and help practitioners make informed decisions in real time. Such tools are especially valuable as clinical decision aids, where they can flag high-risk patients, guide treatment planning, and align care with the latest evidence-based practices.

Accessibility is another critical aspect, particularly for researchers and healthcare systems in low- and middle-income countries (LMICs), where access to advanced computational infrastructure is limited. The development of lightweight, web-based, and open-source tools offers a viable path toward equitable participation in data-driven research and healthcare delivery. These tools can be adapted to different languages and cultural contexts, ensuring their relevance in diverse settings. By supporting localized research efforts and community-based outreach, accessible tools can contribute to closing global gaps in breast cancer outcomes and promote more inclusive research agendas.

Furthermore, interactive tools hold significant potential to foster collaboration across disciplines. Breast cancer care and research involve professionals from various backgrounds, including oncology, epidemiology, public health, and biostatistics. Many of these users may lack programming experience yet still need to engage with complex data. Tools designed with intuitive interfaces, featuring elements like drag-and-drop functionality, sliders, and visual filters, can democratize data exploration and analysis. They also serve as

powerful educational resources, providing hands-on experiences for students, clinicians-in-training, and public health professionals, and enhancing data literacy across the health-care continuum. Another essential function of these tools is to support patient-centered care. Patients tend to understand visual representations of risk and prognosis more effectively than statistical summaries. Interactive platforms that generate personalized risk curves or simulate treatment outcomes can foster shared decision-making between patients and clinicians, helping individuals better understand their condition and participate in their own care. By illustrating the potential impacts of different clinical scenarios, such as treatment delays or specific therapy choices, these tools can improve communication, support informed consent, and build trust in the clinical process.

From a research perspective, interactive tools enable real-time hypothesis generation and exploration. Traditional methods of testing new ideas often require scripting and collaboration with data analysts, which can slow down discovery. By contrast, tools that allow users to filter datasets by variables such as age, tumor subtype, or disease stage, without writing code, empower researchers to quickly identify patterns, test assumptions, and refine their questions. This kind of on-demand analysis is particularly useful for investigating underexplored subpopulations, such as younger patients with triple-negative breast cancer.

Finally, integrating predictive models into these tools adds another layer of value. Visual platforms that embed statistical or machine learning models, such as Cox proportional hazards or decision tree classifiers, allow users to generate live predictions for survival or recurrence. These tools can simulate "what-if" scenarios, where users manipulate clinical variables and observe the projected outcomes. Additionally, the inclusion of explainable AI techniques, such as SHAP values or interpretable decision paths, enhances transparency and helps users understand the rationale behind predictions.

Altogether, the development of interactive and accessible tools is not just a technical enhancement, it represents a fundamental shift toward democratizing clinical and epidemiologic data in breast cancer research. These tools empower diverse stakeholders, promote equity in research participation, enhance communication between patients and providers, and ultimately support better-informed decisions that can improve outcomes across global populations. As the field matures, visual analytics is poised to play a central role in transforming breast cancer data into actionable knowledge, bridging the gap between raw data and clinical insight, and ultimately enhancing the quality and equity of

cancer care.

## 2.5. A Visual Analytics Approach Using R and Shiny

The R programming language was established in the 90s and became well-known in the world of data science and statistical analysis, particularly in fields that rely heavily on large, complex datasets such as epidemiology and clinical research. As it has an ever-expanding library of packages for almost every requirement and offering extensive capabilities for data manipulation, visualization, and modeling, R has seen a big rise in popularity recently.

Shiny, one of those packages, is relatively new in R's history and enables the creation of responsive dashboards that update dynamically based on user input. It allows the creation of interactive, user-friendly web applications for analyzing, modeling, and visualizing data. The platform's web-based deployment ensures broad accessibility of advanced analytics, even for users without programming knowledge. Moreover, the open-source nature of both R and Shiny facilitates customization, extensibility, and community-driven development.

Key features of the proposed tool include modules for survival analysis, group comparison, risk stratification, and scenario simulation. These functionalities will enable users to explore data interactively, test hypotheses, and evaluate decision-making pathways in a user-friendly environment.

## 2.6. Sum up

Prognosis in breast cancer is shaped by a complex interplay of tumor biology, patient demographics, treatment patterns, and structural inequities. While stage at diagnosis, molecular subtype, and treatment response remain key clinical indicators, their prognostic value is often mediated by socioeconomic and systemic factors. This literature review underscores the necessity of multifaceted approaches that integrate clinical insight with real-world variability. Using the large-scale SEER dataset, this study quantifies the key clinical and demographic factors that predict breast cancer survival, building on existing research. Through detailed data characterization and robust statistical and machine learning modeling, it was aimed to build predictive tools that not only mirror the complexities highlighted in the literature but also offer practical applications. Embedding these insights in an interactive Shiny application is intended to help clinicians and researchers visualize prognostic patterns and explore variable interactions. This work reinforces the

importance of integrating data science into cancer epidemiology, and it offers a scalable framework for more equitable and informed decision-making in breast cancer care.

Methodologically, this project adds to breast cancer prognostic modeling by integrating and extending analytical strategies from recent studies. While prior studies have applied traditional Cox regression models or explored isolated predictive tools, this work combines classical survival analysis with modern machine learning techniques to enhance prognostic accuracy using a larger and more recent SEER dataset (2000-2021). Unlike studies that limit their scope to specific subgroups or single risk factors, this project systematically evaluates a wide set of clinical and demographic variables, applies rigorous multivariable modeling, and benchmarks predictive performance across statistical and machine learning approaches. By doing so, it not only tests the replicability of previously identified prognostic relationships but also improves upon existing predictive frameworks through comparative model evaluation. A methodology that is transparent, scalable, and grounded in statistical rigor and predictive accuracy will add value to the development of more precise, data-driven tools for predicting breast cancer survival.

### 3. Materials and Methods

This chapter describes the methods used to evaluate clinical and demographic factors in breast cancer survival and to develop predictive models for 5-year outcomes. It details the SEER dataset used, the variables selected, and the preprocessing steps applied to ensure data quality and analytical validity while also describing the handling of missing data, data cleaning procedures, and the preparation of datasets for both descriptive analyses and predictive modeling.

#### 3.1. SEER Dataset

We performed the 'case listing' function through SEER\*Stat software (version 8.4.4) to export potential cases for analysis. These comprised a total of 1,294,351 breast cancer patients diagnosed between 2000 and 2021, from the SEER 17 database. The SEER 17 database is a population-based cancer registry dataset maintained by the National Cancer Institute's SEER Program, comprising data from 17 regional cancer registries across the United States [83].

##### 3.1.1 Variables

Variables selected for our analysis, along with their descriptions based on the SEER's variable and recode definitions [84] and corresponding data types, are presented in Table 3.1. Among a long list of available variables, these were the ones identified as relevant for the study objectives and served as the basis for subsequent analyses.

Table 3.1: Variables in the dataset with descriptions and corresponding data types.

| Variable Name                       | Description                                                                                                                                           | Type                    |
|-------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| Age recode with single ages and 90+ | Patient's age at diagnosis. Values range from 0 to 90+, with all ages 90 or older grouped together.                                                   | Quantitative (Discrete) |
| Breast Subtype (2010+)              | Breast cancer molecular subtype classification based on hormone receptor (HR) and HER2 expression profiles. Available for diagnoses from 2010 onward. | Categorical (Nominal)   |
| Chemotherapy recode (yes, no/unk)   | Indicates whether the patient received chemotherapy. Values include 'Yes' and 'No/Unknown,' reflecting treatment administration status.               | Categorical (Binary)    |

Table 3.1 – continued from previous page

| Variable Name                                        | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Type                  |
|------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| <b>COD to site recode</b>                            | Cause of death recoded accordingly to the site in the body.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Categorical (Nominal) |
| <b>Combined Summary Stage (2004+)</b>                | Classifies the extent of tumor spread at the time of diagnosis into standardized categories: 'in situ' (present only in the cells where it originated and has not spread to surrounding tissues), 'localized' (confined to the primary site, but it has spread beyond the original cells), regional (spread to regional lymph nodes or surrounding tissues), or distant (metastasis to distant organs). If the extent of spread cannot be determined, or is benign/borderline, the stage is classified as unknown/unstaged. Cases that cannot be staged for certain year ranges appear as blank(s). This variable is available for cases diagnosed from 2004 onward, based on SEER Summary Stage guidelines. | Categorical (Ordinal) |
| <b>Derived HER2 Recode (2010+)</b>                   | HER2 expression status derived from multiple data sources, categorized as positive, negative, or borderline. Available from 2010 onward.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Categorical (Nominal) |
| <b>ER Status Recode Breast Cancer (1990+)</b>        | Estrogen receptor (ER) expression status of the tumor, coded as positive, negative, or borderline. Available from 1990 onward.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Categorical (Nominal) |
| <b>First malignant primary indicator</b>             | Variable indicating whether the tumor was the first malignant primary neoplasm diagnosed in the patient.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Categorical (Binary)  |
| <b>Marital status at diagnosis</b>                   | Patient's marital status at the time of cancer diagnosis. May be relevant for psychosocial and support-related analyses.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Categorical (Nominal) |
| <b>Median household income inflation adj to 2022</b> | County-level median household income adjusted for inflation to 2022 U.S. dollars. Serves as a proxy for socioeconomic status.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Categorical (Ordinal) |
| <b>PR Status Recode Breast Cancer (1990+)</b>        | Progesterone receptor (PR) expression status of the tumor, coded as positive, negative, or borderline. Available from 1990 onward.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Categorical (Nominal) |

Table 3.1 – continued from previous page

| Variable Name                                                      | Description                                                                                                                                                                                                                                                                                                                                                                                                  | Type                      |
|--------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|
| <b>Primary Site - labeled</b>                                      | Anatomical location of the primary tumor, classified according to SEER site codes. For breast cancer, the code 'C50' designates malignant neoplasms of the breast, with the digit following the decimal point indicating the specific region of origin within the breast (e.g., C50.1: Central portion of the breast). Each code is accompanied by a descriptive label of the corresponding anatomical site. | Categorical (Nominal)     |
| <b>Race and origin re-code (NHW, NHB, NHAIAN, NHAPI, Hispanic)</b> | Race and ethnicity classification emphasizing Hispanic origin versus non-Hispanic, and categorizing Non-Hispanic White (NHW), Non-Hispanic Black (NHB), American Indian/Alaska Native (NHAIAN), and Asian/Pacific Islander (NHAPI).                                                                                                                                                                          | Categorical (Nominal)     |
| <b>Radiation recode</b>                                            | Indicates whether the patient received radiation therapy and, if so, the type administered, as recorded in SEER.                                                                                                                                                                                                                                                                                             | Categorical (Nominal)     |
| <b>Rural-Urban Continuum Code</b>                                  | Classification of patients' residential location based on metropolitan and non-metropolitan status, including gradations of population size and proximity to urban centers.                                                                                                                                                                                                                                  | Categorical (Nominal)     |
| <b>SEER cause-specific death classification</b>                    | Variable denoting patient survival status: alive, deceased due to breast cancer, or deceased due to other causes. Derived from SEER cause-of-death coding.                                                                                                                                                                                                                                                   | Categorical (Nominal)     |
| <b>Sex</b>                                                         | Biological sex of the patient as recorded at diagnosis, categorized as Female or Male.                                                                                                                                                                                                                                                                                                                       | Categorical (Binary)      |
| <b>Survival months</b>                                             | Total number of months the patient survived following the initial diagnosis. Used to assess survival duration.                                                                                                                                                                                                                                                                                               | Quantitative (Continuous) |
| <b>Time from diagnosis to treatment in days recode</b>             | Interval, in days, between the date of initial cancer diagnosis and the initiation of first-line treatment.                                                                                                                                                                                                                                                                                                  | Quantitative (Continuous) |
| <b>Total number of benign/borderline tumors for patient</b>        | Cumulative number of benign and borderline tumors diagnosed in the patient.                                                                                                                                                                                                                                                                                                                                  | Quantitative (Discrete)   |
| <b>Total number of in situ/malignant tumors for patient</b>        | Cumulative count of in situ and malignant tumors identified in the patient across their cancer history.                                                                                                                                                                                                                                                                                                      | Quantitative (Discrete)   |
| <b>Tumor Size Over Time Recode (1988+)</b>                         | Largest recorded diameter of the primary tumor (in millimetres), collected from 1988 onward. Used to assess tumor burden at diagnosis.                                                                                                                                                                                                                                                                       | Quantitative (Continuous) |

Table 3.1 – continued from previous page

| Variable Name                     | Description                                                                                                                        | Type                             |
|-----------------------------------|------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|
| <b>Tumor Size Summary (2016+)</b> | Summary measure of the largest tumor dimension (in millimetres) available from 2016 onward, standardized across SEER data sources. | Quantitative (Continuous)        |
| <b>Year of Diagnosis</b>          | Calendar year in which the initial cancer diagnosis was made, based on SEER registry data.                                         | Quantitative (Discrete, Ordinal) |

## 3.2. Data Preprocessing

Data preprocessing is an indispensable step in data mining as it is able to transform raw data, which is collected from the real world and is usually incomplete, noisy, and contains irrelevant and redundant information or errors, into data that is formatted in an understandable way [85].

### 3.2.1 Variable Missingness

Figure 3.1 summarises the proportion of missing values for each variable in the cleaned SEER dataset. The variable *Tumor Size Summary (2016+)* has the highest proportion of missing data, with approximately 68.8% of records coded as 'Blank(s)'. This is expected, as the variable is only recorded from 2016 onward.

The variable *Breast Subtype (2010+)* exhibits approximately 45.4% missing data, distributed across two categories: 'Recode not available' and 'Unknown'. A similar pattern is observed for *Derived HER2 Recode (2010+)*, which shows nearly 41.1% missing values, primarily recorded as 'Recode not available'. Both variables have been collected only since 2010, which likely contributes to the high proportion of missing data.

*Combined Summary Stage (2004+)* is the fourth variable with the largest proportion of missing values (17.8%), divided between 'Blank(s)' and 'Unknown/unstaged'.

Several variables present smaller amounts of missingness. These include *Time from Diagnosis to Treatment in Days Recode* (10.9%, 'Unable to calculate'), *Tumor Size Over Time Recode (1988+)* (7%, 'Unknown or size unreasonable (includes any tumor sizes 401-989)' and 'Tumor Size Not Consistent Over Time or Not Applicable for this Site'), and *Marital Status at Diagnosis* (4.9%, 'Unknown'). *ER Status Recode Breast Cancer (1990+)* and *PR Status Recode Breast Cancer (1990+)* each have approximately 0.6% missingness ('Recode not available'), *SEER Cause-Specific Death Classification* has roughly

0.5% ('N/A not seq 0-59'), and *Survival Months* has approximately 0.4% ('Unknown'). Finally, *Rural Urban Continuum Code* ('Unknown/missing/no match/Not 1990-2022' and 'Unknown/missing/no match (Alaska or Hawaii - Entire State)'), *Median Household Income (inflation-adjusted to 2022)*, and *Total Number of In Situ Malignant Tumors for Patient*, and all have approximately or less than 0.1% missingness.

The two treatment-related variables, *Radiation Recode* and *Chemotherapy Recode*, contain 'None/Unknown' categories. In both cases, these were retained as valid categories, combining individuals with no treatment and those for whom treatment status is unknown. This approach assumes that any absence of treatment, whether confirmed or uncertain, is analytically similar for the purposes of the present study.

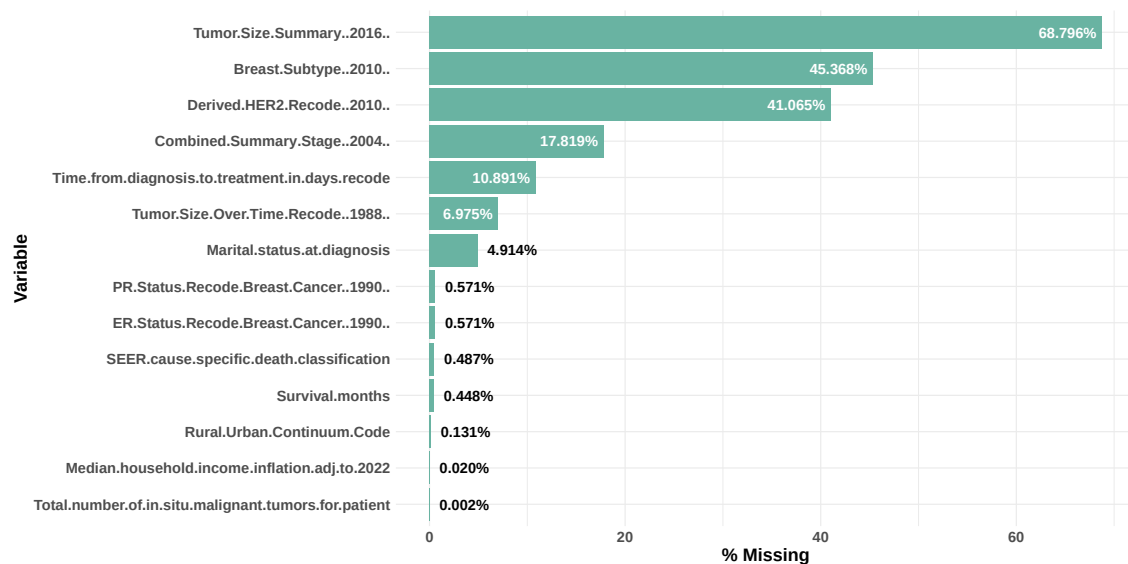


Figure 3.1: Percentage of Missing Data by Variable.

Given the large degree of overlap in meaning and coverage between *Tumor Size Summary* (2016+) and *Tumor Size Over Time Recode* (1988+), and the fact that the former contains data only from 2016 onward, the *Tumor Size Summary* variable was excluded from further analyses.

An additional analysis of the missing data patterns in the dataset, summarized in Table 3.2, reveals that out of the total 1,294,351 rows, 48.24% contain no missing values. Among the remaining rows, 49.62% exhibit a low level of missingness, with up to 10% of data missing. A significantly smaller proportion (2.14%) of rows have missingness between 10% and 20%, while less than 0.01% fall within the 20% to 30% range. Notably, no rows exceed 30% missing data. This indicates that the dataset is largely complete, with over 97% of rows either fully complete or having less than 10% missingness. Rows

with more than 10% missing values represent a negligible fraction (under 3%) and were excluded from analysis without significantly impacting the dataset integrity.

Given that the primary objective of this study is to evaluate 5-year survival, inclusion of cases diagnosed after 2016 would result in incomplete follow-up by the study end date (2021), thereby introducing potential bias in survival estimates. To ensure the validity of the analysis, only cases diagnosed before 2017 were included. This restriction yielded a final analytical dataset comprising 932,726 observations.

Table 3.2: Missingness Distribution.

| Missingness Range | Count of Rows | Percentage |
|-------------------|---------------|------------|
| 0% missing        | 622,879       | 48.12288%  |
| ]0;10] % missing  | 643,285       | 49.69942%  |
| ]10;20] % missing | 28,186        | 2.17762%   |
| ]20;30] % missing | 1             | 0.00008%   |
| ]30;50] % missing | 0             | 0.00000%   |
| >50% missing      | 0             | 0.00000%   |

### 3.2.2 Data Cleaning

The SEER dataset does not represent missing values as 'NA'. Instead, it uses explicit categorical codes such as 'Unknown', 'Not applicable', or 'Unable to determine'. These codes have distinct meanings: some indicate that the information was never collected but could have been, others that the variable is irrelevant for the case, and others that measurement was not possible in real clinical conditions. This design preserves important contextual information and ensures that data handling remains consistent with SEER coding guidelines.

Retaining these coded categories offers several advantages, as it preserves the nuance between different types of absence, avoids unnecessary data loss by preventing the conversion of informative codes to missing values, and maintains compatibility with official documentation, facilitating reproducibility and auditability. At the same time, this approach introduces complications, since statistical models may treat 'Unknown' as a genuine category, potentially biasing predictions, numeric summaries can be distorted if such codes are not excluded, and visualizations may misrepresent the data if these codes are interpreted as standard values.

To balance these considerations, two datasets were maintained in parallel. One dataset preserved SEER’s categorical codes for descriptive and exploratory work. The secondary dataset converted selected codes to ‘NA’ for statistical modeling. For key variables, binary indicator flags were created to capture whether a value represented missingness, allowing investigation of whether the absence of information itself carried predictive value.

With registry data, it is not reasonable to assume that information is missing completely at random (MCAR). MCAR implies that the probability of an observation being missing is unrelated to its value or to the value of any other variables [86]. In the present study, missingness patterns were evaluated for each variable, and appropriate strategies were applied according to their classification as structurally missing, missing at random (MAR), or missing not at random (MNAR).

The handling of missing and categorized SEER codes is summarized below. *Breast Sub-type* and *HER2 Recode* were systematically collected only from 2010 onwards, so analyses involving these variables were limited to cases diagnosed from 2010. Records coded as *Recode not available* (MNAR) were excluded, while *Unknown* values (MAR) were flagged. *Combined Summary Stage*, available from 2004 onwards, excluded *Blank(s)* (MNAR) and flagged *Unknown/unstaged* (MAR). For *ER Status* and *PR Status*, MNAR values were removed, as they occurred in equal numbers for both variables. Variables such as *Marital Status at Diagnosis*, *Median Household Income*, *Time from Diagnosis to Treatment*, and *Total Number of In Situ/Malignant Tumors* contained MAR values, which were flagged. For *Tumor Size Over Time*, cases with ‘Unknown or size unreasonable (401-989 mm)’ were considered MAR and flagged, while ‘Tumor Size Not Consistent Over Time or Not Applicable for this Site’ was MNAR and excluded due to intrinsic links to tumor characteristics, such as multifocal or inflammatory tumors.

Table 3.3 summarizes the transformations applied to each variable in the second dataset, with selected codes recoded as ‘Na’. For the dataset in which the original categories were retained, the only modification involved converting all variables to factors, with consideration given to their nature as binary or ordinal variables.

Table 3.3: Summary of variable transformations during data cleaning.

| Variable                            | Transformation applied                                                                               | Flag created |
|-------------------------------------|------------------------------------------------------------------------------------------------------|--------------|
| Age Recode with Single Ages and 90+ | Removed ‘years’ text; ‘90+’ set to 90 for numeric use (represents $\geq 90$ ); Converted to numeric. | None         |

Table 3.3 – continued from previous page

| Variable                                       | Transformation applied                                                                                                                                                                                                                                                                                                                                                                                                            | Flag created                                      |
|------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| <b>Breast Subtype</b>                          | 'Unknown' set to NA; 'Recode not available' removed; Converted to factor                                                                                                                                                                                                                                                                                                                                                          | subtype_missing_flag<br>subtype_notavailable_flag |
| <b>Chemotherapy Recode</b>                     | Converted to binary factor: Yes(1), No/Unknown(0)                                                                                                                                                                                                                                                                                                                                                                                 | None                                              |
| <b>COD to Site Recode</b>                      | Converted to factor                                                                                                                                                                                                                                                                                                                                                                                                               | None                                              |
| <b>Combined Summary Stage</b>                  | 'Unknown/unstaged' set to NA; Converted to ordered factor ("In situ" < "Localized" < "Regional" < "Distant")                                                                                                                                                                                                                                                                                                                      | stage_missing_flag<br>stage_blanks_flag           |
| <b>Derived HER2 Recode</b>                     | 'Recode not available' removed; Converted to factor                                                                                                                                                                                                                                                                                                                                                                               | HER_notavailable_flag                             |
| <b>ER Status Recode Breast Cancer</b>          | Converted to factor                                                                                                                                                                                                                                                                                                                                                                                                               | ER_notavailable_flag                              |
| <b>First Malignant Primary Indicator</b>       | Converted to binary factor: Yes(1), No(0)                                                                                                                                                                                                                                                                                                                                                                                         | None                                              |
| <b>Marital Status at Diagnosis</b>             | 'Unknown' set to NA; Converted to factor                                                                                                                                                                                                                                                                                                                                                                                          | marital_missing_flag                              |
| <b>Median Household Income (2022 adjusted)</b> | 'Unknown/missing/no match/Not 1990-2022' set to NA; Converted to ordered factor (< \$40,000 < \$40,000 - \$44,999 < \$45,000 - \$49,999 < \$50,000 - \$54,999 < \$55,000 - \$59,999 < \$60,000 - \$64,999 < \$65,000 - \$69,999 < \$70,000 - \$74,999 < \$75,000 - \$79,999 < \$80,000 - \$84,999 < \$85,000 - \$89,999 < \$90,000 - \$94,999 < \$95,000 - \$99,999 < \$100,000 - \$109,999 < \$110,000 - \$119,999 < \$120,000+) | income_missing_flag                               |
| <b>PR Status Recode Breast Cancer</b>          | Converted to factor                                                                                                                                                                                                                                                                                                                                                                                                               | PR_notavailable_flag                              |
| <b>Primary Site - Labeled</b>                  | Converted to factor                                                                                                                                                                                                                                                                                                                                                                                                               | None                                              |
| <b>Race and Origin Recode</b>                  | Converted to factor                                                                                                                                                                                                                                                                                                                                                                                                               | None                                              |
| <b>Radiation Recode</b>                        | Converted to factor                                                                                                                                                                                                                                                                                                                                                                                                               | None                                              |
| <b>Rural Urban Continuum Code</b>              | 'Unknown/missing/no match/Not 1990-2022' set to NA                                                                                                                                                                                                                                                                                                                                                                                | continuum_missing_flag<br>continuum_state_flag    |

Table 3.3 – continued from previous page

| Variable                                                    | Transformation applied                                                                                                                                                                                                    | Flag created                                                     |
|-------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------|
| <b>SEER Cause-Specific Death Classification</b>             | 'N/A not seq 0-59' and 'Dead (missing/unknown COD)' set to NA; Converted to factor                                                                                                                                        | SEERcod_missing_flag                                             |
| <b>Sex</b>                                                  | Converted to factor                                                                                                                                                                                                       | None                                                             |
| <b>Survival Months</b>                                      | Numeric values coerced; 'Unknown' set to NA                                                                                                                                                                               | survival_missing_flag                                            |
| <b>Time from Diagnosis to Treatment in Days Recode</b>      | '731+' set to 731; numeric values coerced; 'Unable to calculate' set to NA                                                                                                                                                | dtc_missing_flag                                                 |
| <b>Total Number of Benign Borderline Tumors for Patient</b> | Converted to numeric                                                                                                                                                                                                      | None                                                             |
| <b>Total Number of In Situ Malignant Tumors for Patient</b> | 'Unknown' set to NA; Numeric values coerced                                                                                                                                                                               | malignant_missing_flag                                           |
| <b>Tumor Size Over Time Recode</b>                          | 998 set to 200 mm; 990 set to 1 mm; numeric values coerced; 'Unknown or size unreasonable (includes any tumor sizes 401-989)' set to NA and 'Tumor Size Not Consistent Over Time or Not Applicable for this Site' removed | tsize_extensive_flag,<br>tsize_micro_flag,<br>tsize_missing_flag |
| <b>Year of Diagnosis</b>                                    | Converted to numeric                                                                                                                                                                                                      | None                                                             |

### 3.2.3 Outcome Variables

Given that one of the main focuses of this analysis is on 5-year survival outcomes, cases with 'Unknown' survival status were excluded, as the outcome for these patients could not be determined and their inclusion could bias results. A binary variable, *survival\_5year*, was then derived from *Survival months*, coded as 0 if the patient survived less than 60 months after diagnosis or died (attributable to breast cancer) and 1 otherwise. This variable facilitates survival analysis and clinically relevant outcome-based comparisons.

To analyse temporal trends, in the *SEER cause-specific death classification* variable, cases coded as 'N/A not seq 0-59' ( $n = 25$ ) were excluded due to indeterminate vital status. A binary variable was then created, coded as 1 if the patient died from breast cancer and 0 otherwise. The latter category included both patients alive at last contact and those who died from causes other than breast cancer.

This restriction yielded a final analytical dataset comprising 915,629 observations.

### 3.2.4 Data Imputation

#### Multiple Imputation by Chained Equations (MICE)

Missing data were addressed using Multiple Imputation by Chained Equations (MICE), implemented in the mice package in R. MICE imputes incomplete variables by iteratively specifying a series of regression models, one for each variable with missing values, and drawing plausible values from the predictive distributions based on the observed data. This process is repeated over multiple iterations until the imputations converge to stable estimates. In the context of survival analysis, particular care must be taken regarding the treatment of outcome variables in the imputation process. In this study, the survival outcome variables, *Survival.months* (time-to-event) and *SEER.cause.specific.death.classification* (event indicator), were not imputed. Imputing these variables could artificially alter event times or misclassify events, thereby introducing bias into the estimation of survival probabilities and hazard ratios. Nonetheless, both outcome variables were retained as predictors in the imputation models for other variables. This approach preserves the associations between covariates and the survival outcomes, which is recommended in the methodological literature to avoid attenuation of effect estimates that may occur when outcomes are excluded from the imputation model. By allowing the outcomes to inform the imputation of covariates, the method ensures that relationships inherent in the complete data are reflected in the imputed values. All other variables with missing values were included as candidates for imputation. It was used for MAR values to preserve statistical power, reduce bias, and account for uncertainty introduced by imputation [87]. Continuous variables were imputed using predictive mean matching (PMM), while categorical variables were imputed using logistic regression for binary outcomes or polytomous regression for nominal variables, as determined automatically by mice. This approach follows the recommendations of [88] and [89], who emphasize that in analyses where the outcome is not imputed, it should still be included as a predictor in the imputation model to preserve important associations and avoid bias in estimated effects. Such an approach is particularly important in survival analysis, where the relationship between covariates and event outcomes is central to the research question.

### 3.2.5 Derived Variables

Several continuous variables were transformed into categorical form for exploratory and bivariate analyses. Importantly, the original continuous versions of these variables were

retained to preserve data integrity and support further modeling. Categorization of continuous variables will be avoided if there is no strong justification, since it can reduce statistical power and obscure nonlinear effects [90]. In cases where categorization was chosen, it was based on clinical or historical rationale, not arbitrary statistical cut-points. [91]

The *Age* variable was categorized into 5-year bins, since the SEER program and many cancer registries use predefined 5-year age groups (for example, 0-4, 5-9, ..., 85-89, 90+) as the basis for age standardization, facilitating consistent comparison of survival rates across different registries, populations, and studies. It is also known that age is a strong confounder in breast cancer survival analysis and because of that, using 5-year age bins allows to control precisely for age-related differences, improving the validity of comparisons between subgroups and ensuring that observed differences in survival are not due to age distribution alone [92].

It is not possible to categorize a tumor's stage considering only its size; it is standard procedure to take into account the TNM system, which classifies tumors by size (T), lymph node involvement (N), and metastasis (M) [93]. However, according to the TNM staging system's tumor size, 'T0' indicates that there is no evidence of primary tumor, 'Tis' is a carcinoma in situ (noninvasive), 'T1' is a tumor with 20 mm or less in greatest dimension, and can be subdivided into 4 different categories. 'T1mi' is a tumor with 1 mm or less, 'T1a' between 1 and 5 mm, 'T1b' between 5 and 10 mm and, lastly, for it to be considered a 'T1c' tumor it has to be greater than 10 mm and smaller than 20 mm. Tumors with more than 20 mm but not more than 50 mm are 'T2', and 'T3' is the stage for tumors with more than 50 mm in the greatest dimension. Finally, T4 is a stage that is not necessarily categorized by tumor size, since it can have any dimension, but rather by a tumor of any size with direct extension to chest wall and/or skin (including inflammatory breast cancer) [94]. It is then important to remember that tumor size alone does not determine the complete stage and that lymph node status and spread are critical for full staging. According to the SEER classification system, this variable includes four additional categories beyond the numerical values representing tumor size: '998 (site-specific code)', '990 (microscopic focus)', 'Unknown or Size Unreasonable (includes any tumor sizes 401-989)', and 'Tumor Size Not Consistent Over Time or Not Applicable for this Site'. The specific definitions and implications of these categories will be detailed further in this work. Accordingly, the tumor size thresholds selected for categorization of the *Tumor Size Over Time* variable were defined as follows: less than or equal to 1 mm, greater than 1 mm and less than or

equal to 5 mm, greater than 5 mm and less than or equal to 10 mm, greater than 10 mm and less than or equal to 20 mm, between 20 mm and 50 mm, and greater than 50 mm. Additionally, the four supplementary categories present in the dataset were retained with their original designations.

To categorize *Year of Diagnosis* in the SEER breast cancer dataset for survival analysis, it is common practice to group years into intervals ranging from 2 to 5 years. This approach balances detail with sufficient sample sizes per group. For this analysis, 3-year intervals were chosen to capture temporal trends in medical practice, treatment advances, and diagnostic techniques while maintaining meaningful group sizes. Specifically, the years 2018 to 2021 were grouped together to ensure adequate representation in the most recent time period.

Following clinical relevance and established research standards, the binning approach for *Time from Diagnosis to Treatment* was informed by commonly used classifications in breast cancer literature and clinical guidelines. Intervals of 0 to 30 days are considered immediate treatment, representing optimal care in line with recommendations for rapid initiation. A period of 31 to 60 days may be regarded as a short delay, but is often still considered timely. Delays of 61 to 90 days are classified as moderate, reflecting care that is acceptable but postponed. Delays exceeding 90 days are consistently associated with worse outcomes for invasive breast cancer [95, 96]. In this dataset, the variable contains both numeric values ranging from 0 to 730 days (equivalent to 24 months) and special codes such as '731+ days' and 'Unable to Calculate'. As the majority of cases occur within 100 days of diagnosis, and to balance the distribution of observations across categories while remaining consistent with one-month intervals reported in the literature, the final classification adopted was: 0-30 days, 31-60 days, 61-90 days, 91-120 days, more than 120 days, and 'Unable to Calculate'.

*Total number of benign/borderline tumors for patient* and *Total number of in situ/malignant tumors for patient* are both discrete numeric variables. However, when categorization is required for certain statistical analyses, these variables can be treated as categorical due to their limited number of unique values. For instance, the variable *Total number of benign/borderline tumors for patient* takes on only five distinct values: 0, 1, 2, 3, and 4. In such cases, each unique value can be considered a separate category and, since the values are inherently ordered ( $0 < 1 < 2$  etc.), it was categorized as ordinal. Following the same reasoning, the variable *Total number of in situ/malignant tumors for patient* was

also categorized as ordinal, with the addition of an “Unknown” category to account for cases where this information was missing or not recorded in the dataset.

This transformation step was essential because many of the statistical methods employed in the following sections, such as contingency tables, the chi-square test, Kaplan-Meier survival curves, and the log-rank test, are designed to work with categorical groupings. The previously mentioned numerical variables were then categorized into clinically meaningful or statistically practical ranges in order to enable the valid application of these methods.

### 3.3. Descriptive Analysis

In the initial phase of the analysis, all variables previously identified directly from the SEER breast cancer dataset underwent univariate analysis to explore their distribution across the study population. This exploratory step was conducted using *R* and several key packages, including *dplyr*, *ggplot2*, *stringr*, *ggrepel*, and *wordcloud*. The dataset was examined in its raw, unprocessed form to observe its natural structure and to identify any underlying patterns or anomalies prior to any transformation or cleaning.

Descriptive statistical functions such as *summarize()*, *mean()*, and *sd()* were applied first to the quantitative variables in order to calculate measures of central tendency and dispersion, including the mean, median, standard deviation, and interquartile range. For categorical variables, frequency tables were created to display both absolute and relative frequencies, thereby offering a clear picture of the prevalence of each category within the dataset.

Following these calculations, a range of graphical visualizations was produced to represent the data in a more intuitive and interpretable manner. Pie charts, bar plots, histograms, density plots, and a word cloud were generated to visually depict the statistical findings, facilitating a clearer and more accessible understanding of the distribution of each variable.

This structured univariate exploration worked as a foundation for the more advanced analytical procedures that followed, ensuring a solid grasp of the dataset's basic properties before engaging in further modeling and inferential statistics.

### 3.4. Traditional survival/statistical analysis

Continuous variables were not used directly and, instead, their categorized versions, created as described previously, were employed to enable appropriate comparison using categorical methods. This approach facilitated the identification of relevant relationships to inform feature selection and guide further modeling within the SEER breast cancer dataset.

For analyses of 5-year survival proportions (contingency tables, chi-square), only patients diagnosed between 2000 and 2016 were included to ensure complete follow-up. For time-to-event analyses (Kaplan-Meier, Cox regression), all patients diagnosed between 2000 and 2021 were included, with right-censoring applied to those with shorter follow-up.

#### 3.4.1 Contingency Tables

Contingency tables were constructed to summarize the frequency distributions of two or more categorical variables, enabling the assessment of their joint distributions. These tables facilitate the identification of potential associations, such as correlations between cancer subtypes and demographic characteristics. When combined with inferential tests, contingency tables support subgroup analyses and hypothesis testing relevant to clinical decision-making [97]. The primary objective of this step is to evaluate whether each variable has a significant influence on 5-year survival.

#### 3.4.2 Chi-square Test of Independence

The Chi-square test of independence was employed to evaluate whether relationships exist between pairs of categorical variables. The null hypothesis assumes no association between variables, while the alternative hypothesis posits a relationship. An underlying assumption is that expected cell frequencies exceed five since this test relies on large-sample approximations that become invalid with small expected cell counts, leading to inaccurate p-values. The test aids in identifying variables with significant associations to cancer outcomes, thereby guiding subsequent analyses [98].

#### 3.4.3 Kaplan-Meier Survival Analysis

Survival probabilities over time were estimated using the Kaplan-Meier method, a non-parametric approach suitable for time-to-event data without requiring distributional assumptions. Survival curves represent the estimated probability of event-free survival on

the y-axis over time on the x-axis, with stepwise decreases at observed event times [99]. Censored observations, indicating incomplete follow-up or absence of events by study end, were appropriately accounted for and marked on the curves. Median survival time was derived from the point where the survival probability equals 0.5, and comparative survival analyses were conducted by plotting Kaplan-Meier survival curves for individual variables.

#### 3.4.4 Log-rank Test

To statistically compare survival distributions between groups, the log-rank test was applied. This nonparametric test evaluates the null hypothesis of equal survival functions by comparing observed versus expected event counts across groups at each event time, accounting for censored data. The test statistic approximates a chi-square distribution under the null hypothesis. The log-rank test assumes proportional hazards; violations of this assumption, such as crossing survival curves, reduce its sensitivity and warrant alternative methods (for example, the Wilcoxon test or Cox proportional hazards modeling). Together with Kaplan-Meier curves, the log-rank test provides a robust framework for the analysis of time-to-event outcomes [100].

#### 3.4.5 Cox Proportional Hazards Models

The Cox Proportional Hazards model is a semiparametric regression method widely used in survival analysis to investigate the effect of multiple covariates on the hazard, or risk, of death over time. Unlike univariate methods such as Kaplan-Meier analysis or chi-square tests, this approach allows simultaneous adjustment for confounding variables, providing a clearer understanding of independent prognostic factors. In clinical research, where multiple patient characteristics can influence survival, multivariable modeling is essential for accurate inference. The variable selection process was guided by both statistical and clinical criteria. Initially, variables that demonstrated statistically significant associations with 5-year survival in chi-square tests, or with overall survival in Kaplan-Meier log-rank tests, and variables with a  $p$ -value less than 0.05 in these analyses, were considered for inclusion. However, statistical significance alone was not the sole criterion and clinically important variables were retained regardless of univariate results to avoid excluding critical predictors due to sample variability.

Given the dataset contained 187,442 events (deaths), statistical power was more than adequate to include multiple covariates in the model. Although the theoretical upper bound

for the number of predictors (based on the common rule of one variable per 10 events) was high, a more conservative and interpretable approach was adopted. The modeling strategy aimed to include approximately 15 to 25 covariates, balancing comprehensiveness with interpretability and minimising the risk of overfitting. In this study, all variables showed statistical significance in bivariate analysis and, therefore, a Cox regression was applied to evaluate clinical and demographic factors associated with overall survival in breast cancer patients, using all variables that underwent the bivariate analysis. In the multivariable model, numeric variables were included in their continuous form, as opposed to the categorical transformations used in the bivariate analysis. Following variable selection, a multivariable Cox regression model was fitted and the proportional hazards assumption was carefully assessed.

To address the structural missingness of both *Breast Subtype* and *HER2 Recode* variables, two complementary Cox proportional hazards models were developed. The first model (Model A) included all cases across the entire study period and only variables available consistently from the earliest diagnosis years. This maximised the sample size and follow-up duration, providing robust estimates of long-term survival patterns, though without molecular stratification. The second model (Model B) was restricted to cases diagnosed from 2010 onwards, allowing inclusion of *Breast Subtype* and *HER2 Recode* alongside other predictors. While this model incorporated detailed molecular data and reflected contemporary clinical practice, it involved a smaller cohort and shorter follow-up, which could result in wider confidence intervals (CI) and less stable effect estimates.

By fitting both models, hazard ratios for variables present in both analyses could be directly compared to assess consistency across historical and contemporary cohorts. Sensitivity analyses were performed to evaluate whether differences in effect sizes were attributable to the inclusion of molecular predictors or to temporal changes in patient characteristics and treatment practices. This dual-model approach ensured transparency in handling structural missingness while balancing the trade-off between maximising longitudinal data coverage and incorporating clinically relevant molecular information in recent cohorts. The final multivariable models provided adjusted hazard ratios with 95% confidence intervals for each covariate, offering robust evidence for understanding survival dynamics in breast cancer patients.

### 3.5. Machine Learning Models

To compare traditional statistical approaches with predictive models, supervised machine learning algorithms were developed to predict 5-year survival as a binary outcome.

#### 3.5.1 Data Preparation

The outcome variable was defined as a binary indicator of whether a patient survived at least 60 months (5 years). All predictors were included in the analysis, as they are both clinically relevant and previous exploratory results demonstrated statistical significance. Categorical variables were encoded using one-hot encoding to ensure consistency across models. Although Random Forest and XGBoost can natively handle categorical variables, the same encoding was applied for comparability. Numerical predictors were standardized to zero mean and unit variance, as Logistic Regression is sensitive to feature scaling. Although Random Forest and XGBoost are not affected by variable scaling, standardization was applied uniformly across all models to maintain methodological consistency.

#### 3.5.2 Train/Test Split

The dataset was randomly divided into training (80%) and test (20%) subsets. To preserve the proportion of patients surviving at least 60 months across both subsets, stratified sampling was applied based on the outcome variable. The training set was used for model development, including hyperparameter tuning, while the test set was held out for final model evaluation.

#### 3.5.3 Model Selection

XBoost is, as backed up by several studies, a top-performing algorithm in structured clinical datasets like SEER, however a diverse mix of machine learning models will be applied, offering different strengths. Logistic Regression was the first to be applied since it is a simple and interpretable model for binary classification. This model will, hopefully, help to understand the additive effects of predictors and compare ML model performance against the traditional statistical classifiers applied earlier. Random Forest, known for handling nonlinearities and interactions well, is also robust to outliers and missing values. Given this, it is expected to provide good accuracy and variable importance metrics, useful for explainability. Lastly, Gradient Boosting Machines (GBMs), such as XGBoost, was applied. These are powerful ensemble methods based on decision trees that usually

achieve top results in structured datasets and efficiently handle missing values, outliers and class imbalance.

#### 3.5.4 Handling Class Imbalance

The distribution of the outcome variable, five-year survival, was moderately imbalanced, with a greater proportion of patients surviving beyond 60 months compared to those who did not. Class imbalance can adversely affect model performance by biasing predictions toward the majority class and, therefore, tailored strategies were implemented to mitigate this issue according to the requirements of each algorithm.

For Logistic Regression, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE generates synthetic instances of the minority class by interpolating between existing minority observations, thereby reducing the risk of overfitting associated with random duplication. The resampled training set was subsequently used to fit the model, with Logistic Regression specified under a binomial family.

For the Random Forest model, data-level resampling was not applied. Instead, misclassification costs were incorporated by assigning higher weights to minority class observations, enhancing the model's ability to identify minority outcomes without altering the original data distribution.

For XGBoost, class imbalance was addressed using the 'scale\_pos\_weight' parameter, defined as the ratio of negative to positive class frequencies. This adjustment modifies the gradient updates during training to increase the emphasis placed on the minority class. The model was trained using a logistic objective function, with performance monitored through the area under the receiver operating characteristic curve (ROC AUC).

#### 3.5.5 Evaluation Metrics

Model performance was assessed using multiple complementary metrics derived from the confusion matrix, including accuracy, sensitivity, specificity, and G-mean, providing a comprehensive evaluation of both overall correctness and the ability to correctly identify positive and negative outcomes.

Receiver operating characteristic (ROC) curves were plotted for each model, and the area under the curve (ROC AUC) was calculated as the primary measure of discrimination. Additionally, precision-recall (PR) curves and the corresponding Precision-Recall Area

Under the Curve (PR AUC) were computed to evaluate model performance in the context of imbalanced classes. The Youden index was applied to determine the optimal threshold for classification.

Feature importance was assessed to identify the most influential predictors contributing to survival outcomes, allowing interpretation of the relative impact of clinical and demographic variables. Finally, the predictive performance of the best-performing machine learning model was compared with the Cox Proportional Hazards model to assess the potential improvements offered by machine learning approaches over traditional statistical methods in breast cancer prognosis.

### 3.6. Shiny App

An interactive web-based application was developed using the Shiny framework [101] in R to enable visualization of the SEER breast cancer dataset, statistical analyses, and predictive model outputs.

The full source code is available in the project repository: [GitHub Repository](#)

### 3.7. Implementation Details

#### Hardware

All tests were run on a system running a 64-bit operating system with an Intel® Core™ i3-1005G1 CPU @ 1.20 GHz, 8 GB RAM, and integrated Intel UHD Graphics.

#### Software and Libraries

All analyses were performed using R version 4.4.0 [102], with the following packages:

- Data handling and preprocessing: *readxl*, *dplyr*, *stringr*, *purrr*, *car*
- Data visualization: *ggplot2*, *ggrepel*, *wordcloud*, *tm*
- Missing data imputation: *mice*
- Classical statistics: *survival*, *survminer*
- Machine learning: *caret*, *randomForest*, *smotefamily*, *xgboost*, *e1071*, *ranger*
- Model evaluation and interpretation: *pROC*, *PRROC*

## 4. Results

### 4.1. Univariate Analysis

Codes and definitions follow the official SEER codebooks.

#### 4.1.1 Age

In SEER, the “Age Recode” variable (single ages and 90+) categorizes individuals by each year of age from infancy through 89, with a combined category for those aged 90 years and older.

As shown in Figure 4.1, *Age at Diagnosis* in the study cohort spans from birth to 90 years; the latter including all individuals aged 90 or above. The average age at diagnosis was 61.9 years (SD: 13.8), with a median of 62 years, a mode of 65 years, and an interquartile range of 52-72 years. The distribution is mildly right-skewed, with the majority of cases clustering between ages 50 and 75. Notably, a sharp rise in frequency occurs at age 90, reflecting the dataset’s top-coding of this group.

Figure 4.2 presents the same data as in Figure 4.1, but grouped into categorical intervals rather than displayed as a continuous ordinal variable. Ages were aggregated into five-year bins, with the ‘00 years’ and ‘90+ years’ categories retained as separate groups. As this is a reclassification of the same variable, the summary statistics remain unchanged from those previously reported.

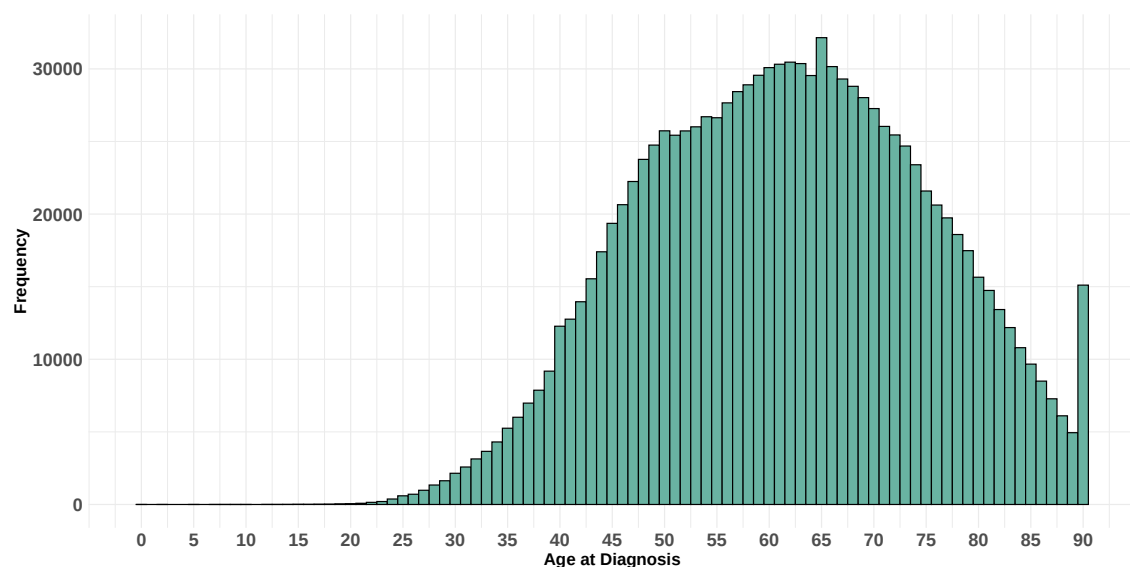


Figure 4.1: Distribution of Age at Diagnosis in Breast Cancer Cases (SEER, 2000-2021).

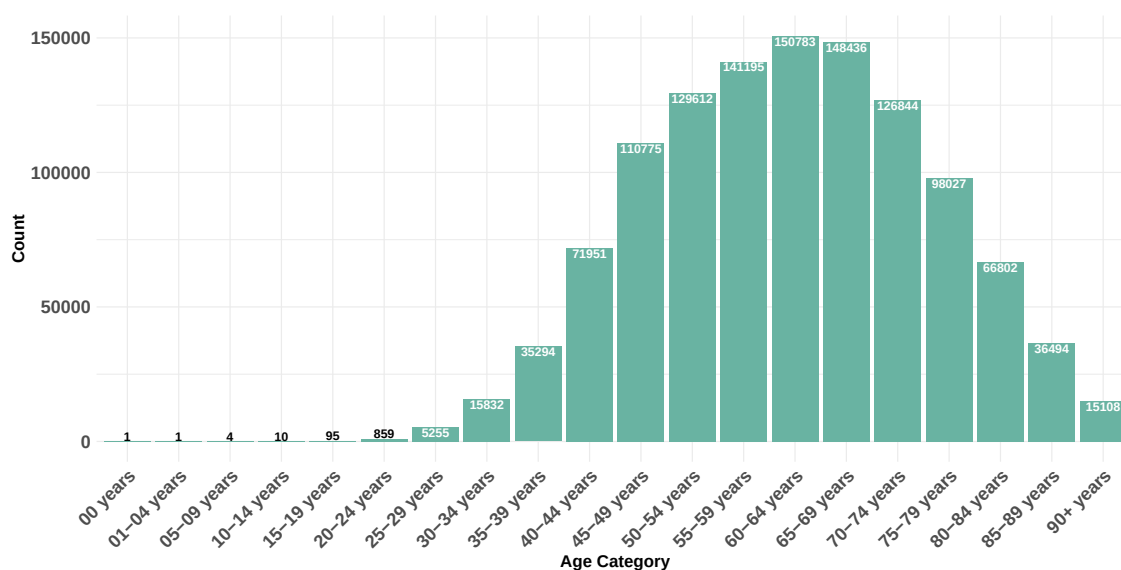


Figure 4.2: Distribution of Age at Diagnosis as a categorical variable with 5-year bins in Breast Cancer Cases (SEER, 2000-2021).

#### 4.1.2 Breast Subtype

Since 2010, SEER has collected data on human epidermal growth factor receptor 2 (HER2) status for breast cancer cases. Breast cancer molecular subtypes are defined by the joint assessment of hormone receptor (HR: estrogen receptor [ER] and progesterone receptor [PR]) status and HER2 status. To facilitate the analysis of trends in these subtypes, SEER\*Stat includes the *Breast Subtype (2010+)* variable [103].

In this variable, as seen in Figure 4.3 'Recode not available' (531,528 cases) indicates that the cancer was diagnosed prior to 2010, before HER2 status was recorded. The 'HER2+/HR+' category, with 73,318 recorded cases, includes tumors with positive HER2 expression and at least one positive hormone receptor (ER or PR). 'HER2+/HR-' includes tumors that are HER2 positive but negative for both ER and PR, and represents 30,604 cases of this dataset. Conversely, 'HER2-/HR+' (the most prominent category in the dataset, with 526,452 cases) and 'HER2-/HR-' (with a frequency of 76,759 and representing triple-negative breast cancer) indicate HER2 negativity, with the first group having at least one positive hormone receptor and the second group negative for both. Cases with insufficient information for classification are recorded as 'Unknown'.

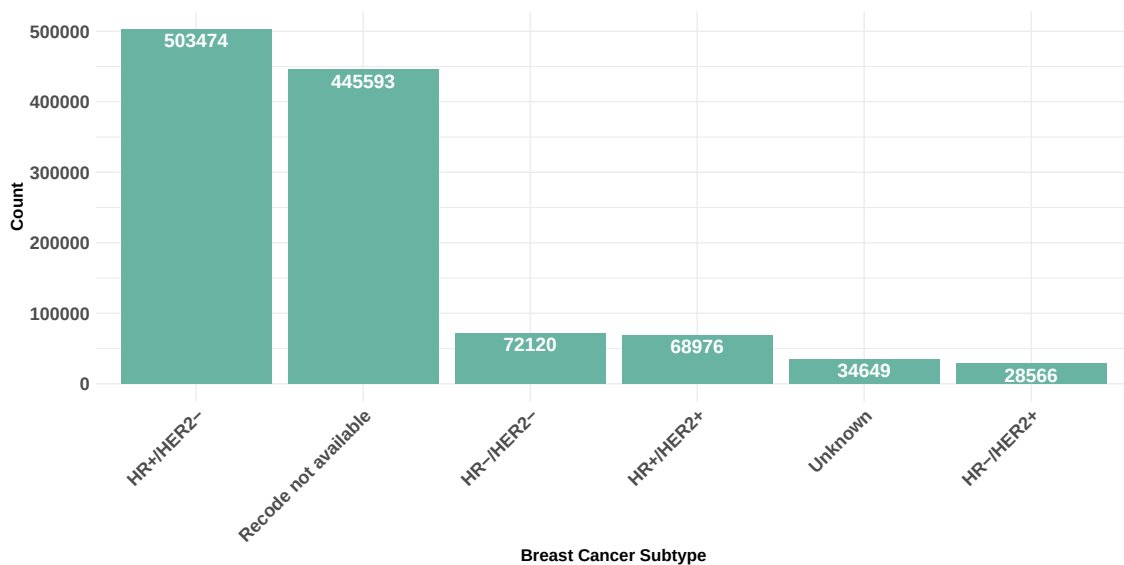


Figure 4.3: Distribution of Breast Cancer Subtypes (SEER, 2000-2021). The bar plot displays the frequencies of breast cancer cases across molecular subtypes defined by hormone receptor (estrogen and progesterone) status and HER2 (human epidermal growth factor receptor 2) expression.

#### 4.1.3 Chemotherapy

In this cohort, 493,933 patients (38%) were recorded as having received chemotherapy, while 800,418 (62%) were classified as 'No/Unknown' (Figure 4.4). This latter category combines patients who did not receive chemotherapy with those for whom treatment information was missing or unrecorded, limiting the ability to distinguish between treatment avoidance and incomplete data capture.

The *Chemotherapy Recode* variable, a simplified binary-like classification commonly used in large clinical datasets, reflects whether chemotherapy was part of a patient's treatment. However, SEER acknowledges a key limitation: for cases without recorded treatment, it is unclear whether chemotherapy was truly not given or whether it was missed in the registry. Moreover, interpretation of the 'No/Unknown' group requires caution, as it merges two clinically distinct subpopulations with potentially different prognoses. Most importantly, survival differences between patients who did and did not receive chemotherapy cannot be attributed solely to treatment efficacy or effectiveness without accounting for the clinical and demographic factors influencing treatment receipt. Similarly, these differences cannot be generalized to estimate the benefit an individual patient might expect from chemotherapy [104].

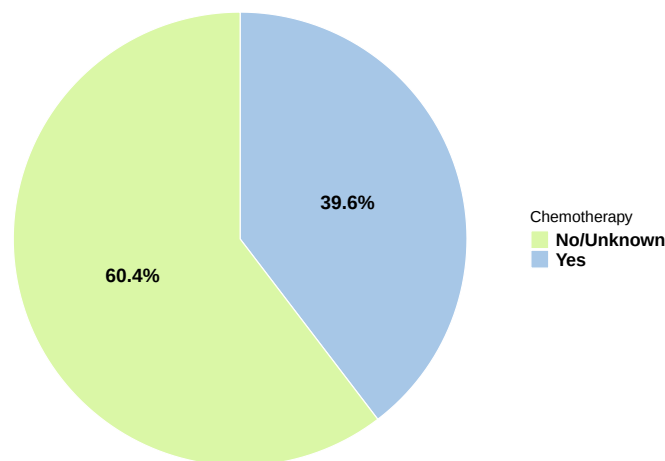
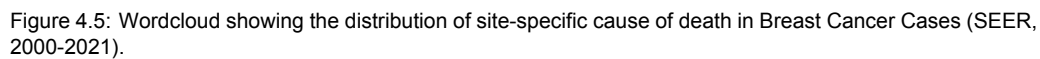


Figure 4.4: Distribution of Chemotherapy Status in Breast Cancer Cases (SEER, 2000-2021). The pie chart illustrates the proportion of breast cancer patients who received chemotherapy versus those who did not or for whom the status is unknown, as recorded in the SEER database.

4.1.4 COD to Site

As shown in Figure 4.5, the most frequently recorded value for the variable *COD to Site Recode* is 'Alive', indicating patients who were alive at last contact. The second most prevalent category, highlighted in red, is 'Breast', representing patients whose recorded cause of death was breast cancer, confirming that this remains a leading cause of mortality within this cohort, consistent with its clinical significance. All other values represent deaths attributed to causes other than breast cancer, including other neoplasms and non-neoplastic diseases. Highlighting the heterogeneity of mortality causes in the cohort, and underscoring that patients diagnosed with breast cancer may die from a range of competing factors is an important consideration to have in survival analyses, as they may influence overall prognosis and the interpretation of treatment effects.



The variable *Combined Summary Stage (2004+)* categorizes the extent of disease at diagnosis, indicating how far the cancer had spread at the time it was first identified. Developed for standardized surveillance and comparison purposes, this staging framework classifies tumors into the following broad groups: 'In situ', 'Localized', 'Regional', 'Distant', 'Unknown/unstaged', and 'Blank(s)'. The classification is based on the SEER staging manual [105], which incorporates definitions from the AJCC Cancer Staging Manual [106].

The second most frequent category was 'Regional' stage (n = 284,972), representing cancers that had spread to nearby lymph nodes or surrounding tissues but not to distant organs. These cases typically require more aggressive treatment and carry a moderately worse prognosis compared to localized disease.

Cancers diagnosed at the 'Distant' stage (n = 49,226), those that had metastasized to remote organs, comprised a smaller portion of the dataset. Although less common, this

category is associated with the poorest prognosis and reflects advanced disease at presentation, highlighting the continuing challenge of detecting breast cancer before systemic spread occurs.

Only three cases were classified as 'In situ', where malignant cells remain confined to the ducts or lobules without invasion into surrounding tissue. This finding is notably lower than expected; in 2009, approximately 20% of new breast cancer diagnoses in the United States were in situ [107]. This discrepancy likely reflects either under-reporting, exclusion during preprocessing, or the dataset's focus on invasive cancers. SEER and many cancer registries often report in situ breast carcinoma separately [107]. Prognosis for in situ cases is excellent, with very low risks of recurrence or progression [108], making them the ideal stage for early detection initiatives.

A substantial proportion of the dataset lacked definitive stage information: 168,088 cases were labelled 'Blank(s)', and 4,548 were 'Unknown/unstaged'. 'Unknown/unstaged' category captures cases where staging could not be determined and, excluding them from analyses, may lead to biased results, so they are often tracked separately in cancer registry studies as a measure of data quality [109]. However, 'Blank(s)', typically correspond to years or circumstances where stage data were not collected, are, in most analyses using SEER data, excluded from stage-specific frequency calculations, as it does not add meaningful staging information [110].

In summary, the distribution of *Combined Summary Stage (2004+)* in this dataset indicates that most breast cancer cases were diagnosed at a localized stage, offering a favourable outlook for many patients. However, the considerable number of regional and distant diagnoses, along with the large proportion of missing stage data, underscores the need for improvements in both early detection strategies and the completeness of cancer registry records.

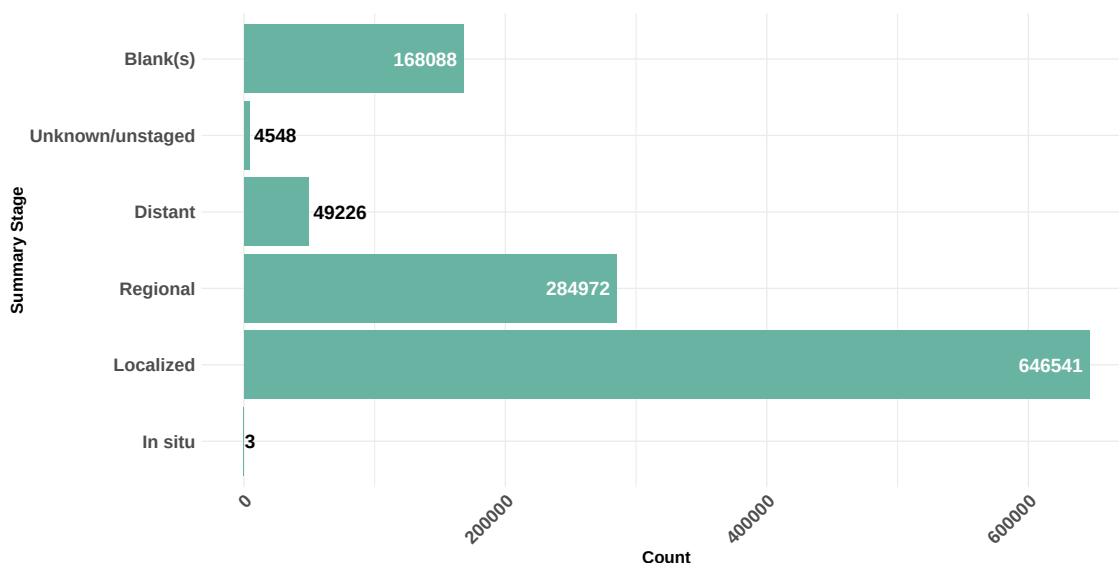


Figure 4.6: Distribution of Stage at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the number of breast cancer cases by stage at diagnosis, as recorded in the SEER database.

#### 4.1.6 Derived HER2

The SEER *Derived HER2 Recode (2010+)* consolidates HER2 status from multiple registry sources to support consistent longitudinal analysis. In this cohort, as depicted in Figure 4.7, Negative cases dominate (604,212), followed by Recode not available (531,528), Positive (104,299), and Borderline/Unknown (54,312).

Similar to what was seen in the *Breast Subtype (2010+)*, cases with 'Recode not available' reflect pre-2010 diagnoses, prior to routine HER2 data capture. The 'Borderline/Unknown' group reflects diagnostic uncertainty or missing data; SEER's harmonization efforts post-2018 combine these codes because of evolving assay interpretation protocols. The data handling and coding rules in SEER suggest that borderline HER2 cases should be carefully identified and managed as a separate group, potentially flagged for further diagnostic work to clarify HER2 expression before making treatment decisions [111].

The predominance of HER2-negative tumors in this dataset aligns with known breast cancer epidemiology, while the substantial number of HER2-positive cases emphasizes the importance of HER2-targeted treatments. The relatively high frequency of cases with indeterminate or missing HER2 information highlights limitations in data completeness and potential challenges for analyses involving HER2 status. Given HER2's pivotal role in prognosis and therapeutic decision-making, accurate and comprehensive HER2 classification is essential for reliable epidemiological research and for evaluating trends in treatment outcomes.

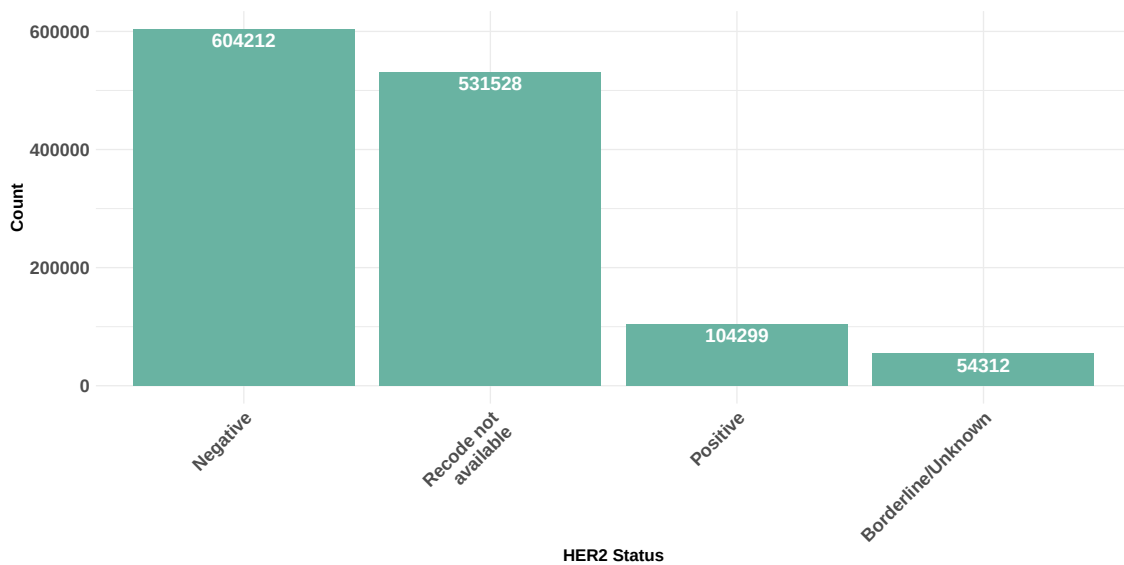


Figure 4.7: Distribution of Derived HER2 at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the status of HER2 at diagnosis, as recorded in the SEER database.

#### 4.1.7 ER Status

The ER status (Estrogen Receptor status) in breast cancer refers to whether the cancer cells have receptors for the hormone estrogen. It is a key biomarker used to classify breast cancer types and guide treatment decisions. As observed in Figure 4.8 the *ER Status* variable is categorized, from most to least frequent, as follows: ‘Positive’ (n = 974,774) identifies tumors with estrogen receptor expression, suggesting potential responsiveness to endocrine therapies; ‘Negative’ (n = 215,518) includes tumors lacking ER expression, which are less likely to benefit from hormone-based treatments. Cases labeled ‘Borderline/Unknown’ (96,673) represent instances where ER testing was performed but yielded inconclusive or borderline results, flagging uncertainty in receptor status that may require cautious clinical interpretation. Conversely, ‘Recode Not Available’ (7,386 cases) indicates that no ER data were recorded or available for recoding, reflecting missing information rather than equivocal test results. This distinction is important for analytic approaches, as borderline/unknown cases reflect ambiguous biomarker expression that may influence treatment, whereas recode not available cases represent absent data and pose challenges for classification and outcome analyses [86].

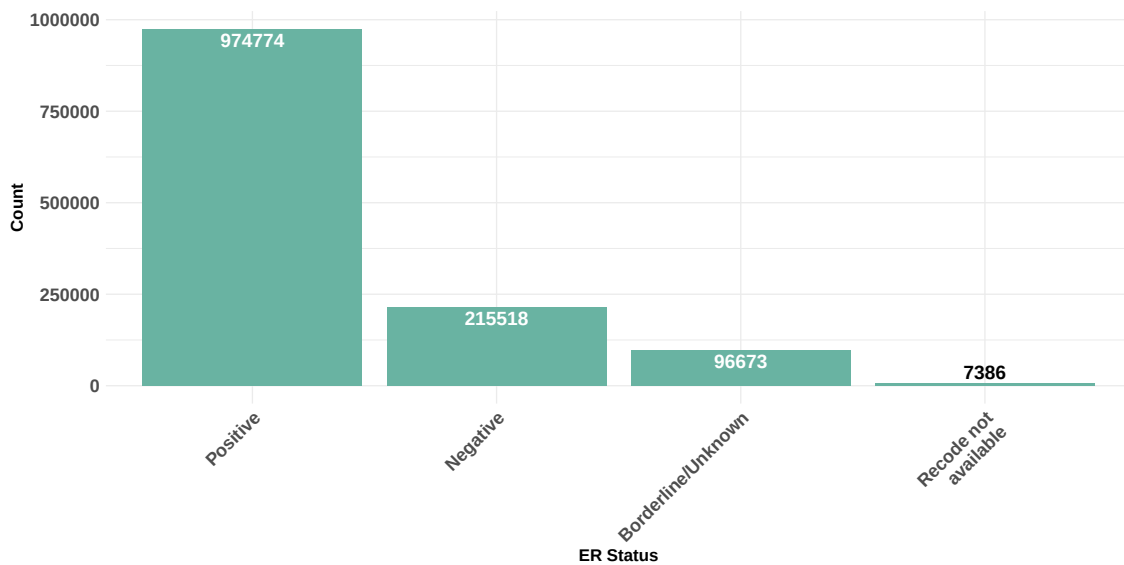


Figure 4.8: Distribution of ER Status at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the status of ER at diagnosis, as recorded in the SEER database.

#### 4.1.8 First Malignant Primary Indicator

The *First Malignant Primary Indicator* in the SEER breast cancer dataset identifies whether the cancer record corresponds to the patient's first malignant primary cancer in the registry. As shown in Figure 4.9, the majority of cases (84.3%) are marked 'Yes', indicating that this is the patient's initial primary malignancy recorded. The remaining 15.7% of cases are labeled 'No', representing subsequent primary cancers diagnosed in the same individual.

This variable is important for distinguishing incident primary cancers from recurrences or second primaries, which can influence prognosis and treatment decisions.

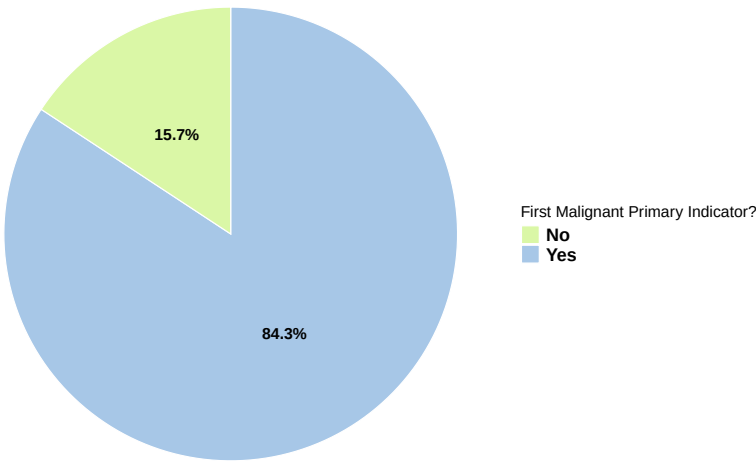


Figure 4.9: Distribution of First Malignant Primary Tumor Status in Breast Cancer Cases (SEER, 2000-2021). The pie chart illustrates whether the recorded breast tumor was the patient's first malignant primary tumor, as documented in the SEER database. This indicator helps estimate cancer incidence by identifying initial malignancies, even if prior cancers are not recorded in the registry.

4.1.9 Marital Status at Diagnosis

Figure 4.10 displays the distribution of *Marital Status at Diagnosis* among patients in the dataset. The most common category is 'Married (including common law)', with 636,173 cases. This is followed by 'Widowed' (168,745 cases), 'Single (never married)' (162,939), 'Divorced' (121,511), 'Unknown' (48,351), 'Separated' (12,193), and the smallest group, 'Unmarried or Domestic partner' (3,466 cases).

Marital status is recorded as a categorical variable in the SEER database and has been studied as a potential social determinant influencing cancer outcomes. Previous research has suggested that married patients often experience better survival rates compared to unmarried groups, potentially due to greater social support, better access to care, or adherence to treatment. Thus, analysing marital status in relation to cancer prognosis can provide important insights into the role of social factors in patient outcomes.

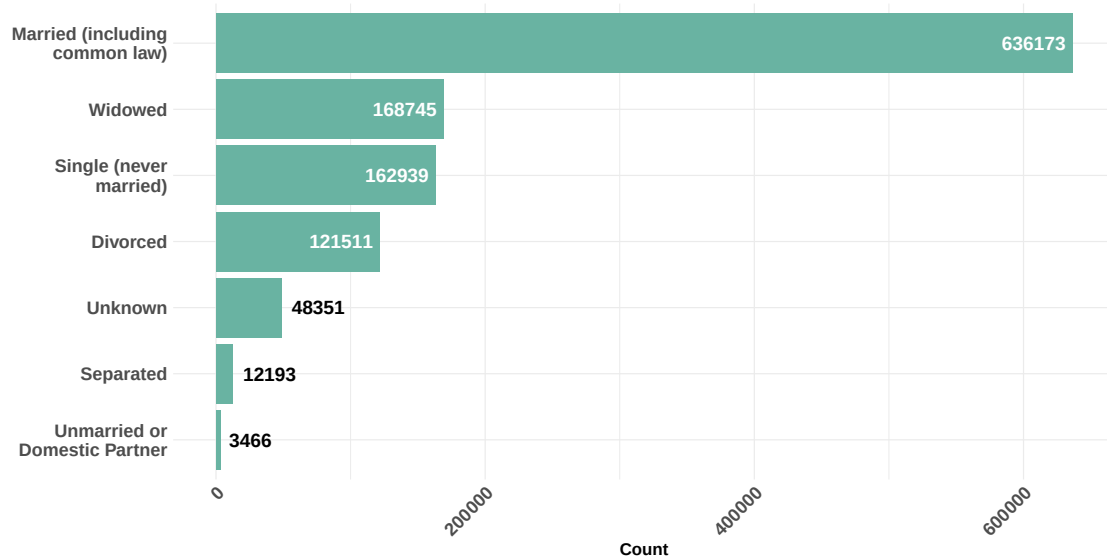


Figure 4.10: Distribution of Marital Status at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the number of breast cancer cases by marital status at the time of diagnosis, as recorded in the SEER database.

#### 4.1.10 Median Household Income (Inflation-Adjusted to 2022)

The distribution of *Median household income (inflation adjusted to 2022)* values, reveals distinct patterns across different income ranges. The data indicates a substantial concentration of households within the mid-income brackets, specifically between \$60,000 and \$85,000. The highest count of households falls within the \$70,000 - \$74,999 range, with 187,363 households, representing the mode of the income distribution.

Households earning between \$80,000 and \$84,999 also constitute a significant portion of the sample, with 144,514 households. Similarly, income brackets \$75,000 - \$79,999 and \$100,000 - \$109,999 show relatively high household counts, with 122,763 and 110,007 households respectively, indicating a sizeable middle to upper-middle-income group.

Lower income brackets (below \$60,000) show a gradual increase in household counts as income rises, starting from 14,031 households earning less than \$40,000, up to 91,048 households earning between \$60,000 and \$64,999. This suggests a relatively smaller proportion of low-income households in the distribution.

Conversely, higher income brackets beyond \$110,000 show a notable decrease, with counts dropping to approximately 58,800 in the \$110,000 - \$119,999 range and 56,881 households earning \$120,000 or more, illustrating an attenuation in the number of high-income households.

A small number of households (261) fall under the "Unknown/Missing" category, indicating minimal data gaps.

Overall, the income distribution indicates a predominantly middle-income demographic, with a strong central tendency around \$70,000 to \$85,000, reflecting the economic characteristics of the surveyed population in the given time frame.

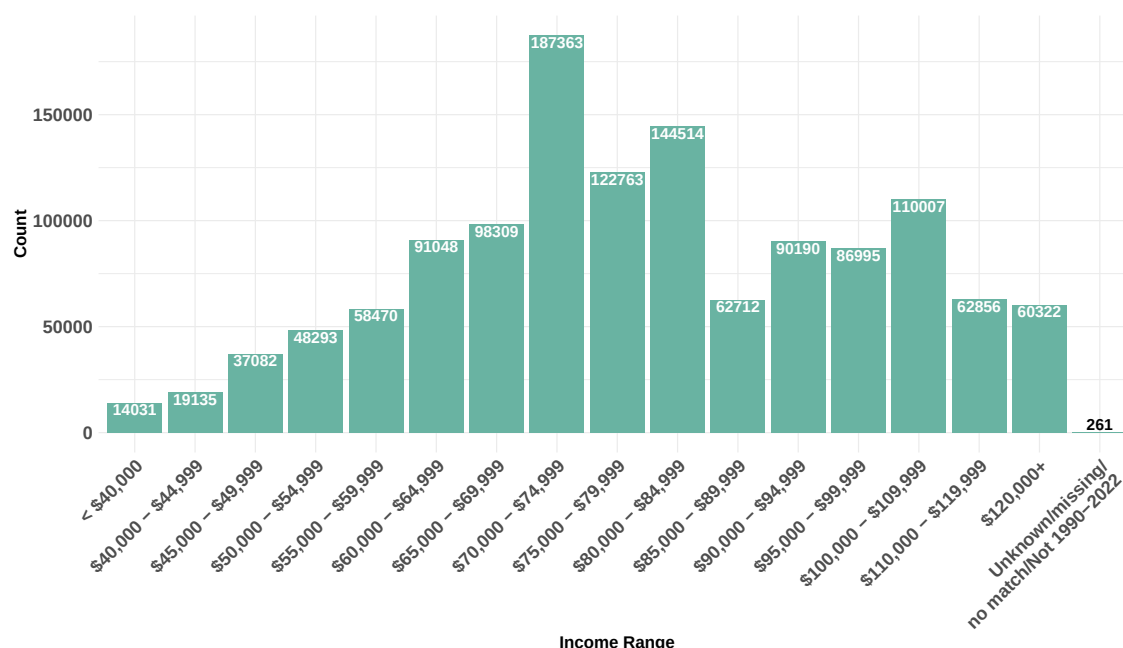


Figure 4.11: Distribution of Median Household Income Among Breast Cancer Patients (SEER, 2000-2021). The bar plot illustrates the distribution of breast cancer cases by median household income, with values inflation-adjusted to 2022 dollars.

#### 4.1.11 PR Status

The status of progesterone receptors (PR) in breast cancer indicates whether the cancer cells contain proteins capable of binding to the hormone progesterone, which may stimulate their growth. Similar to estrogen receptors (ER), it serves as a vital biomarker utilized in the classification of breast cancer types and in formulating treatment strategies. As illustrated in Figure 4.12, the *PR Status* is classified by frequency as follows: 'Positive' (n = 833,238), representing tumors with progesterone receptor expression and suggesting a potential response to hormone therapies; 'Negative' (n = 345,608), encompassing tumors without PR expression, which are less likely to benefit from hormone-based treatments; 'Borderline/Unknown' (108,119), indicating cases where PR testing provided inconclusive or borderline outcomes, necessitating careful clinical interpretation due to uncertain receptor status. 'Recode Not Available' (7,386 cases, mirroring ER data) indicates the absence of PR data for recoding, reflecting unavailable information rather than unclear test

outcomes. Differentiating between these categories is critical for analytical approaches since borderline/unknown instances depict ambiguous biomarker expression that might affect treatment, while recode unavailable cases highlight missing data, complicating classification and outcome analysis [86].

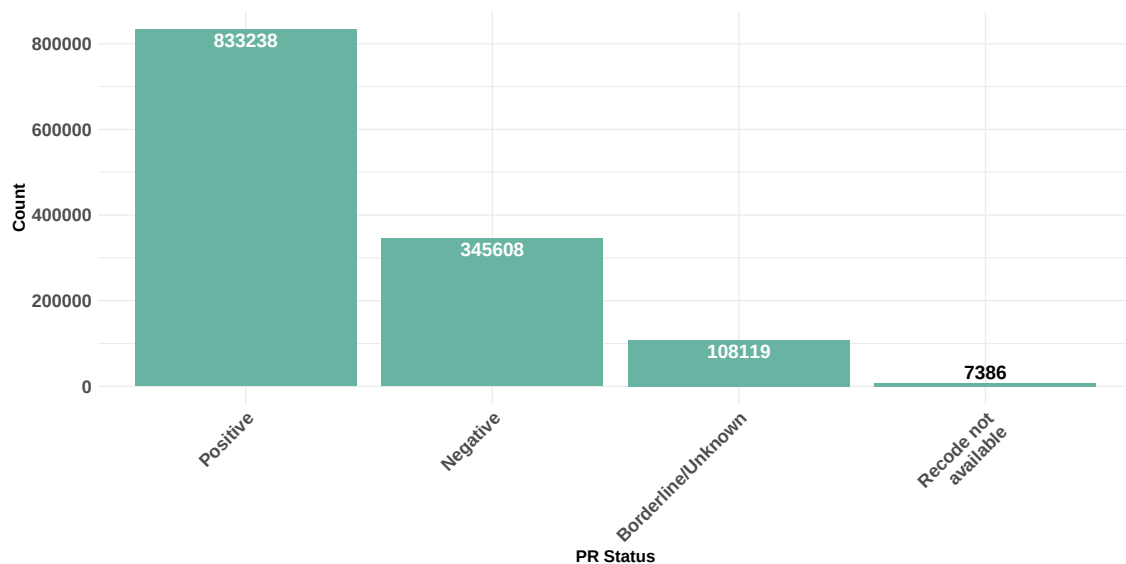


Figure 4.12: Distribution of PR Status at Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the status of PR at diagnosis, as recorded in the SEER database.

#### 4.1.12 Primary Site

The variable *Primary Site - labeled* identifies the specific anatomical location within the breast where the primary tumor was diagnosed. This categorical variable is coded according to the ICD-O (International Classification of Diseases for Oncology) conventions, with each code representing a distinct breast subsite. All classifications follow the SEER codebooks for breast tumors [112].

The most frequently occurring tumor location in this dataset, as shown in Figure 4.13, is the upper-outer quadrant of the breast (C50.4), with 422,223 cases. This accounts for a substantial proportion of all diagnoses and is consistent with clinical evidence that this quadrant contains a higher concentration of glandular tissue, making it more susceptible to malignant transformation [113]. The second most common category is overlapping lesions of the breast (C50.8), comprising 281,033 cases. This designation refers to a single tumor that spans multiple contiguous quadrants or regions of the breast. Such cases themselves do not convey information about tumor aggressiveness or prognosis, which depend on other clinical factors.

The “Breast, NOS” (Not Otherwise Specified) category (C50.9) includes 202,058 cases. This label is typically used when the exact location within the breast is unknown or cannot be determined, but it may also apply to non-contiguous tumors in different quadrants of the same breast, or to diffuse presentations such as inflammatory carcinoma. The upper-inner quadrant (C50.2) and lower-outer quadrant (C50.5) follow with 146,141 and 90,870 cases, respectively.

Other subsites, including the lower-inner quadrant (C50.3) and the central portion (C50.1), show moderate frequencies of 68,939 and 67,609 cases, respectively. The nipple (C50.0) and axillary tail (C50.6) are the least common tumor locations, with only 7,808 and 7,670 cases. The lower incidence in these regions aligns with their smaller size and reduced amount of glandular breast tissue, which is a key factor in breast cancer risk [114].

In conclusion, the distribution of tumor locations within the breast in this dataset closely reflects known anatomical and pathological patterns. The predominance of tumors in the upper-outer quadrant underscores the role of glandular tissue density in cancer development, while the high frequency of overlapping lesions and “NOS” classifications highlights ongoing challenges in precise tumor localization.

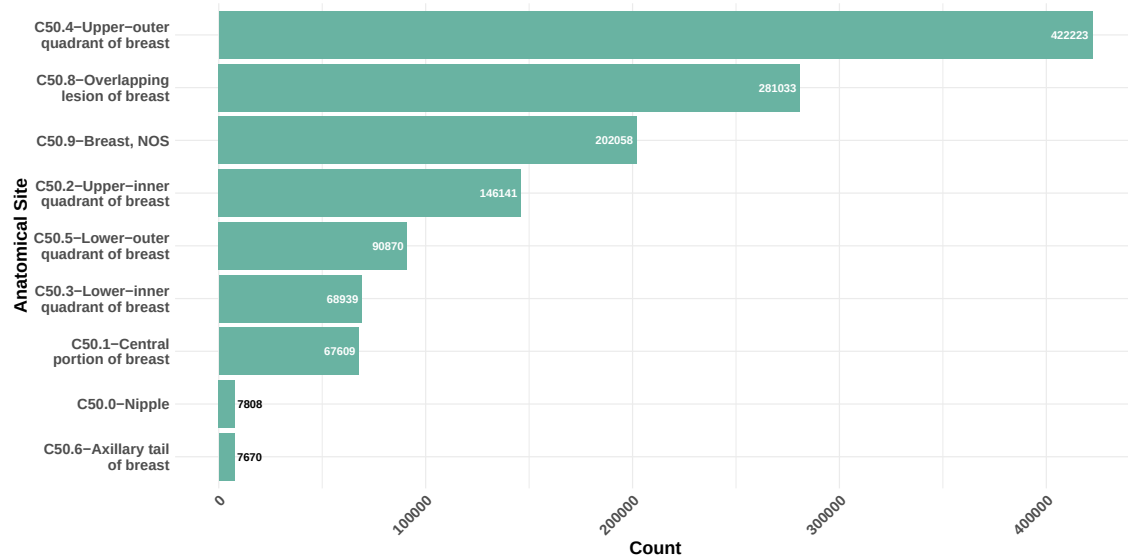


Figure 4.13: Distribution of primary tumor sites in breast cancer cases (SEER, 2000-2021). The bar plot illustrates the frequency of tumor locations coded according to SEER's anatomical site classification.

#### 4.1.13 Race and Origin

The *Race and Origin* variable is a classification used in the SEER cancer registry to group patients into five mutually exclusive race and ethnicity categories. This standardization

follows SEER’s recommended reporting guidelines, ensuring comparability of cancer incidence and survival statistics across diverse demographic groups. By consolidating data from multiple sources and harmonizing classification methods, the variable supports consistent epidemiologic analyses and reduces the risk of misclassification bias.

As shown in Figure 4.14, the dataset contains 899,812 cases of Non-Hispanic White, 143,670 cases of Hispanic (All Races), 129,696 cases of Non-Hispanic Black, 107,855 cases of Non-Hispanic Asian or Pacific Islander, 7,187 cases of Non-Hispanic American Indian/Alaska Native, and 6,131 cases categorized as Non-Hispanic Unknown Race. These figures highlight the predominance of Non-Hispanic White cases in the registry, followed by Hispanic and Non-Hispanic Black populations, while American Indian/Alaska Native and unknown race categories represent smaller proportions.

The distribution of cases across racial and ethnic groups is critical for understanding disparities in cancer incidence and outcomes. Certain cancers have been shown to vary significantly in prevalence, stage at diagnosis, and survival rates by race and ethnicity—differences often driven by a complex interplay of genetic, socioeconomic, environmental, and healthcare access factors.

Moreover, accurate race and ethnicity classification is essential for monitoring progress toward health equity goals. The SEER Race and Origin variable enables researchers and policymakers to track disparities over time, identify high-risk groups, and allocate resources effectively.

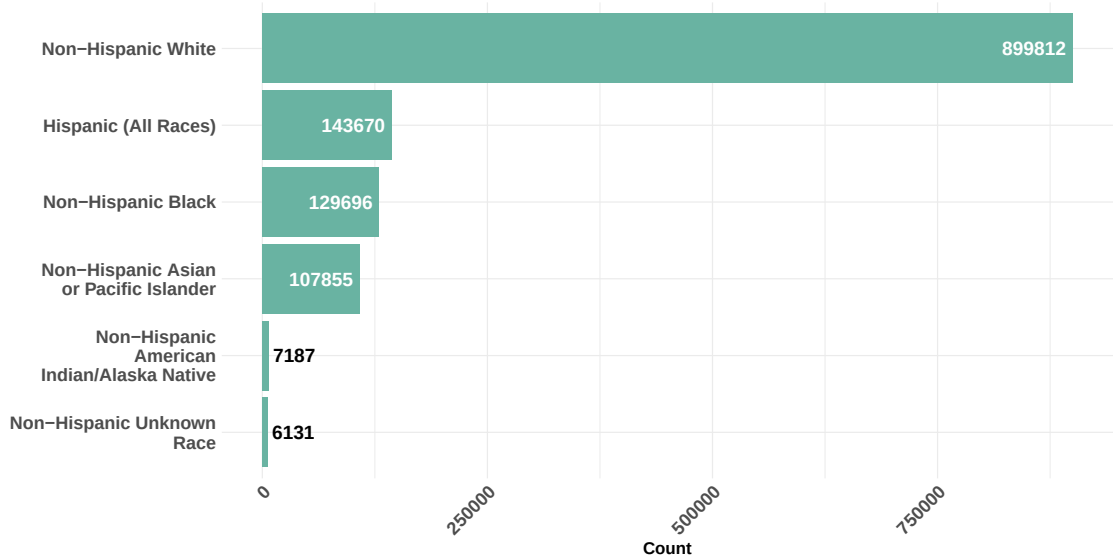


Figure 4.14: Distribution of Race and Origin in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the frequency of breast cancer cases across different racial and ethnic groups as recorded in the SEER database.

#### 4.1.14 Radiation

The variable *Radiation Recode* captures the type of radiation therapy administered to patients as part of their cancer treatment, including both specific modalities and administrative details.

As The most frequent category is 'None/Unknown', encompassing 621,282 patients, including both patients who did not receive radiation therapy and those for whom treatment status was not recorded. The inability to distinguish between these groups introduces uncertainty and potential bias in analyses examining the relationship between radiation therapy and outcomes, as missing or misclassified data may distort results.

The second most common category is 'Beam radiation' (588,356 patients). This modality, typically referring to external beam radiation therapy (EBRT), is the standard and most widely used form of radiation for breast cancer treatment. Its high prevalence is consistent with established clinical practice, particularly following breast-conserving surgery or in cases with lymph node involvement [115].

Follows radiation 'Recommended, unknown if administered' with 29,126 patients. This classification likely reflects incomplete follow-up or documentation, possibly due to patient transfer, loss to follow-up, or delays in care reporting. Radiation therapy 'Refused' was reported for 27,618 patients and may stem from personal preference, concerns about side effects, or clinical recommendations based on comorbidities, frailty, or age. 'Radiation, NOS (method or source not specified)' appears in 60,795 patient records and, while this indicates radiation was given, the lack of detail on the method decreases specificity in treatment-specific analyses.

Less common treatment modalities include 'Radioactive implants (includes brachytherapy)' (19,509 patients), occasionally used in early-stage, localized disease or as part of partial-breast irradiation protocols, 'Combination therapy (beam radiation with implants or radioisotopes)' (1,962 patients), indicating selective use in specific clinical situations, and lastly, radioisotopes (410 patients), representing a rare treatment approach in this population.

In summary, while beam radiation dominates among patients who received therapy, a substantial proportion of patients fall into categories marked by absent, incomplete, or ambiguous treatment data. The 'None/Unknown' and 'Radiation recommended, unknown if administered' groups, in particular, should be interpreted with caution in any survival

or comparative effectiveness analyses, as they may contain both treated and untreated individuals. The frequency distribution of these categories is illustrated in Figure 4.15, which highlights the predominance of beam radiation among known treatments and the large proportion of patients with unrecorded or absent radiation data.

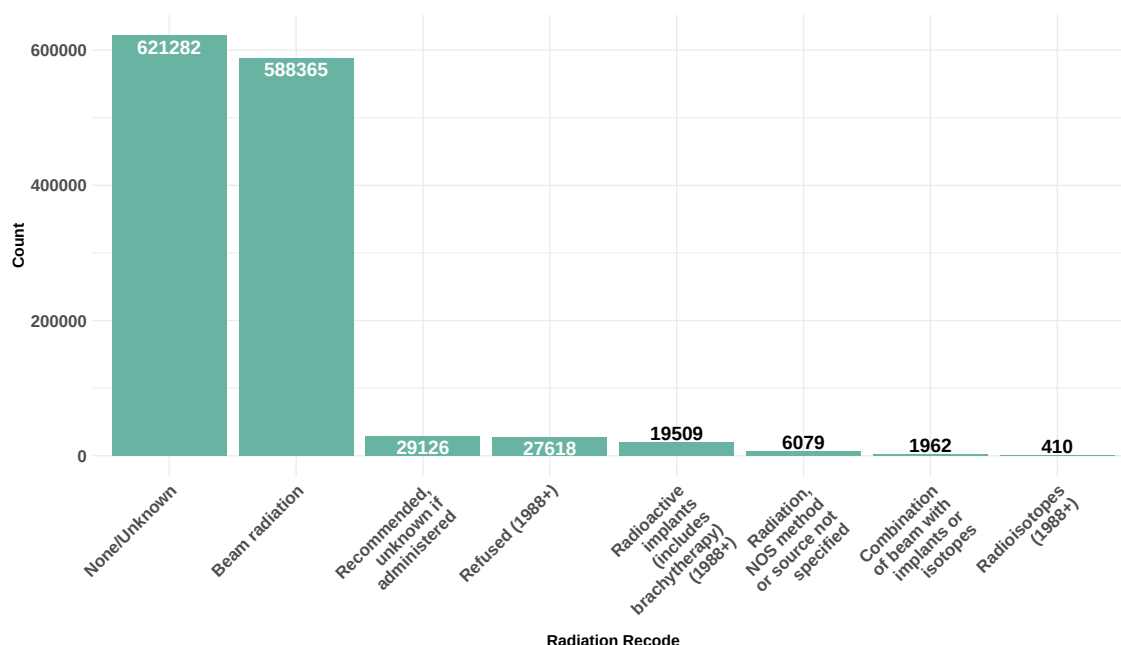


Figure 4.15: Distribution of Radiation Types in Breast Cancer Cases (SEER, 2000-2021). The bar plot displays the frequencies of different radiation therapies received by patients, as recorded in the SEER database.

#### 4.1.15 Rural-Urban Continuum Code

The SEER variable *Rural-Urban Continuum Code* classifies counties by degree of urbanization and proximity to metropolitan areas, providing a geographic and socioeconomic context for cancer epidemiology. In this cohort, the most frequent category was counties in metropolitan areas with a population of at least one million, accounting for 785,708 cases. This was followed by counties in metropolitan areas with 250,000 to one million population (269,597 cases) and counties in metropolitan areas with fewer than 250,000 population (98,343 cases). Outside metropolitan areas, nonmetropolitan counties adjacent to a metropolitan area represented 82,490 cases, while nonmetropolitan counties not adjacent to a metropolitan area accounted for 40,627 cases. Two small special categories were also present: “Unknown/Missing/No Match (Alaska and Hawaii, entire state)” comprised 1,435 cases, reflecting the fact that SEER reports for these states at the state level rather than by county, and “Unknown/Missing/No Match/Not 1990-2022” accounted for only 261 cases, representing records where classification could not be determined.

The distribution of cases across categories reflects the broader U.S. population pattern, with the majority of cases arising from large metropolitan areas. This variable is particularly valuable for exploring disparities in cancer diagnosis, treatment, and outcomes, as geographic setting often correlates with healthcare access, screening availability, socioeconomic status, and environmental exposures. However, there are several considerations when interpreting these results. The high concentration of cases in large metropolitan areas may overshadow trends in smaller or rural populations, requiring caution when comparing groups of unequal size. Rural counties, especially those not adjacent to metropolitan areas, may have longer travel times to treatment facilities, fewer oncology specialists, and lower screening rates, which could contribute to differences in both incidence and survival outcomes. The aggregated reporting for Alaska and Hawaii prevents assessment of within-state geographic variation, and the small “Not 1990-2022” category reflects temporal inconsistencies in coding that could affect longitudinal analyses. Additionally, because the classification is based on county-level USDA codes, within-county differences such as suburban-urban variation are not captured, potentially masking finer-scale disparities. Overall, while the *Rural-Urban Continuum Code* provides an important framework for examining geographic influences on cancer care and outcomes, careful handling of special categories and awareness of urban-rural imbalances are essential for robust interpretation.

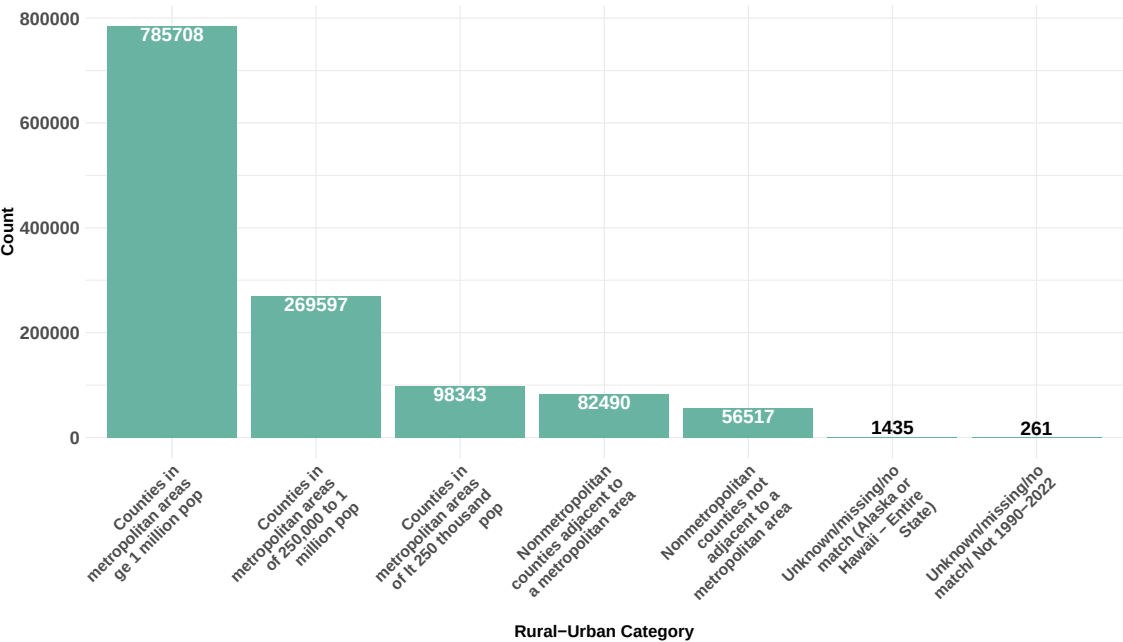


Figure 4.16: Distribution of Rural/Urban Continuum Codes in Breast Cancer Cases (SEER, 2000-2021). The bar plot displays the frequencies of breast cancer cases across different geographic classifications based on the Rural/Urban Continuum Codes (RUCC), which reflect the degree of urbanization and proximity to metropolitan areas of patients’ residences.

#### 4.1.16 SEER Cause-Specific Death Classification

The variable *SEER cause-specific death classification* captures vital status information and whether a patient's death was attributable to their cancer diagnosis. The majority of patients, 1,096,055 individuals, were classified as either alive or deceased due to causes unrelated to breast cancer. This category includes those who remained alive throughout the follow-up period or whose deaths were attributed to external or comorbid conditions, rather than the primary cancer under investigation.

A substantial portion of the remaining cases, 191,994 patients, were recorded as having died specifically as a result of the diagnosed cancer. This group represents cancer-attributable mortality and is particularly relevant for survival analysis and outcome prediction modeling. A smaller group of 6,277 individuals were categorized as deceased, but with missing or unknown cause of death, limiting the certainty regarding whether breast cancer was the underlying cause. Finally, a negligible number of cases (only 25 records), were labeled as 'N/A not seq 0-59', which likely corresponds to incomplete or invalid sequencing of tumors or a data collection artifact. While this group is statistically insignificant, its presence indicates a minor degree of data irregularity that warrants acknowledgment.

Overall, this distribution underscores that while the vast majority of patients survive or die from other causes, a meaningful fraction experiences cancer-specific mortality, affirming the clinical importance of identifying prognostic factors related to survival outcomes in breast cancer patients.

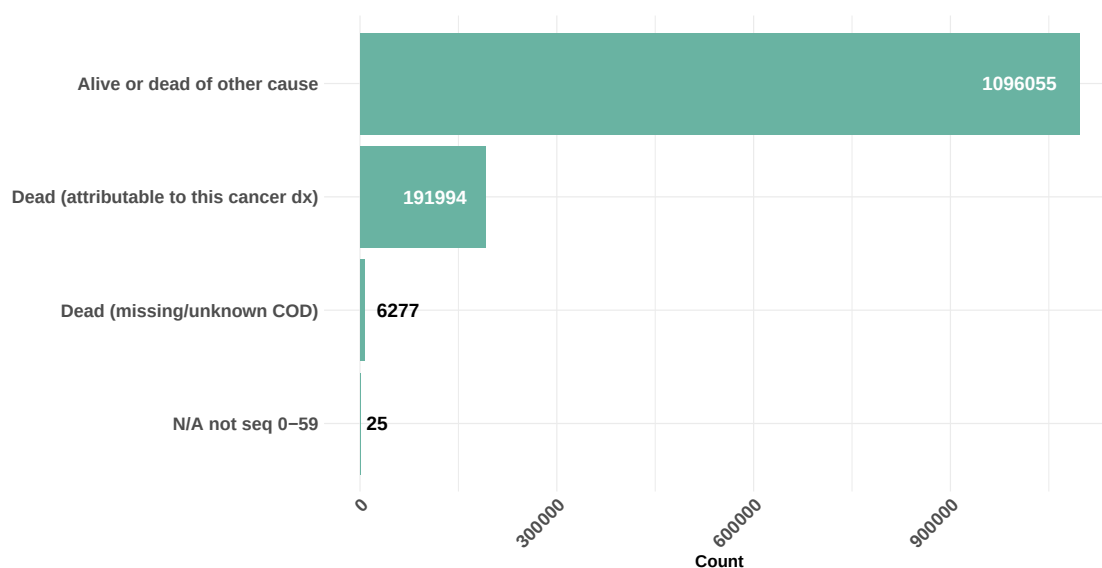


Figure 4.17: Distribution of Cause-Specific Death Classification in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the number of patients by cause-specific death classification, as recorded in the SEER database. This classification distinguishes between deaths due to breast cancer and those alive, or dead due to other causes.

#### 4.1.17 Sex

The cohort was overwhelmingly female, with 1,144,709 cases (99.2%), while males represented only 8,669 cases (0.8%), as seen in Figure 4.18. This distribution reflects the underlying epidemiology of breast cancer, which occurs far more frequently in women due to biological, hormonal, and genetic factors. Although male breast cancer is rare, it is clinically significant, often diagnosed at a later stage and associated with different prognostic considerations compared to female cases. The low proportion of males in this dataset is consistent with national incidence patterns, where men account for less than 1% of all breast cancer diagnoses. The marked sex disparity should be taken into account in statistical modeling, as male-specific analyses may be limited by small sample size, and sex-stratified results may lack precision for the male subgroup.

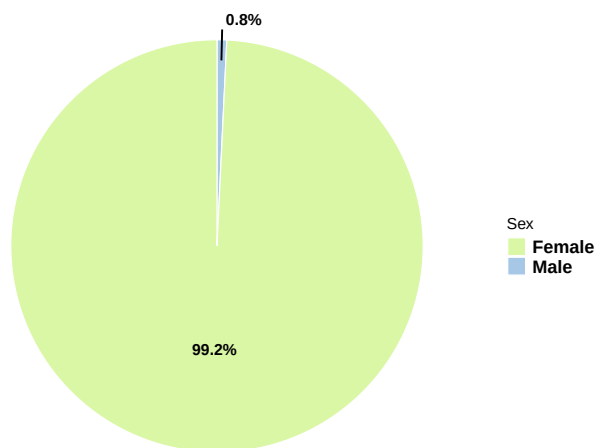


Figure 4.18: Distribution of Patients' Sex in Breast Cancer Cases (SEER, 2000-2021). The pie chart illustrates the proportion of breast cancer cases by gender, as recorded in the SEER database.

#### 4.1.18 Survival Months

The survival time in the cohort, measured in months from diagnosis to either death or last follow-up, ranged from 0 to 263 months. The first quartile was 33 months, the median survival was 76 months, and the third quartile was 138 months, indicating that half of the patients survived between approximately three and eleven and a half years. The mean survival time was 90.98 months, slightly higher than the median, suggesting a right-skewed distribution influenced by a subset of long-term survivors.

The minimum survival of 0 months reflects patients who died shortly after diagnosis, likely due to advanced disease at presentation or comorbid conditions, while the maximum observed survival exceeded 21 years, representing long-term survivors or patients with indolent disease. This wide range underscores the heterogeneity in breast cancer prognosis, influenced by tumor biology, stage at diagnosis, treatment modalities, and patient factors such as age and comorbidity burden. The distribution of *Survival Months* is illustrated in Figure 4.19, which highlights the central tendency, spread, and presence of potential outliers in the data.



Figure 4.19: Distribution of Survival Months in Breast Cancer Cases (SEER, 2000-2021). The density plot illustrates the distribution of the number of months that breast cancer patients survived, as recorded in the SEER database.

#### 4.1.19 Time from Diagnosis to Treatment

The variable *Time from Diagnosis to Treatment (in days)* captures the number of days between a patient's cancer diagnosis and the initiation of their treatment. It is a continuous numerical variable that provides insights into the efficiency and responsiveness of the healthcare system in addressing newly diagnosed cancer cases.

According to the summary statistics, the minimum time to treatment is 0 days, suggesting that for some patients, treatment began on the same day as the diagnosis, likely in cases of emergent care or where diagnostic confirmation and therapeutic intervention occurred simultaneously. The 1st quartile is 13 days, the median is 27 days, and the 3rd quartile is 43 days, indicating that half of the patients began treatment within approximately four weeks of diagnosis. The mean time to treatment is 32.05 days, slightly higher than the median, which, along with the density plot (see Figure 4.20), suggests a right-skewed distribution. The density plot visually confirms this skewness. A large proportion of patients initiated treatment within the first 50 days, with a peak occurring between 10 and 30 days. After that, the frequency of treatment initiation declines sharply. However, there is a long tail extending out to the maximum value of more than 731 days, roughly two years, highlighting that either a small number of patients experienced substantial delays or, a more plausible explanation, some errors were introduced in the dataset. This long-tailed distribution raises several important considerations. While most patients are treated relatively

quickly, delays beyond the median can be clinically significant, especially in cancer care where time-to-treatment is known to influence prognosis.

When examined categorically, the distribution of time to treatment across the groups is further elucidated in the box plot (Figure 4.21). This corroborates what is observed in the density plot (Figure 4.20), showing that while most patients start treatment relatively quickly, there is a subset experiencing longer delays. The increasing median times across categories confirm the trend toward prolonged treatment initiation in some patients, which may have important clinical implications despite representing a smaller proportion of the population. Additionally, there are 140,973 cases categorized as 'Unable to Calculate', indicating missing or incomplete data for this variable, which may affect analyses and warrants careful consideration in interpretation.

In summary, the data shows that most patients receive treatment in a timely manner, typically within a month of diagnosis. However, the presence of outliers with prolonged delays underscores the need to investigate the causes of such delays, as they may disproportionately affect certain subgroups and have critical implications for outcomes. This variable will be particularly useful in downstream survival analyses and quality of care evaluations.

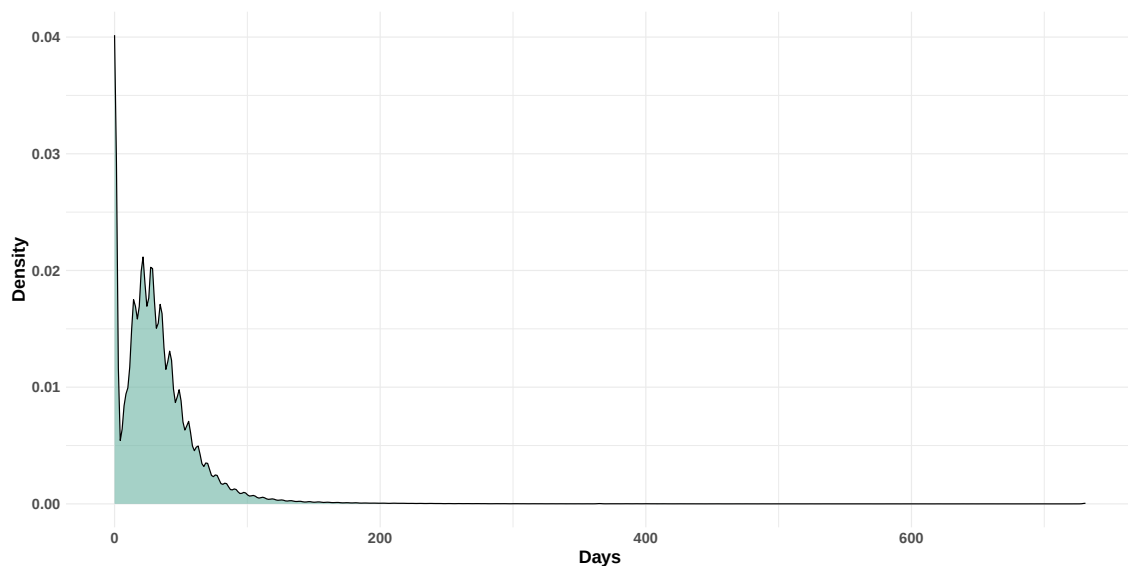


Figure 4.20: Distribution of Time from Diagnosis to Treatment in Breast Cancer Cases (SEER, 2000-2021). The density plot illustrates the distribution of breast cancer cases based on the number of days between diagnosis and the initiation of treatment, as recorded in the SEER database.

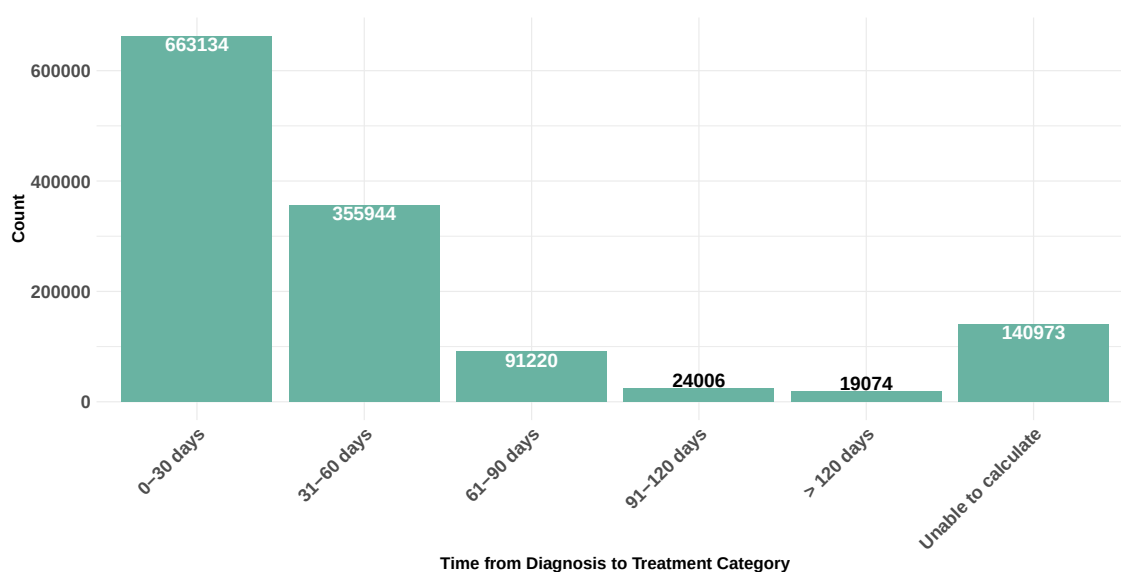


Figure 4.21: Distribution of Time from Diagnosis to Treatment in Breast Cancer Cases (SEER, 2000-2021) by category.

#### 4.1.20 Total Number of Benign/Borderline Tumors

The distribution of *Total Number of Benign/Borderline Tumors* per patient was highly skewed toward zero. Across the cohort, the number of benign/borderline tumors ranged from 0 to 5, with a mean of 0.008 tumors per patient and a standard deviation of 0.09. The median, first quartile, and third quartile were all 0, indicating that the vast majority of patients did not present with any benign or borderline tumors.

As shown in Figure 4.22, the distribution is extremely right-skewed, with over 1.1 million patients (approximately 99.2%) having no benign or borderline tumors. Only a very small proportion of patients presented with one or more such tumors: 8,412 patients had one, 302 had two, 24 had three, and just a single patient each had four or five tumors. This skewness is expected given that while benign and borderline tumors are captured in breast tumor registries, they comprise a small minority compared to malignant tumors, which form the majority of breast cancer cases reported in population registries like SEER [116].

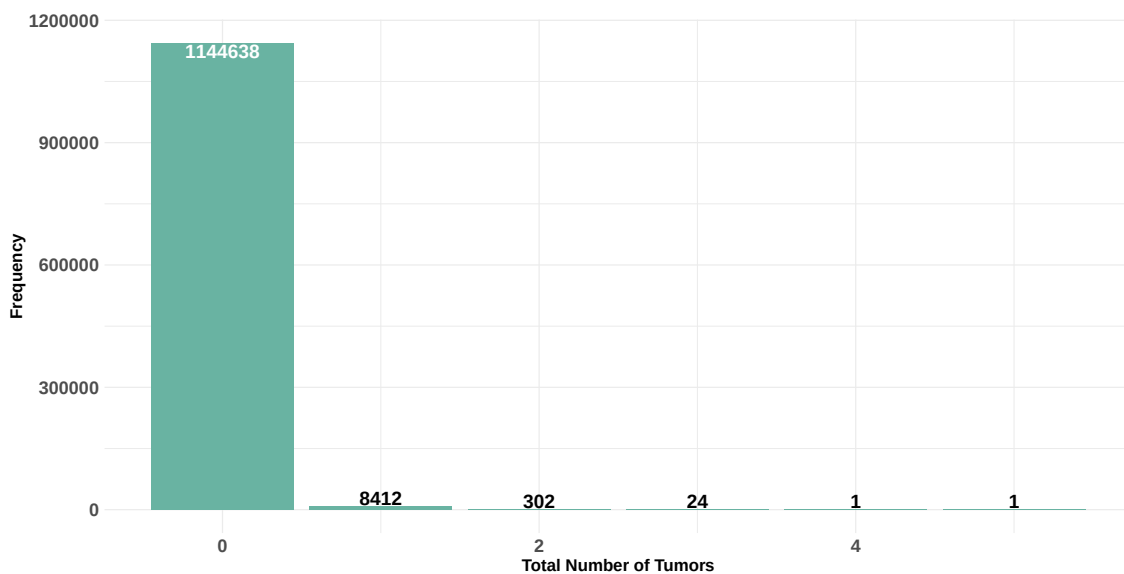


Figure 4.22: Distribution of Total Number of Benign/Borderline Tumors in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the distribution of breast cancer cases based on the number of benign or borderline tumors reported per patient, as recorded in the SEER database.

#### 4.1.21 Total Number of In Situ/Malignant Tumors

The distribution of *Total number of In situ/malignant tumors* was markedly skewed toward lower values. The most frequent category was 1 tumor (911,605 cases), followed by 2 tumors (299,877 cases). Cases with 3 tumors numbered 66,302, while 4 tumors accounted for 13,264. The frequency continued to decline sharply for higher tumor counts: 2,453 cases for 5 tumors, 594 for 6 tumors, 133 for 7 tumors, 50 for 8 tumors, 18 for 9 tumors, and 6 each for 10, 11 and 13 tumors. Higher counts were extremely rare, with only 2 cases each for 12, 14, 16, 20 and 22 tumors, and a single case for 18 tumors. An additional 250 cases were classified as 'Unknown'.

The steep decline in frequency as tumor counts increase is clearly visible in Figure 4.23, highlighting that multiple in situ or malignant tumor occurrences beyond two are rare in the dataset. This distribution aligns with clinical expectations, as most patients present with a single malignant or in situ tumor, and higher counts typically reflect either rare syndromic cases, metastatic misclassification, or data anomalies.

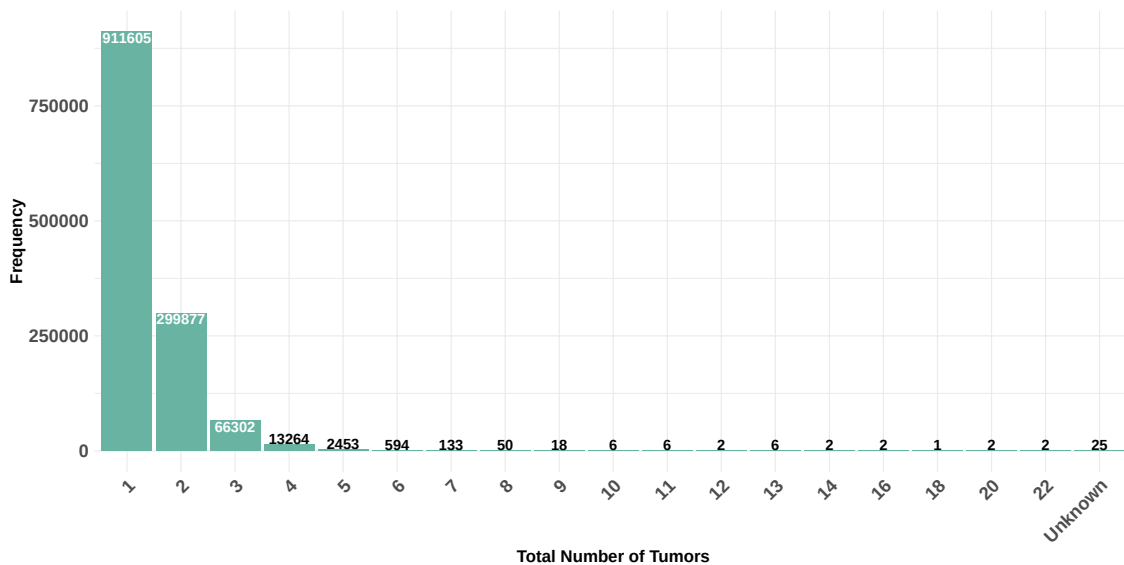


Figure 4.23: Distribution of Total Number of In Situ/Malignant Tumors in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the distribution of breast cancer cases based on the number of in situ or malignant tumors reported per patient, as recorded in the SEER database.

#### 4.1.22 Tumor Size Over Time

The variable *Tumor Size Over Time* exhibits a right-skewed distribution, with values ranging from 0 to 200 mm, as shown in the density plot in Figure 4.24. Summary statistics confirm this skewness, with a median tumor size of 17 mm and a mean of 21.98 mm. The first quartile is 10 mm, and the third quartile is 27 mm, indicating that 50% of tumors fall within this interquartile range.

When tumor sizes are categorized, the absolute frequencies reflect a similar pattern of right skewness, as illustrated in Figure 4.25. The number of tumors gradually increases from 13,452 cases with tumors  $\leq 1$  mm to 84,130 cases for tumors  $> 1$  mm and  $\leq 5$  mm. This upward trend continues with 209,253 cases for tumors between 5 and 10 mm, peaking at 418,785 cases for tumors  $> 10$  mm but  $\leq 20$  mm. Frequencies then decline for larger tumors, with 377,042 cases measuring between 20 and 50 mm, and only 82,346 cases exceeding 50 mm.

Special categories with smaller absolute frequencies capture cases where tumor size was missing, ambiguous, or recorded using site-specific or clinical codes. The category 'Unknown or size unreasonable (including sizes 401-989 mm)' contains 87,235 cases, representing outliers or absent size data. Additionally, 13,046 cases were classified as '990 (microscopic focus)', indicating microscopic tumor foci not measurable macroscopically. The '998 (site-specific code)' category, comprising 6,022 cases, reflects extensive tumor involvement, such as inflammatory carcinoma affecting  $\geq \frac{3}{4}$  of the breast, applied to

specific anatomic sites (C50.0-C50.6 and C50.8-C50.9). Finally, 3,040 cases were coded as 'Tumor Size Not Consistent Over Time or Not Applicable for this Site', indicating either inconsistent measurements or cases where tumor size is not relevant.

These findings highlight the heterogeneous distribution of tumor sizes within the cohort, with a majority of tumors clustered in the small to moderate size ranges and a long tail toward larger sizes. The presence of a substantial number of cases with unknown or clinically coded tumor sizes emphasizes the importance of carefully handling these categories in any analysis to avoid bias. The skewed distribution and wide range of tumor sizes reflect biological variability and differences in stage at diagnosis, which are critical factors influencing treatment decisions and prognosis. Furthermore, categorizing tumor size facilitates stratified analyses and supports clinical interpretation, especially when assessing survival outcomes or response to therapy.

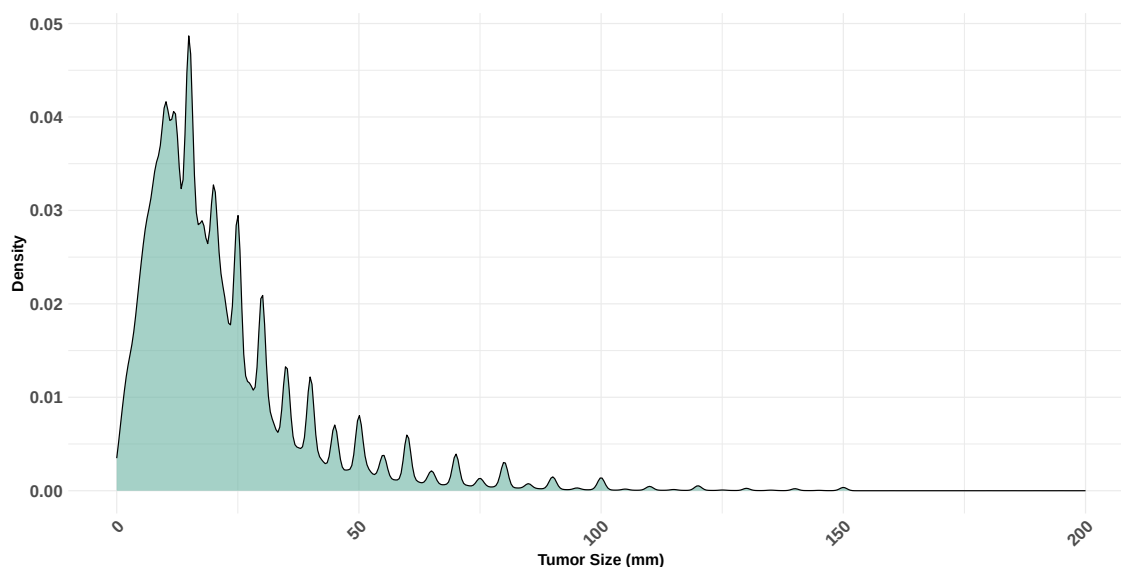


Figure 4.24: Distribution of Tumor Size in Breast Cancer Cases (SEER, 2000-2021). The density plot illustrates the distribution of tumor sizes, measured in millimetres, among breast cancer cases as recorded in the SEER database.

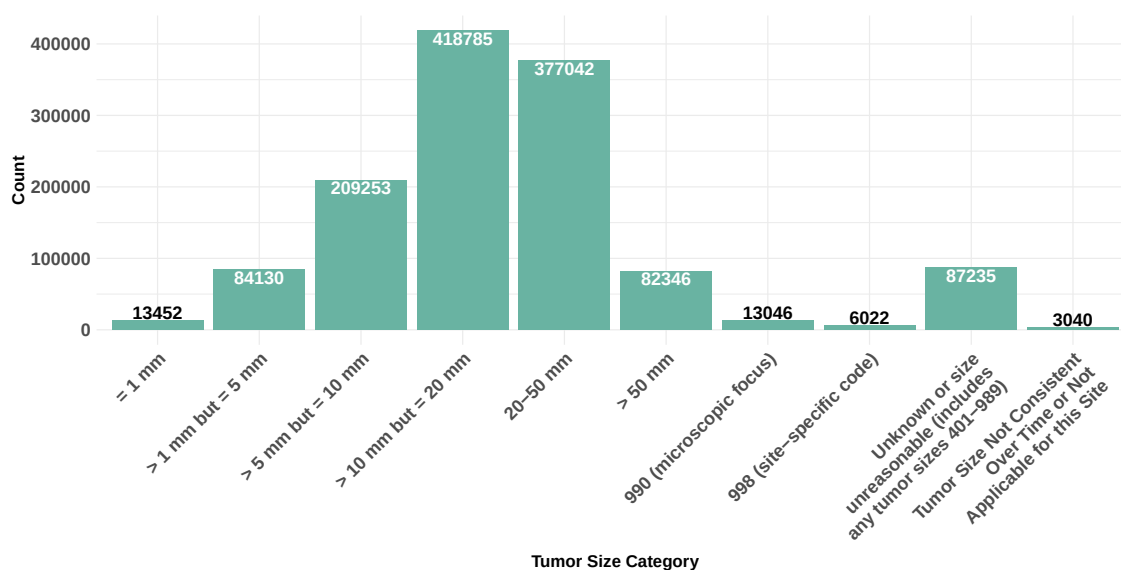


Figure 4.25: Distribution of Tumor Size in Breast Cancer Cases (SEER, 2000–2021) by Category. The bar plot illustrates the distribution of tumor sizes, measured in millimetres, among breast cancer cases as recorded in the SEER database.

#### 4.1.23 Tumor Size Summary

The variable *Tumor Size Summary (2016+)* describes the recorded tumor size for cases diagnosed from 2016 onward. As shown in the density plot in Figure 4.26, values range from 0.0 to 999.0 mm, with a first quartile of 11.0 mm, a median of 18.0 mm, a mean of 83.3 mm, and a third quartile of 32.0 mm. The notably large maximum value of 999.0 mm corresponds to a special code indicating unknown or unreasonable tumor size. The mean being substantially higher than the median indicates a strong right skew, reflecting the influence of extreme or coded values.

When categorized (Figure 4.27), the most prevalent value is *Blank(s)*, with 890,464 cases. This is expected, as the variable only contains valid data for diagnoses made in 2016 or later, meaning the majority of earlier cases are missing and coded as blank. Among the cases with recorded sizes, the frequency distribution follows the same trend observed in the density plot: tumors  $\leq 1$  mm accounted for 6,088 cases, tumors  $> 1$  mm and  $\leq 5$  mm for 28,137 cases, and tumors  $> 5$  mm and  $\leq 10$  mm for 65,235 cases. The peak category was tumors  $> 10$  mm and  $\leq 20$  mm, with 126,281 cases, followed by a decline to 122,229 cases for tumors  $> 20$  mm and  $\leq 50$  mm. The smallest group among recorded sizes was tumors  $> 50$  mm, with 55,917 cases.

This distribution highlights that, aside from the large proportion of missing values due to temporal limitations of the variable, most tumors fall within the 10–50 mm range, a clinically relevant size bracket influencing surgical planning, staging, and prognosis. The skewness

and high frequency of special codes underscore the importance of cautious interpretation, especially when integrating this variable into longitudinal or pre-2016 datasets.

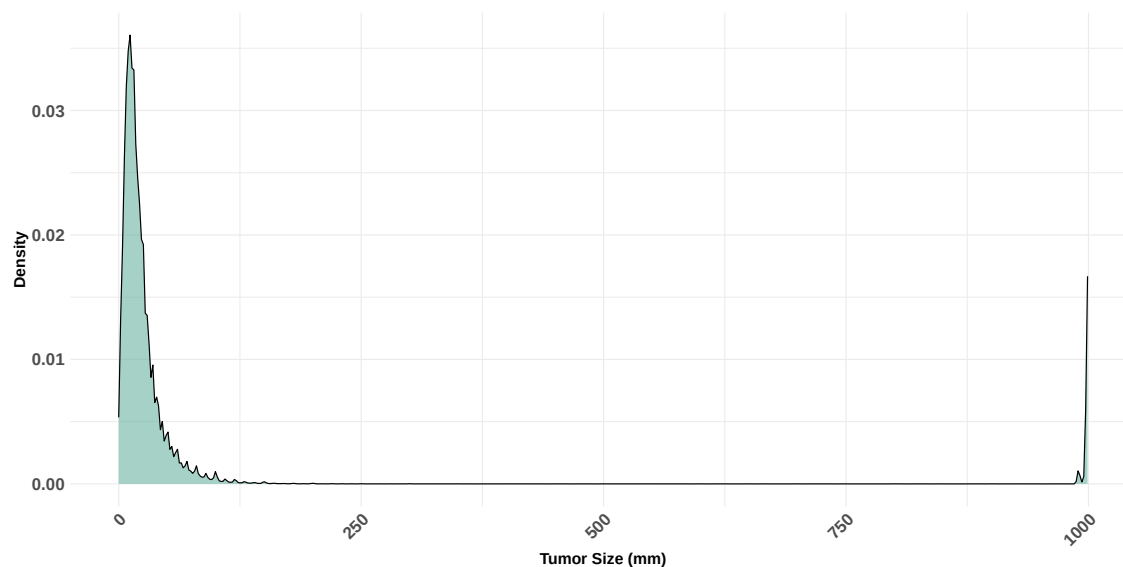


Figure 4.26: Distribution of Tumor Size in Breast Cancer Cases (SEER, 2016-2021). The density plot illustrates the distribution of tumor sizes, measured in millimetres, among breast cancer cases as recorded in the SEER database.

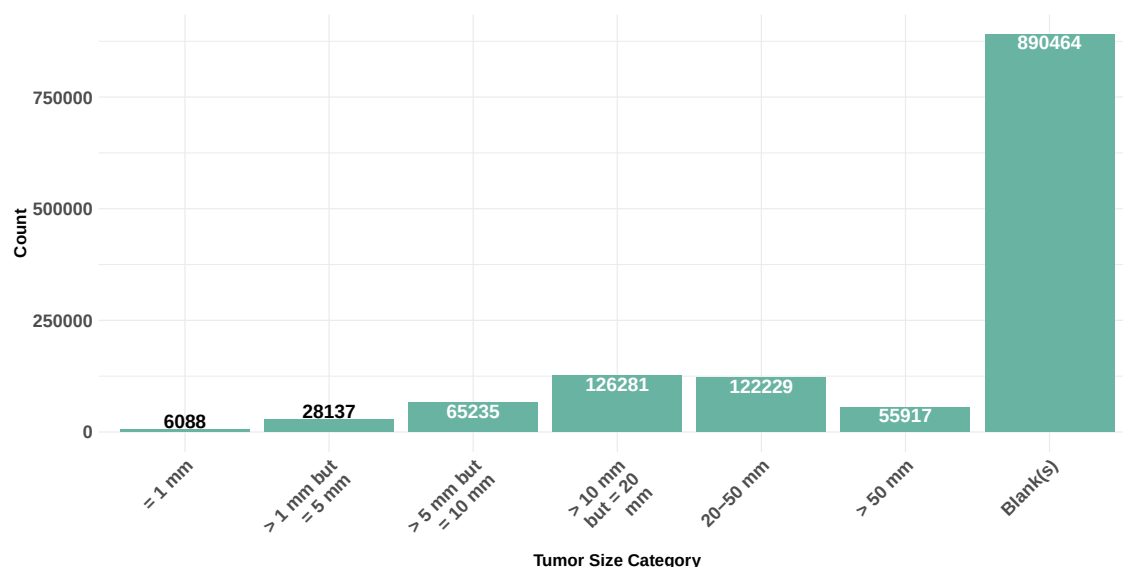


Figure 4.27: Distribution of Tumor Size in Breast Cancer Cases (SEER, 2016-2021) by Category. The bar plot illustrates the distribution of tumor sizes, measured in millimetres, among breast cancer cases as recorded in the SEER database.

#### 4.1.24 Year of Diagnosis

The variable *Year of Diagnosis* represents the calendar year in which patients received their diagnoses. This is a discrete numerical variable and serves as a proxy for understanding temporal patterns in diagnosis frequency over the observed period. According to the summary statistics, the earliest year of diagnosis in the dataset is 2000 and the latest

is 2021. Both the median year of diagnosis is 2012, and the mean 2011. The interquartile range, from 2006 (1st quartile) to 2017 (3rd quartile), indicates that most diagnoses occurred within this 11-year window.

Visual inspection of the distribution through the horizontal bar plot (Figure 4.28 reveals a clear upward trend in the number of diagnoses over time. In the early 2000s, the number of diagnoses hovered just above 51,000 cases per year. From around 2005 onward, there is a noticeable and consistent increase in counts, culminating in a peak of 72,194 diagnoses in 2021. This increasing trend likely reflects a combination of factors that include population growth, improvements in healthcare access, increased public awareness, better screening programs, or improvements in reporting practices and data completeness over the years. The visibly smaller amount of diagnosis made in 2020 probably reflect the consequences of SARS-Cov2.

Although the increase in diagnoses over time may reflect positive developments in health systems and disease detection, it introduces important considerations for longitudinal and survival analyses. For instance, individuals diagnosed in later years, particularly those from 2019 to 2021, have not had enough follow-up time to observe long-term outcomes, such as five-year survival. Including these recent cases in time-to-event analyses without proper censoring or adjustment could lead to biased estimates. Therefore, it is advisable to treat these years cautiously in downstream analyses, either by applying censoring methods or excluding them altogether when analysing five-year outcomes.

Overall, the *Year of Diagnosis* variable not only offers valuable insights into trends in disease detection over time but also plays a critical role in shaping the methodological approach to survival and outcome analyses. It highlights the need for careful temporal adjustments when interpreting patterns and comparing outcomes across different diagnosis periods.

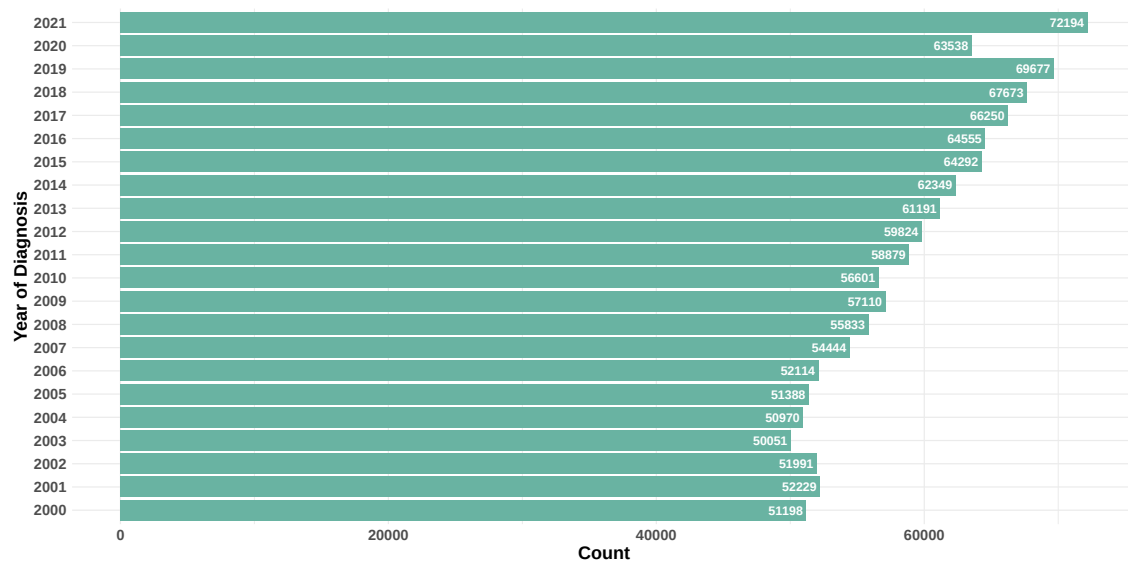


Figure 4.28: Distribution of Year of Diagnosis in Breast Cancer Cases (SEER, 2000-2021). The bar plot illustrates the distribution of when the tumor was diagnosed, based on year, among breast cancer cases as recorded in the SEER database.

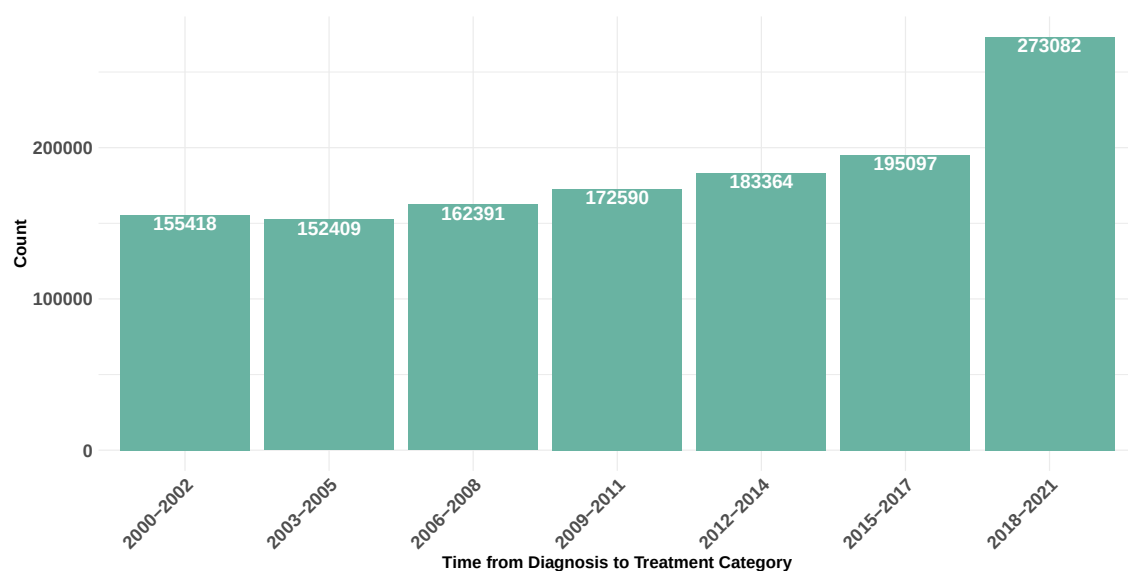


Figure 4.29: Distribution of Year of Diagnosis in Breast Cancer Cases (SEER, 2000-2021) by Category. The bar plot illustrates the distribution of when the tumor was diagnosed, based on year, among breast cancer cases as recorded in the SEER database.

## 4.2. Contingency Tables and Chi-square Tests

Prior to conducting the chi-squared analyses, several data adjustments were made to ensure reliable statistical testing and to avoid violations of minimum cell count assumptions. Specifically, *age* categories below 20 years were combined due to the small number of cases ( $n=62$ ). For *tumor size*, the category “Not consistent over time or not applicable for this site” ( $n=9$ ) was excluded. In the analysis of *benign/borderline tumors*, patients with

2, 3, or 4 tumors were grouped into a single “2 or more tumors” category, and for *in situ*/malignant tumors, cases with more than 5 tumors were similarly combined. Finally, in the *Combined Summary Stage* variable, the “In situ” category was removed because it contained fewer than 5 observations. These adjustments ensured more robust chi-squared approximations and enhanced the reliability of subsequent survival analyses.

Chi-square analyses revealed statistically significant associations between all examined clinical, demographic, and socioeconomic variables and 5-year breast cancer survival (all  $p < 0.001$ ). However, the strength and direction of associations varied considerably across categories.

### **Demographic and socioeconomic factors**

Age showed a strong association with survival, with the best outcomes among patients aged 45-69 ( $\approx 84$ -85%) and lower survival both at younger ( $> 30$  years,  $\approx 73\%$ ) and older ages ( $\geq 90$  years, 70.6%). Female patients had better survival than males (82.7% vs. 77.4%). Racial/ethnic differences were pronounced: Asian/Pacific Islander patients had the highest survival (86.4%), while non-Hispanic Black patients had the lowest (74.1%). Marital status and income were also predictive: married/partnered individuals and those living in higher-income areas had consistently higher survival rates. The rural-urban differences were smaller, although patients in large metropolitan areas performed slightly better.

### **Tumor characteristics**

Prognosis was strongly determined by disease stage and tumor size. Localized cancers had the highest 5-year survival (92.6%), compared to only 24.1% in distant-stage disease. Similarly, survival decreased with increasing tumor size, ranging from 94.9% for tumors  $\geq 5$  mm to 57.7% for tumors  $> 50$  mm. Molecular markers also played an important role: ER/PR positivity was linked to substantially higher survival ( $\approx 85$ -87%), HER2 negativity conferred a modest advantage, and triple-negative disease showed the poorest outcomes ( $\approx 75\%$ ). Primary tumor site also showed variation, with upper-inner/outer quadrants associated with more favourable survival compared to “Breast, NOS”.

### **Treatment-related factors**

Patterns of survival varied across treatment modalities. Patients who received radiation therapy, especially brachytherapy or multimodal approaches, had superior outcomes compared to those who refused treatment (95.2% vs. 74.1%). Chemotherapy recipients

had lower raw survival (77.3% vs. 86.2% without chemotherapy), but this likely reflects confounding by indication, since chemotherapy is typically administered to higher-risk patients. Timely initiation of treatment was associated with better outcomes, with survival decreasing sharply when delays exceeded 120 days (76.5%).

### **Other clinical variables**

Patients with breast cancer as a first primary malignancy had slightly better survival (83.2%) than those with prior cancers (79.9%). Interestingly, the presence of benign/borderline or multiple in situ tumors did not worsen survival and may be linked to earlier detection or closer monitoring.

### **Temporal trends**

Survival improved steadily across the study period, from 78.1% in 2000-2002 to 88.8% in 2015-2016, possibly reflecting advances in diagnosis and treatment over time.

## **4.3. Survival Analysis: Kaplan-Meier Curves and Log-rank Test**

Breast cancer survival exhibited significant variation across demographic, clinical, tumor-related, treatment, socioeconomic, geographic, and temporal factors. Age at diagnosis was a strong determinant of survival, with younger patients demonstrating lower than expected mortality and older patients, particularly those aged 85 years and above, experiencing substantially higher mortality. This pattern likely reflects both the influence of comorbidities and differences in disease aggressiveness or treatment tolerance with increasing age. Male patients similarly exhibited poorer outcomes than females, consistent with previous literature indicating later-stage diagnosis and more aggressive disease in male breast cancer. Survival also varied across racial and ethnic groups, with non-Hispanic Black patients experiencing markedly higher observed mortality relative to expected values, while Hispanic and non-Hispanic White patients demonstrated survival closer to expectations, suggesting potential contributions of disparities in tumor biology, access to care, and socioeconomic conditions. Marital status at diagnosis was associated with survival, as widowed and single individuals exhibited higher mortality than married patients, indicating that social support and household structure may influence treatment adherence and early detection.

Socioeconomic and geographic factors further influenced survival outcomes. Patients residing in lower-income areas, particularly those with median household incomes below

\$40,000, experienced disproportionately higher mortality, highlighting the impact of socioeconomic disadvantage on access to timely and effective care. Rural-urban status also played a role, with patients living in nonmetropolitan counties not adjacent to metropolitan areas exhibiting poorer survival, consistent with limited availability of specialized oncology services and cancer screening programs in these regions.

Tumor biology strongly predicted prognosis. Hormone receptor status, HER2 expression, and breast cancer subtype all demonstrated significant associations with survival. Estrogen receptor (ER) and progesterone receptor (PR) positive tumors were associated with markedly better outcomes, whereas receptor-negative and triple-negative (HR-/HER2-) tumors were linked to worse survival. Similarly, HER2-positive tumors exhibited outcomes near expected values, likely reflecting the benefit of targeted therapies, while borderline or unknown HER2 status was associated with excess mortality. The anatomical location of the primary tumor also influenced survival, with tumors located in the upper-inner, lower-inner quadrants, or overlapping lesions associated with higher-than-expected mortality, whereas nipple tumors demonstrated outcomes close to expected. Stage at diagnosis remained one of the strongest prognostic indicators; patients with localized disease experienced substantially better-than-expected survival, whereas distant-stage disease was associated with dramatically higher mortality, emphasizing the importance of early detection.

Treatment-related factors were critical in modifying survival risk. Receipt of radiation therapy, chemotherapy, and timely initiation of treatment significantly influenced outcomes. Patients who refused recommended radiation therapy or experienced treatment delays beyond 120 days exhibited the worst survival, underscoring the importance of treatment adherence and rapid initiation. Chemotherapy administration, although often reserved for higher-risk patients, conferred survival benefits, as evidenced by observed outcomes relative to expected deaths.

Tumor characteristics and history further refined prognostic stratification. Prior benign or borderline tumors were modestly associated with improved survival, likely reflecting increased surveillance and earlier detection in these patients. Conversely, multiple prior in situ or malignant tumors were linked to incrementally poorer survival. Tumor size emerged as a particularly strong predictor, with smaller tumors ( $\leq 10$  mm) associated with substantially better survival, while large tumors ( $> 20$  mm) demonstrated markedly higher mortality, highlighting tumor burden as a key determinant of outcome.

Finally, temporal trends in survival suggested gradual improvements over the study period. Patients diagnosed in later years, particularly 2009-2016, exhibited lower observed mortality than expected, consistent with advances in early detection, screening, therapeutic strategies, and supportive care. These findings collectively illustrate the multifactorial determinants of breast cancer survival, emphasizing the interplay between patient demographics, tumor biology, treatment, socioeconomic context, and temporal improvements in cancer care. The results highlight the necessity of individualized management strategies, equitable access to timely and effective treatment, and the continued importance of early detection programs in improving breast cancer outcomes.

## 4.4. Cox Proportional Hazards Model Results

### 4.4.1 2000-2021 Diagnoses

A multivariable Cox proportional hazards regression was fitted to evaluate the association between clinical, demographic, and socioeconomic variables and breast cancer-specific survival in the SEER cohort ( $n = 915,629$ ; number of events = 162,653). The results highlight several key prognostic factors, spanning tumor characteristics, treatment modalities, biological markers, and sociodemographic influences.

Sex was a significant predictor of survival. Male patients exhibited a 3% higher hazard of death compared to females ( $HR = 1.31$ , 95% CI: 1.25-1.38,  $p < 0.001$ ). This aligns with prior evidence that male breast cancer, though rare, is often diagnosed at a later stage and carries a worse prognosis. Age at diagnosis showed a small but highly significant effect ( $HR = 1.02$  per year,  $p < 0.001$ ), indicating progressively worse outcomes with increasing age. Year of diagnosis was inversely associated with mortality ( $HR = 0.99$  per year,  $p < 0.001$ ), suggesting steady improvements in survival outcomes over the 2000-2016 period, likely reflecting advances in screening, therapies, and care delivery.

Tumor primary site within the breast was associated with differential survival. Compared with tumors of the nipple (reference category), lesions located in the lower-inner quadrant ( $HR = 1.31$ ), lower-outer quadrant ( $HR = 1.30$ ), axillary tail ( $HR = 1.33$ ), and breast NOS ( $HR = 1.40$ ) conferred significantly higher hazards of death. Tumor stage had the strongest impact. Patients with regional disease had a hazard nearly 2.5 times higher than those with localized disease ( $HR = 2.49$ ,  $p < 0.001$ ), while those with distant disease had a more than tenfold higher hazard ( $HR = 10.23$ ,  $p < 0.001$ ). Tumor size was a continuous predictor, with increasing size associated with worse survival ( $HR = 1.02$ ,

$p < 0.001$ ). Importantly, extensive tumors (`tsize_extensive_flag`) were associated with a markedly reduced hazard ( $HR = 0.18$ ), which likely reflects model coding nuances and should be interpreted cautiously. Conversely, missing tumor size information was associated with substantially worse survival ( $HR = 1.67$ ), suggesting either more aggressive disease or incomplete documentation.

Breast cancer molecular subtype had strong prognostic significance. Compared with triple-negative (reference group), patients with HR+/HER2- tumors demonstrated a 26% reduction in hazard ( $HR = 0.74$ ,  $p < 0.001$ ). By contrast, HR+/HER2+ tumors were associated with worse outcomes ( $HR = 1.26$ ,  $p < 0.001$ ). HR-/HER2+ tumors showed no significant survival difference relative to triple-negative. ER and PR receptor status followed expected patterns. ER+ ( $HR = 0.85$ ,  $p < 0.001$ ) and PR+ ( $HR = 0.82$ ,  $p < 0.001$ ) were both protective, while negative receptor status increased mortality risk. HER2 positivity (when defined independently) was associated with a protective effect ( $HR = 0.61$ ,  $p < 0.001$ ), reflecting the transformative impact of HER2-targeted therapies in the modern treatment era.

Chemotherapy was associated with a modest but significant increase in hazard ( $HR = 1.11$ ,  $p < 0.001$ ). This finding likely reflects confounding by indication, as patients with more advanced or aggressive disease were more likely to receive chemotherapy. Radiation therapy showed heterogeneous effects. Refusal of radiation was strongly associated with worse survival ( $HR = 1.96$ ,  $p < 0.001$ ) and patients with “none/unknown” radiation also had poorer outcomes ( $HR = 1.32$ ,  $p < 0.001$ ). Conversely, brachytherapy was protective ( $HR = 0.73$ ,  $p < 0.001$ ). Time from diagnosis to treatment did not show a statistically significant association with survival, suggesting that modest variations in treatment initiation timing did not substantially influence long-term outcomes once other factors were accounted for.

Median household income was inversely associated with hazard of death, since, patients residing in higher-income areas had progressively better survival, with hazard ratios decreasing steadily across income quintiles. For example, those in the  $\geq \$120,000$  bracket experienced a 33% reduction in hazard compared to patients from areas with  $< \$40,000$  median income ( $HR = 0.67$ ,  $p < 0.001$ ). Marital status also exerted an influence, and it was shown that being married was protective ( $HR = 0.83$ ,  $p < 0.001$ ), consistent with social support advantages documented in oncologic outcomes. By contrast, being widowed ( $HR = 1.09$ ,  $p < 0.001$ ) or separated ( $HR = 1.09$ ,  $p < 0.001$ ) was associated with

worse survival. Race/ethnicity displayed marked disparities, and when compared to non-Hispanic Whites, non-Hispanic Black patients experienced a 25% higher hazard of death (HR = 1.25,  $p < 0.001$ ). Conversely, Asian or Pacific Islander patients had better survival (HR = 0.85,  $p < 0.001$ ). American Indian/Alaska Native patients did not differ significantly, while “unknown race” was associated with substantially worse outcomes (HR = 0.29,  $p < 0.001$ ), likely reflecting misclassification. Rural-urban status showed modest effects, as living outside metropolitan areas was associated with slightly worse survival, although effect sizes were small (HR  $\approx$  0.93-0.94). Missing rural-urban information, however, was linked with higher hazard (HR = 1.28,  $p = 0.003$ ).

#### 4.4.2 2010-2021 Diagnoses

A Cox proportional hazards regression restricted to patients diagnosed from 2010 onwards was performed ( $n = 520,859$ ; number of events = 71,432). This time frame was chosen because reliable data on molecular subtype and derived HER2 receptor status became available only after 2010. The restricted analysis not only provides more detailed biological insights but also allows a comparison with the full-period model (2000-2021).

The effect of sex remained strong and males had a 26% higher hazard compared to females (HR = 1.26, 95% CI: 1.17-1.35,  $p < 0.001$ ). While still elevated, this hazard ratio was slightly lower than in the full-period model (HR = 1.31), suggesting a relative improvement in male survival in the modern era, likely due to better diagnostic awareness and access to systemic therapies. Age at diagnosis again showed a small but consistent effect (HR = 1.02,  $p < 0.001$ ), similar to the earlier model. Year of diagnosis continued to be protective (HR = 0.95,  $p < 0.001$ ), with a stronger effect size than in the full-period model (HR = 0.99). This reflects sharper survival improvements in the 2010-2021 decade, aligning with therapeutic advances in HER2-targeted agents, endocrine therapies, and immunotherapies.

The primary tumor site associations largely mirrored those of the full dataset with tumors in the breast NOS category remaining strongly adverse (HR = 1.46, vs. HR = 1.40 previously), and quadrant-specific variations persisted but were less pronounced compared to the earlier model. Stage at diagnosis was again the most powerful determinant of survival with regional stage carrying a hazard nearly 2.8 times higher than localized disease (HR = 2.76 vs. 2.49 in the full period), while distant stage had an elevenfold higher hazard

(HR = 11.26 vs. 10.23). These slightly higher HRs suggest that, despite treatment advances, late-stage presentation continues to dramatically worsen prognosis. Tumor size maintained its predictive value (HR = 1.02,  $p < 0.001$ ). However, unlike the earlier model, missing tumor size data had an even greater negative impact (HR = 2.01 vs. 1.67 previously), highlighting the importance of complete documentation in the modern registry era.

The inclusion of molecular subtype data (available from 2010) sharpened prognostic differentiation. Compared with triple-negative tumors, HR+/HER2- tumors had a 40% reduction in hazard (HR = 0.60,  $p < 0.001$ ), a stronger protective effect than in the full model (HR = 0.74), HR+/HER2+ tumors carried a 15% higher hazard (HR = 1.15,  $p < 0.001$ ), milder than the earlier estimate (HR = 1.26), and, finally, HR-/HER2+ tumors had significantly worse survival (HR = 1.27,  $p < 0.001$ ), whereas this group had shown no significant difference in the full-period model. These differences likely reflect both biological heterogeneity and evolving treatment landscapes. The strong protective effect of HR+/HER2- cancers highlights the effectiveness of endocrine therapy. The improved outcomes in HER2+ cancers relative to the past emphasize the transformative role of newer HER2-directed agents introduced after 2010. Independent receptor status analyses were consistent with ER+ (HR = 0.84) and PR+ (HR = 0.83) remaining protective, while HER2 positivity continued to reduce hazard (HR = 0.63), confirming the benefit of HER2-targeted therapy in the modern treatment period.

Chemotherapy again appeared associated with higher mortality (HR = 1.14,  $p < 0.001$  vs. 1.11 previously), reflecting treatment confounding by disease severity and radiation therapy patterns were largely similar to the full-period model. Refusal remained strongly adverse (HR = 1.72 vs. 1.96 previously), while brachytherapy was protective (HR = 0.69 vs. 0.73). The effect sizes were slightly attenuated compared with the earlier cohort, suggesting more uniform radiation benefits in the modern treatment landscape. Finally, time from diagnosis to treatment remained nonsignificant, consistent with earlier findings.

Given socioeconomic gradients, higher median household income persisted as protective, with patients from the highest bracket ( $\geq \$120,000$ ) experiencing a 41% reduction in hazard (HR = 0.59,  $p < 0.001$ ), an even stronger gradient than in the full-period model (HR = 0.67). Marital status continued to influence survival since being married was, once again, protective (HR = 0.84, similar to before), while being widowed (HR = 1.11) or separated (HR = 1.11) remained detrimental. With race/ethnicity disparities persisting, black

patients experienced worse survival (HR = 1.24, nearly identical to 1.25 previously), while Asian/Pacific Islanders retained their survival advantage (HR = 0.83 vs. 0.85). These consistent disparities across time underscore structural inequities in access and care. Rural-urban status exerted minimal influence, with hazard ratios close to 1 and missing geographic information carried, again, higher risk (HR = 1.21 vs. 1.28 previously).

#### 4.4.3 Proportional Hazards Assumptions

In summary, the protective effect of HR+/HER2- cancers was stronger in the post-2010 cohort, while HR-/HER2+ cancers emerged as a high-risk group, unlike in the full-period model. It was also possible to observe the impact of HER2-targeted therapy, as HER2+ cases showed relatively better survival rates in 2010 and beyond, consistent with the widespread use of targeted therapies. Late-stage diagnosis still hinders therapeutic progress, as evidenced by the fact that hazard ratios for both regional and distant disease increased when analyzing the cohort only since 2010. Similar to this, socioeconomic disparities also widened as income-related gradients became even steeper after 2010, with high-income groups benefiting disproportionately from modern therapies. On the other hand, sex disparities narrowed slightly and the male-female survival gap, while still large, diminished compared to earlier years, reflecting improved awareness and management of male breast cancer. Finally, restricting analyses to 2010-2021 reinforced the dominant prognostic importance of tumor stage and molecular subtype, while highlighting the growing influence of HER2-directed therapy and the persistent role of socioeconomic and racial disparities. Compared with the 2000-2021 model, subtype-based prognostic became more precise, HER2+ outcomes improved, and survival disparities based on income widened, underscoring both the therapeutic progress and ongoing inequities in breast cancer care.

Assessment of the proportional hazards assumption using the `cox.zph()` test revealed that many covariates violated this assumption ( $p < 0.05$ ), indicating that their effects on the hazard of death vary over time, and only the missing-data flags for stage, marital status, and income satisfied the assumption ( $p > 0.05$ ). Since the global test was highly significant ( $\chi^2 = 6280$ ,  $df = 75$ ,  $p < 2e-16$ ), this suggests that the proportional hazards assumption does not hold for the model overall.

While the proportional hazards assumption was not met for most covariates, the model nonetheless provides important insights into factors influencing breast cancer survival. The violations indicate that the impact of several predictors on survival is time-dependent,

reflecting dynamic clinical and demographic influences over the course of follow-up and, for instance, treatment effects or tumor progression may change hazard rates at different time points. These findings underscore the complexity of breast cancer prognosis and suggest that future analyses could incorporate time-dependent covariates or stratified Cox models to capture these temporal effects more precisely. Importantly, despite these assumption violations, the identified covariates remain highly significant predictors, highlighting their relevance for understanding survival patterns in the studied population.

## 4.5. Machine Learning Models

### Logistic Regression Model

The logistic regression model achieved an overall accuracy of 81.5% (95% CI: 81.3-81.7). This value was lower than the No Information Rate (NIR = 88.6%), which represents the accuracy of always predicting the majority class (survival). This apparent reduction is expected in imbalanced datasets, as accuracy disproportionately favours the majority class and penalizes attempts to correctly identify minority cases (non-survivors). Therefore, accuracy is not an appropriate measure of model performance in this context. Instead, metrics that account for class imbalance provide a more informative evaluation.

From the confusion matrix, sensitivity (true positive rate) was 82.8%, while specificity (true negative rate) was 71.8%, with a balanced accuracy of 77.3%. The positive predictive value (PPV) was high (95.8%), but the negative predictive value (NPV) was low (34.9%), reflecting difficulty in correctly identifying patients who did not survive. McNemar's test ( $p < 2e-16$ ) further indicated significant misclassification asymmetry, with more false negatives than false positives.

Variable importance analysis identified several predictors as particularly influential. The most important features included tumor size over time (importance = 19,835.7), age at diagnosis (11,210.2), time from diagnosis to treatment (10,696.2), summary stage (8,344.1), and median household income (inflation-adjusted to 2022) (7,358.9). Other relevant factors included year of diagnosis (7,116.9), combined summary stage (7,940.7), primary site (5,458.6), radiation treatment (2,549.5), and hormone receptor status (ER, PR, and HER2 recodes). Demographic and socioeconomic factors such as race/ethnicity (2,842.9) and marital status at diagnosis (3,392.4) also contributed meaningfully to the model.

The model's kappa statistic was 0.37, indicating only fair agreement beyond chance. This reflects the challenge of imbalanced classification, where predicting minority outcomes is

more difficult. However, discrimination performance was assessed using receiver operating characteristic (ROC) and precision-recall (PR) analyses with the model achieving an ROC AUC of 0.855, indicating good ability to distinguish between survivors and non-survivors. The PR AUC was notably high at 0.974, reflecting strong predictive performance in the context of class imbalance, suggesting it was effective at distinguishing between survivors and non-survivors despite limitations in raw accuracy.

In summary, while logistic regression provided good discriminatory performance (ROC AUC = 0.855, PR AUC = 0.974), it struggled with imbalanced classification, particularly in predicting non-survivors (low NPV). Altogether, the logistic regression model identified prognostic factors that are consistent with established clinical knowledge, tumor burden, disease stage, treatment timeliness, therapeutic modality, and hormone receptor status, while also highlighting the importance of social determinants of health, such as income, marital status, and race/ethnicity. This underscores the multifactorial nature of breast cancer survival, where biological, clinical, and socio-demographic variables jointly shape patient outcomes.

### **Random Forest Model**

The Random Forest model achieved an overall accuracy of 90.8% (95% CI: 90.7-90.9), which was significantly higher than the No Information Rate (NIR = 88.6%;  $p < 2.2e-16$ ). Unlike logistic regression, this indicates that the Random Forest classifier improved upon simply predicting the majority class. The kappa statistic was 0.42, reflecting a moderate level of agreement beyond chance.

From the confusion matrix, the model showed very high sensitivity (98.1%), indicating strong ability to correctly identify patients who survived. However, specificity was limited (34.7%), showing difficulty in detecting patients who did not survive. This trade-off resulted in a balanced accuracy of 66.4%. The positive predictive value (PPV) was strong (92.1%), meaning that predicted survivors were generally correct, while the negative predictive value (NPV) was moderate (69.8%), reflecting some success in identifying non-survivors, but still with room for improvement. McNemar's test ( $p < 2.2e-16$ ) again revealed significant asymmetry in misclassifications, largely driven by the imbalance between survivors and non-survivors.

The model achieved an ROC AUC of 0.863, slightly higher than logistic regression, and a PR AUC of 0.977, which confirms excellent performance under class imbalance. These

results suggest that the Random Forest model effectively captured non-linear patterns in the data, improving overall classification compared to logistic regression.

Variable importance analysis highlighted similar prognostic factors to those identified in logistic regression. The most influential predictors included tumor size over time (19,835.7), age at diagnosis (11,210.2), and time from diagnosis to treatment initiation (10,696.2), underscoring the significance of tumor burden, patient age, and timely treatment in survival outcomes. Disease staging variables, including summary stage (8,344.1) and combined summary stage (7,940.7), also ranked highly, confirming their established prognostic value. Socio-demographic indicators such as median household income (7,358.9), race/ethnicity (2,842.9), and marital status (3,392.4) contributed meaningfully, suggesting survival disparities linked to socioeconomic context and social support. Treatment-related variables, including radiation therapy (2,549.5), chemotherapy (1,482.9), and hormone receptor status (ER, PR, HER2), were likewise important, consistent with their clinical role in therapeutic decision-making.

In summary, the Random Forest model outperformed logistic regression in overall accuracy and discrimination, particularly by leveraging non-linear relationships among prognostic factors. While it still struggled with specificity, its high sensitivity and PR AUC underscore its potential as a reliable predictive model for breast cancer survival in the presence of class imbalance.

### **XGBoost**

The XGBoost classifier achieved an overall accuracy of 80.3% (95% CI: 80.1-80.5). Similar to logistic regression, this was lower than the No Information Rate (NIR = 88.6%;  $p = 1$ ), reflecting the limitations of accuracy in the context of severe class imbalance. The kappa statistic was 0.37, again indicating fair agreement beyond chance.

The model demonstrated a balanced performance between sensitivity and specificity. Sensitivity was 80.8%, showing that XGBoost effectively identified the majority of survivors, while specificity was 76.3%, considerably higher than that of the Random Forest model. This resulted in a balanced accuracy of 78.6%, the highest among the models evaluated. The positive predictive value (PPV) was high (96.4%), indicating reliable identification of survivors, whereas the negative predictive value (NPV) remained low (33.9%), highlighting persistent challenges in detecting non-survivors. McNemar's test ( $p < 2e-16$ ) again indicated significant asymmetry in misclassifications.

In terms of discrimination, XGBoost achieved an ROC AUC of 0.870, the highest of all models, and a PR AUC of 0.978, reflecting excellent predictive ability under class imbalance. These metrics suggest that XGBoost provided the strongest overall discriminatory performance in this study, despite its lower raw accuracy compared to Random Forest.

Variable importance analysis revealed patterns consistent with the other models. The top predictors included tumor size over time (19,835.7), age at diagnosis (11,210.2), and time from diagnosis to treatment initiation (10,696.2), all of which are well-established prognostic factors in breast cancer survival. Staging variables, including summary stage (8,344.1) and combined summary stage (7,940.7), further underscored the importance of disease extent at presentation. Socio-demographic predictors such as median household income (7,358.9), race/ethnicity (2,842.9), and marital status at diagnosis (3,392.4) contributed significantly, supporting the role of socioeconomic disparities in survival outcomes. Treatment-related factors, such as radiation therapy (2,549.5), chemotherapy (1,482.9), and hormone receptor status (ER, PR, HER2), also emerged as important features.

In summary, XGBoost provided the best balance between sensitivity and specificity, with the highest ROC AUC (0.870) and PR AUC (0.978) across all models tested. While accuracy remained below the NIR due to class imbalance, the strong discriminative performance and identification of key prognostic factors reinforce XGBoost as the most robust predictive model in this analysis.

While both Random Forest and XGBoost are ensemble tree-based methods, their underlying learning strategies differ in ways that explain the superior performance of XGBoost in this study. Random Forest relies on bagging, where multiple decision trees are trained independently on random subsets of data and features, and their predictions are aggregated based on a majority vote. This reduces variance and improves generalizability but does not directly optimize for misclassified cases or class imbalance. As a result, Random Forest achieved high sensitivity (98.1%) but poor specificity (34.7%), indicating a tendency to over-predict the majority class (survivors). In contrast, XGBoost is based on gradient boosting, which builds trees sequentially, with each new tree focusing on the residual errors of the previous ones. This approach allows the model to iteratively correct mistakes, particularly improving classification of harder-to-predict cases such as non-survivors. The use of a gradient-based optimization framework, regularization, and weighted handling of errors helps XGBoost strike a better balance between sensitivity (80.8%) and specificity

(76.3%). This balance translated into the highest balanced accuracy (78.6%), ROC AUC (0.870), and PR AUC (0.978) of all models tested. Taken together, these theoretical advantages explain why XGBoost achieved superior discriminatory performance, despite having a lower overall accuracy than Random Forest. Random Forest maximized correct predictions for the majority class but failed to generalize well for minority outcomes, whereas XGBoost achieved a better trade-off, making it the most robust and clinically useful model in this study. It is yet important to note that RF and XGBoost might have performed even better when trained with their native handling of categorical variables, but it was chosen a uniform preprocessing pipeline to ensure fair comparisons.

## 4.6. Breast Cancer Data Explorer

The home page of the Breast Cancer Data Explorer constitutes the central entry point to the application. It was designed to introduce the dataset, present key summary information, and provide a structured overview of the analytical workflow implemented in the tool with a layout that emphasizes clarity and usability, ensuring that users can engage with the application regardless of their level of statistical expertise.

At the top of the page, as it can be seen in Figure 4.30 and Figure 4.31, a jumbotron highlights the title of the tool, its scope, and the covered time frame, while also offering direct access to the official SEER website for further information about the data source. Immediately below, a set of summary value boxes displays essential dataset characteristics, namely the total number of breast cancer cases, the number of variables available, and the temporal coverage (2000-2021). These indicators provide a first orientation regarding the size and scope of the data under analysis. The descriptive content is organized into interactive components that can be expanded or collapsed according to the user's needs. A map of SEER registry coverage illustrates the geographical distribution of contributing registries across the United States, which can be explored interactively and maximized to facilitate detailed inspection. Alongside this, a line plot depicts the annual distribution of breast cancer diagnoses, enabling the rapid identification of temporal trends across the two decades of observation. In addition, dynamic value boxes are generated according to the type of each variable, displaying the most frequent category in the case of categorical data or the median value when variables are numeric.

To provide orientation for the analytical process, an analysis pipeline is also presented on the home page. This component not only summarizes the sequential structure of the application, beginning with dataset description and progressing through exploratory analysis,

survival analysis, prognostic factor assessment, prediction, and concluding with decision-support functionalities, but it also offers users a clear overview of the available modules and a logical order in which they can be employed. Complementing this, a dedicated section entitled 'How to Use This App' outlines the purpose of each module in concise terms, thereby facilitating navigation and promoting knowledge as well as effective use of the tool.

Finally, the home page includes an acknowledgments section that situates the application within its academic context, specifying the goals of the project, acknowledging the SEER Program as the data source, and identifying the institutional setting in which the tool was developed, namely as part of a Master's thesis in Bioinformatics and Computational Biology at Faculty of Sciences of the University of Porto (FCUP). Contact information is also provided to enable feedback and future communication. As one of the main goals of this tool, the design of the main page prioritizes adaptability and ease of use. To honor that, all descriptive cards can be collapsed to minimize visual clutter, while the map and plots can be expanded to full-screen mode when a more detailed inspection is required. This flexibility allows the interface to accommodate both users seeking a high-level overview and those interested in conducting a more detailed examination of the data.

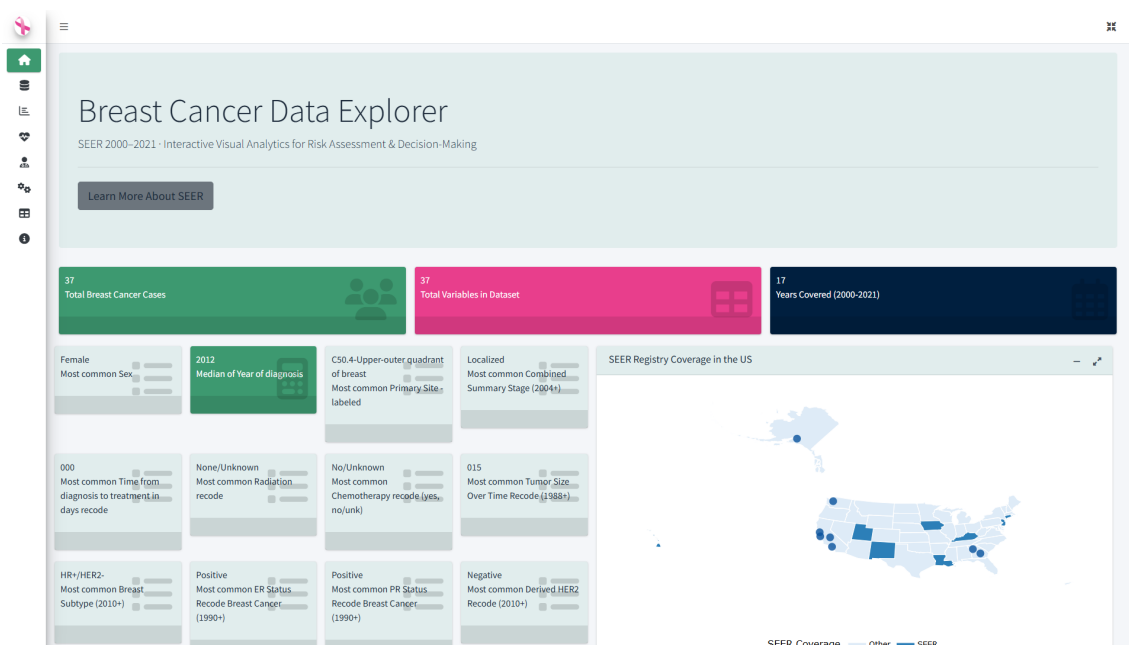


Figure 4.30: **Home page of the Breast Cancer Data Explorer.** The interface introduces the SEER dataset, summarizes key characteristics, and provides interactive components such as the registry coverage map and diagnosis trends over time. It also presents the analysis pipeline, usage instructions, and acknowledgments, offering users a clear overview of the tool's scope and functionality.

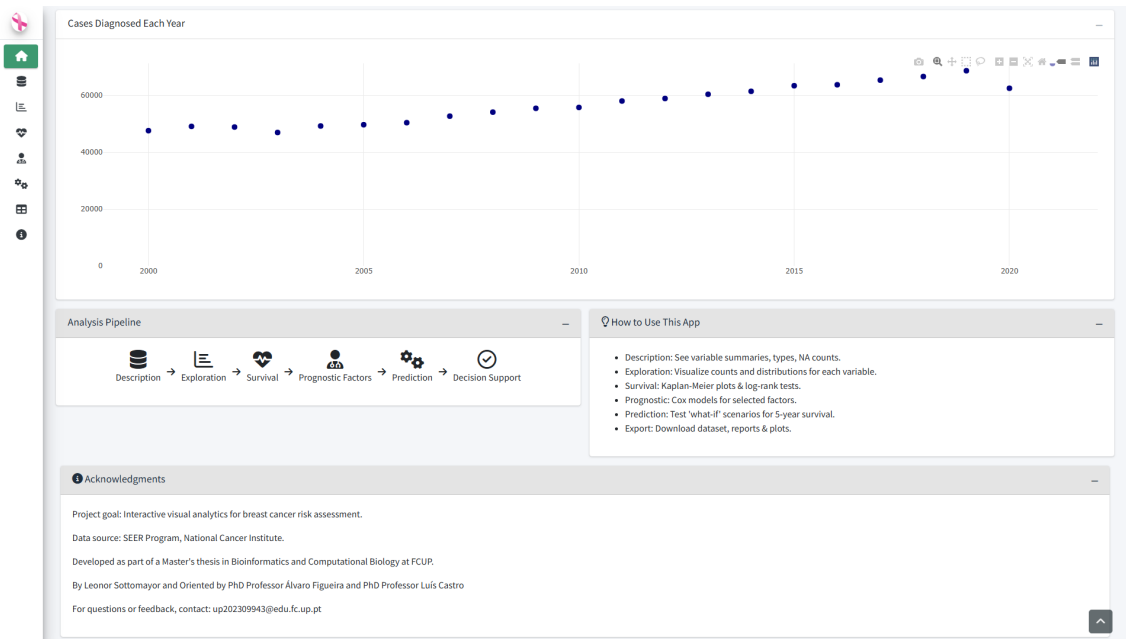


Figure 4.31: **Home page of the Breast Cancer Data Explorer (2).** The interface introduces the SEER dataset, summarizes key characteristics, and provides interactive components such as the registry coverage map and diagnosis trends over time. It also presents the analysis pipeline, usage instructions, and acknowledgments, offering users a clear overview of the tool's scope and functionality.

The Breast Cancer Data Explorer's sidebar is used as the app's primary navigation system, enabling users to move seamlessly between the different analytical modules. It is designed in a collapsible format to optimize screen space while maintaining ease of access to the available sections. The sidebar (Figure 4.32) includes entries for the Home page, Dataset Description, Exploratory Analysis, Survival Analysis, Prognostic Factor Analysis, Prediction, Data, and About/Help. This structure reflects the logical progression of the analytical workflow, guiding users from initial dataset characterization through exploratory and inferential analyses, and ultimately to prediction and decision-support functionalities. By organizing the modules in a coherent and sequential manner, the sidebar ensures intuitive navigation and contributes to a smooth user experience.

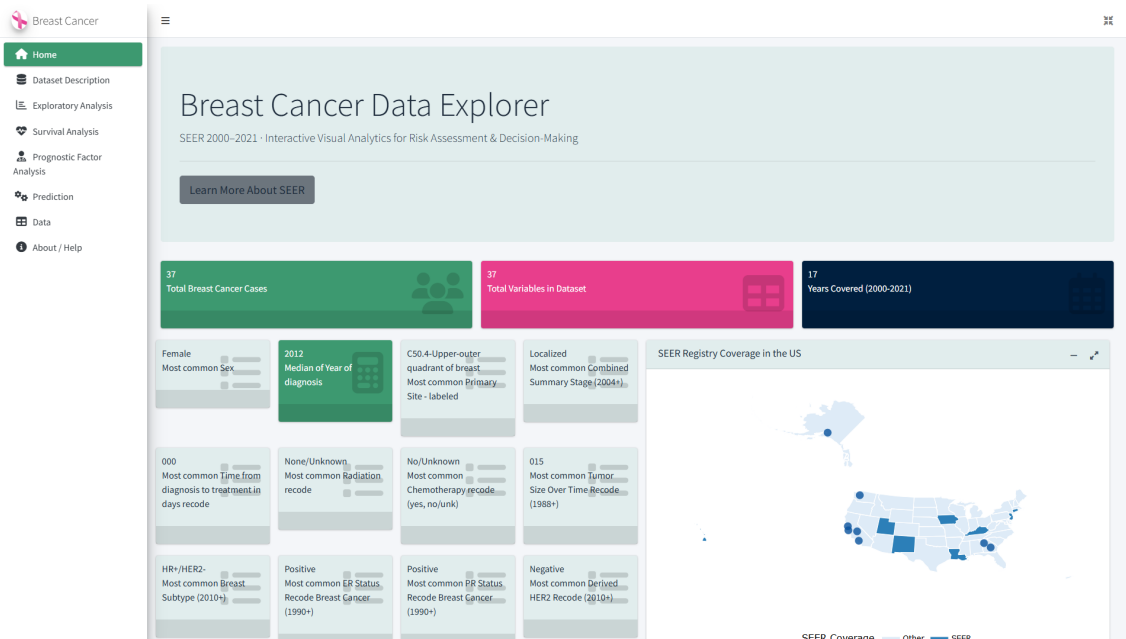


Figure 4.32: **Sidebar navigation menu of the Breast Cancer Data Explorer.** The sidebar provides access to all major modules of the application, including the Home page, Dataset Description, Exploratory Analysis, Survival Analysis, Prognostic Factor Analysis, Prediction, Data, and About/Help sections. Its collapsible design ensures intuitive navigation while maintaining a clean interface.

The dataset description tab (Figure 4.33) provides an overview of the raw data prior to preprocessing, summarizing its fundamental characteristics. At the top of the page, three information boxes report the total number of entries, the number of variables under study, and the total number of missing values. To facilitate navigation, the interface allows users to search for a specific variable by name or to sort variables alphabetically, as well as by ascending or descending number of missing values. Each variable is displayed in an interactive card containing key information, including its type (numeric or categorical), the number and percentage of missing values, and the number of unique values observed. For all variables, the six most frequent categories are visualized, with all remaining values grouped into an “Others” category, and for numeric variables, descriptive statistics such as the minimum, maximum, mean, median, and standard deviation are provided. Each card also contains an information button that, when activated, displays a modal dialog (Figure 4.34) with the variable’s definition, enabling users to better understand its meaning and relevance. Consistent with the overall design of the application, variable cards can be collapsed to reduce visual clutter or expanded to full-screen mode for more detailed inspection. This combination of summary statistics, search and sorting functionality, and interactive visualization ensures that the tab accommodates both users seeking a high-level overview of the dataset and those requiring a more in-depth exploration of individual variables.

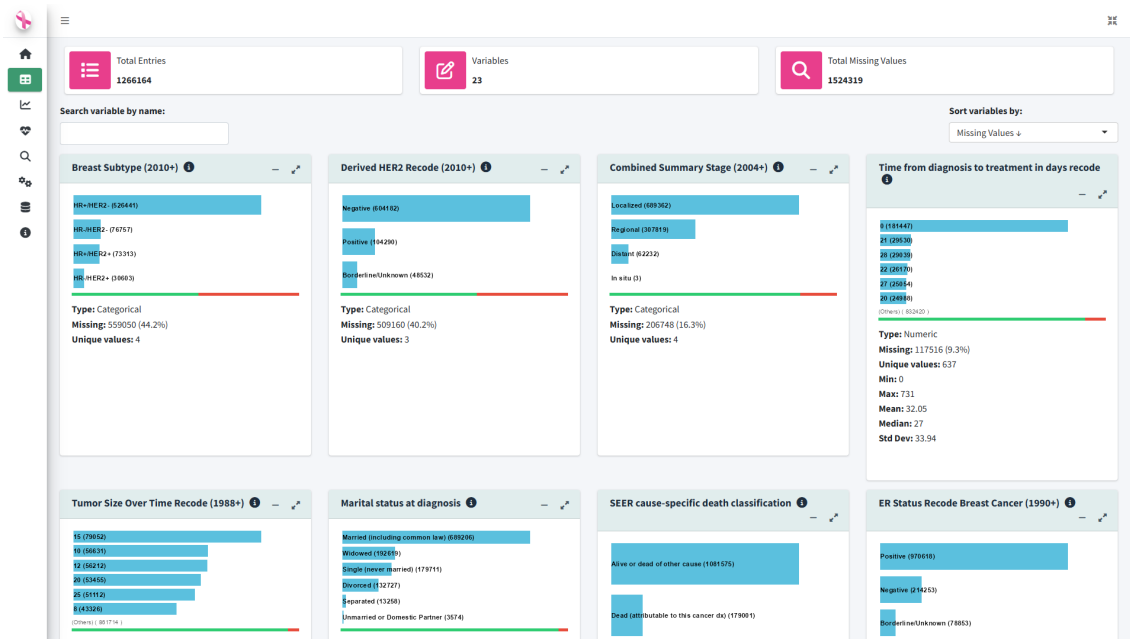


Figure 4.33: **Dataset Description tab of the Breast Cancer Data Explorer.** This section summarizes the dataset before preprocessing, displaying the number of entries, variables, and missing values. Users can search or sort variables, and each variable card provides type, missingness, unique values, and descriptive statistics, with options to collapse or expand for detailed inspection.

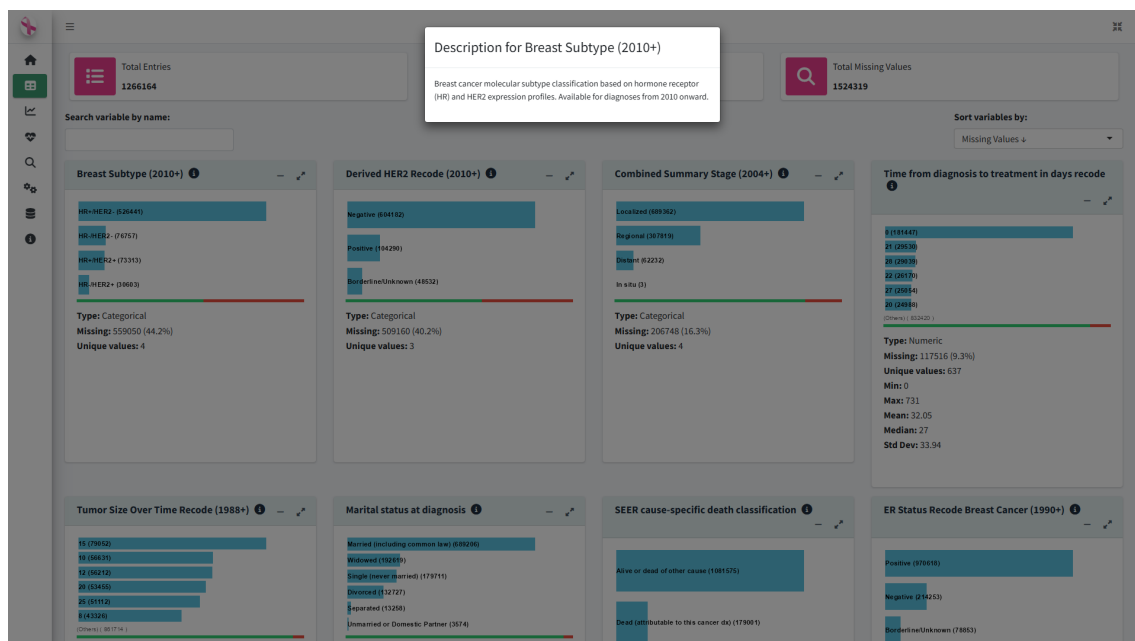


Figure 4.34: **Dataset Description tab of the Breast Cancer Data Explorer.** This section summarizes the dataset before preprocessing, displaying the number of entries, variables, and missing values. Users can search or sort variables, and each variable card provides type, missingness, unique values, and descriptive statistics, with options to collapse or expand for detailed inspection.

Upon navigating to the Exploratory Analysis tab of the Breast Cancer Data Explorer (Figure 4.35), users are presented with a series of sortable boxes, each corresponding to a different variable within the preprocessed dataset. The preprocessing applied, as described in the Methods section, ensures that the data are standardized, cleaned, and ready for

visualization. Each variable is represented using a plot type chosen to optimize clarity and accessibility, including pie charts for categorical distributions with a small amount of unique values, bar charts for the other categorical variables, as well as grouped counts, line plots for temporal trends, histograms for continuous variables, and a word cloud for textual data. This variety of simple visualizations supports an intuitive interpretation of the data, particularly for users without advanced statistical training.

Interactive controls within each box enhance the flexibility of exploration, with users being able to filter data by gender, compare distributions between males and females, or combine them to observe overall values. For numeric variables, an additional toggle allows users to switch between viewing continuous numeric data and grouped categorical representations, consistent with the binning strategy described in the Methods. All plots are interactive, allowing users to hover over data points or bars to view the exact count or value for each category, which improves interpretability and facilitates detailed inspection of the distributions.

In addition, the sortable design of the boxes allows users to rearrange the order in which variables are displayed, facilitating side-by-side comparison of related metrics. Consistent with the overall interface design, each box can be collapsed to reduce visual clutter or expanded to full-screen mode for detailed examination. This combination of interactive controls, dynamic plots, and flexible layout ensures that the Exploratory Analysis tab caters to both users seeking a high-level overview of the dataset and those conducting a more detailed variable-by-variable investigation. By providing an accessible yet comprehensive exploration environment, this tab forms a critical component of the application, bridging raw data inspection and subsequent statistical or predictive analyses.

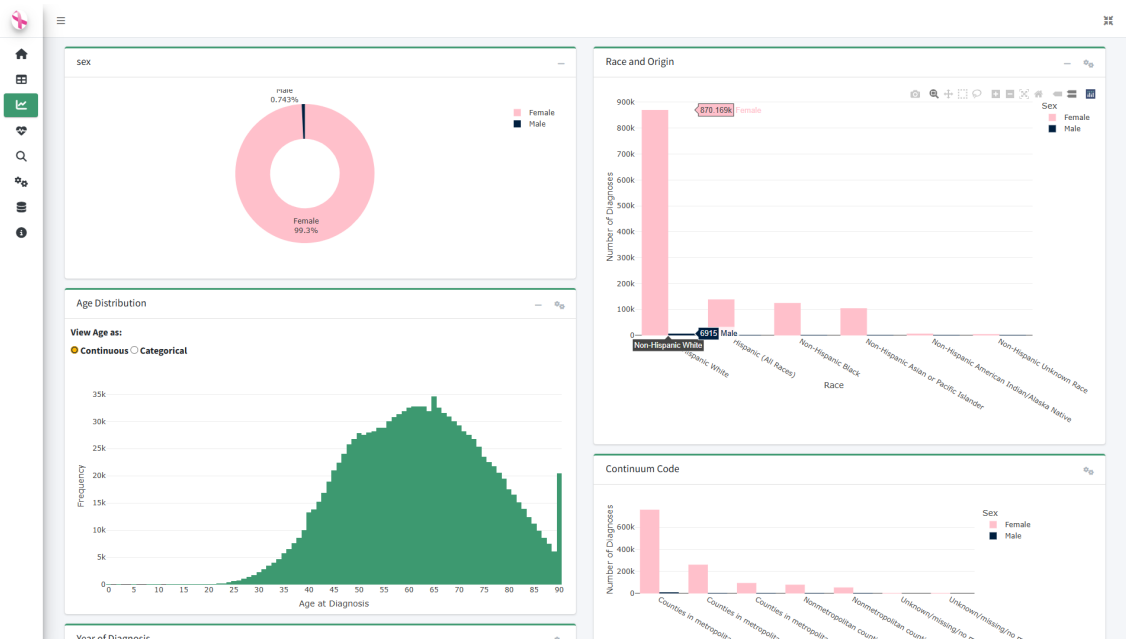


Figure 4.35: **Exploratory Analysis tab of the Breast Cancer Data Explorer.** The tab displays each variable of the preprocessed dataset in interactive, sortable boxes. Users can visualize distributions and trends through bar charts, line plots, pie charts, histograms, and word clouds, filter by gender, or toggle between numeric and categorical representations for numeric variables. All plots are interactive, allowing users to hover over elements to view exact counts or values. Boxes can be collapsed or expanded for detailed inspection, enabling both high-level overviews and in-depth exploration.

The Survival Analysis tab of the Breast Cancer Data Explorer enables users to explore patient survival outcomes interactively through Kaplan-Meier curves and associated risk tables, as can be observed in Figure 4.36. At the top of the page, users can select whether to display overall survival or cause-specific survival, and stratify the analysis by any categorical variable available in the dataset. The Kaplan-Meier plot is generated dynamically based on these selections, displaying survival probabilities over time with confidence intervals for each group. The plot is fully interactive, allowing users to hover over any time point to view detailed information, including the group, survival probability, number at risk, and number of events, thereby facilitating precise interpretation of survival trends.

Beneath the Kaplan-Meier plot, an interactive risk table presents the number of patients at risk at standardized time intervals, providing a complementary summary of the underlying cohort dynamics. The log-rank test is performed automatically for the selected stratification variable, with the resulting p-value displayed in the plot subtitle, allowing users to quickly assess statistically significant differences between survival curves.

The tab design emphasizes flexibility and clarity since users can switch between overall and cause-specific survival, select any grouping variable, and inspect detailed survival information directly from the interactive plots. By integrating visual and numerical

summaries, this tab supports both high-level overviews and in-depth analyses of survival outcomes, making it accessible to users with varying levels of statistical expertise.

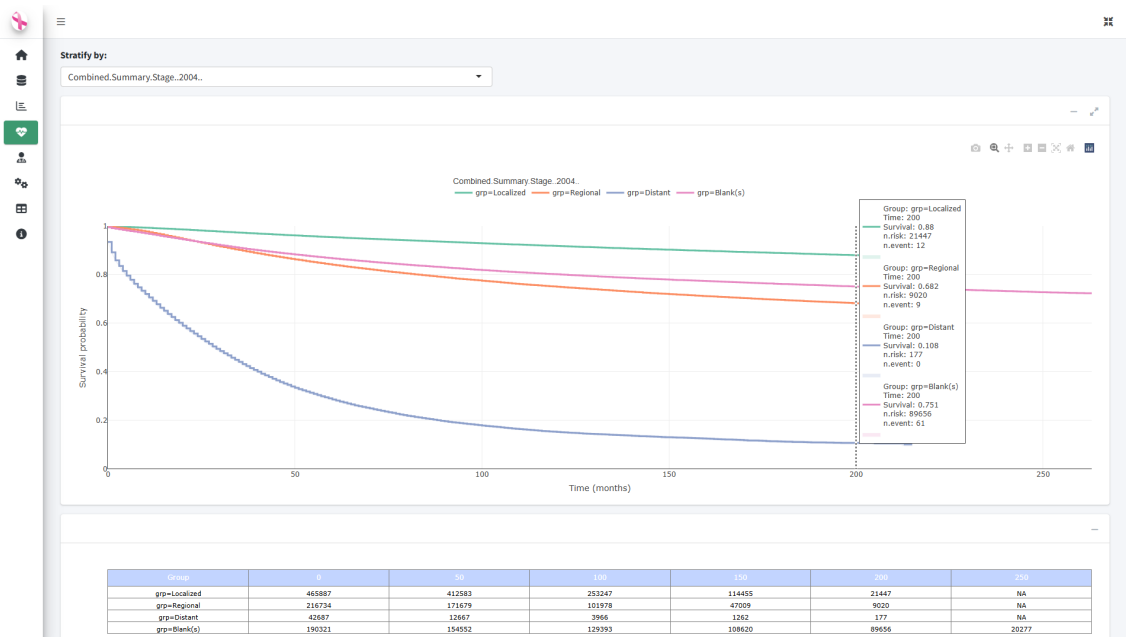


Figure 4.36: **Survival Analysis tab of the Breast Cancer Data Explorer.** Users can explore Kaplan-Meier survival curves and associated risk tables for overall or cause-specific survival. Survival can be stratified by any categorical variable in the dataset. Plots are interactive, allowing users to hover over time points to view group, survival probability, number at risk, and number of events. Log-rank p-values are displayed to assess differences between groups.

To explore the influence of clinical and demographic characteristics on patient survival, the Prognostic Factor Analysis tab provides an interactive interface for fitting and interpreting Cox Proportional Hazards models (Figure 4.37). Users begin by selecting one or more variables from a comprehensive list that includes patient demographics, tumor characteristics, treatment details, and socioeconomic indicators, meaning all variables available in the SEER dataset analyzed here. Upon clicking the “Fit Cox Model” button, the application dynamically generates a multivariable Cox model, allowing users to assess the independent effect of each selected factor on survival outcomes.

The tab presents the results in multiple complementary formats. A summary table displays the hazard ratio, 95% confidence interval, and p-value for each variable, highlighting those factors that are statistically significant predictors of survival. To facilitate intuitive interpretation, a forest plot visualizes these hazard ratios and confidence intervals, providing a clear overview of relative risks across variables. Additionally, the tab generates Cox-adjusted survival curves, which illustrate the predicted survival probabilities for different levels of a selected variable while accounting for the effects of all other covariates.

in the model. These curves include confidence intervals and an accompanying risk table, offering detailed information on the number of patients at risk at specific time points.

The interactive design of the tab allows users to explore complex relationships between multiple prognostic factors without needing advanced statistical expertise. By combining numerical summaries and visualizations, this tab provides both a high-level assessment of important prognostic factors and a detailed, patient-level perspective on survival outcomes. This functionality is particularly valuable for identifying key clinical and demographic predictors that may guide research, risk stratification, and decision-making in breast cancer management.

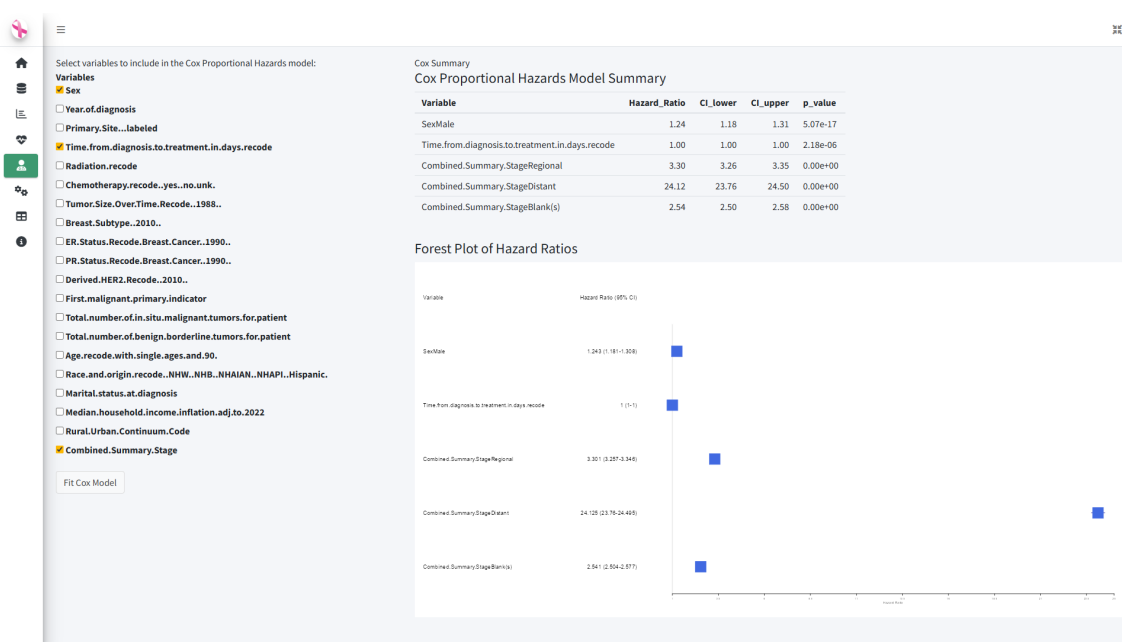
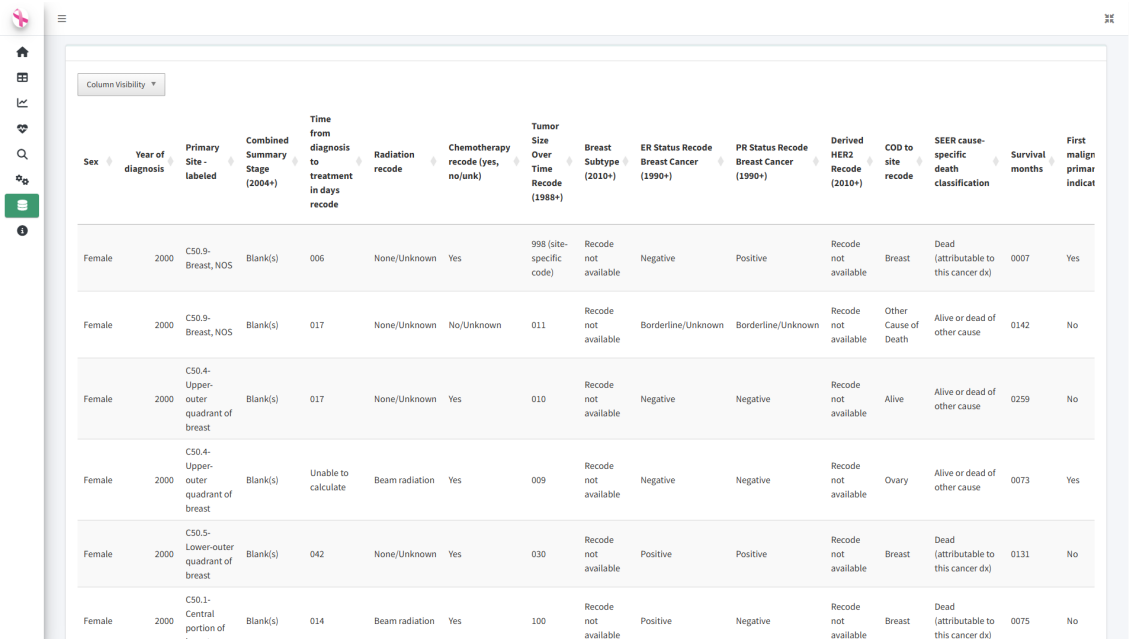


Figure 4.37: **Prognostic Factor Analysis tab of the Breast Cancer Data Explorer.** Users can select variables to include in a Cox Proportional Hazards model to examine their impact on survival. The tab provides a summary table of hazard ratios with confidence intervals and p-values, a forest plot visualizing hazard ratios, and Cox-adjusted survival curves with risk tables. Plots are interactive, allowing users to hover to view exact values and facilitating detailed exploration of the influence of multiple covariates on patient survival.

To provide users with direct access to the underlying SEER dataset, the Data tab offers a flexible and interactive interface for exploring the patient-level records included in this study (Figure 4.38). The tab displays the dataset in a searchable and scrollable table format, with pagination controls that allow users to navigate through the large number of observations efficiently. Columns can be shown or hidden according to user preference, enabling a tailored view that highlights only the variables of interest, whether demographic, clinical, or survival-related.



| Sex    | Year of diagnosis | Primary Site labeled                  | Combined Summary Stage (2004+) | Time from diagnosis to treatment in days recode | Radiation recode | Chemotherapy recode (yes, no/unk) | Tumor Size Over Time Recode (1988+) | Breast Subtype (2010+) | ER Status Recode Breast Cancer (1990+) | PR Status Recode Breast Cancer (1990+) | Derived HER2 Recode (2010+) | COD to site recode   | SEER cause-specific death classification | Survival months | First malign primar indicat |
|--------|-------------------|---------------------------------------|--------------------------------|-------------------------------------------------|------------------|-----------------------------------|-------------------------------------|------------------------|----------------------------------------|----------------------------------------|-----------------------------|----------------------|------------------------------------------|-----------------|-----------------------------|
| Female | 2000              | C50.9- Breast, NOS                    | Blank(s)                       | 006                                             | None/Unknown     | Yes                               | 998 (site-specific code)            | Recode not available   | Negative                               | Positive                               | Recode not available        | Breast               | Dead (attributable to this cancer dx)    | 0007            | Yes                         |
| Female | 2000              | C50.9- Breast, NOS                    | Blank(s)                       | 017                                             | None/Unknown     | No/Unknown                        | 011                                 | Recode not available   | Borderline/Unknown                     | Borderline/Unknown                     | Recode not available        | Other Cause of Death | Alive or dead of other cause             | 0142            | No                          |
| Female | 2000              | C50.4- Upper-outer quadrant of breast | Blank(s)                       | 017                                             | None/Unknown     | Yes                               | 010                                 | Recode not available   | Negative                               | Negative                               | Recode not available        | Alive                | Alive or dead of other cause             | 0259            | No                          |
| Female | 2000              | C50.4- Upper-outer quadrant of breast | Blank(s)                       | Unable to calculate                             | Beam radiation   | Yes                               | 009                                 | Recode not available   | Negative                               | Negative                               | Recode not available        | Ovary                | Alive or dead of other cause             | 0073            | Yes                         |
| Female | 2000              | C50.5- Lower-outer quadrant of breast | Blank(s)                       | 042                                             | None/Unknown     | Yes                               | 030                                 | Recode not available   | Positive                               | Positive                               | Recode not available        | Breast               | Dead (attributable to this cancer dx)    | 0131            | No                          |
| Female | 2000              | C50.1- Central portion of             | Blank(s)                       | 014                                             | Beam radiation   | Yes                               | 100                                 | Recode not available   | Positive                               | Negative                               | Recode not available        | Breast               | Dead (attributable to this cancer dx)    | 0075            | No                          |

Figure 4.38: **Data tab of the Breast Cancer Data Explorer.** The tab presents the SEER dataset in an interactive table format, allowing users to browse, filter, and sort patient-level records. Pagination and horizontal scrolling support navigation across a large number of observations, while column visibility controls enable customization of the displayed variables. Together, these features ensure transparency of the dataset and provide a flexible, user-friendly environment for exploring the variables.

## 4.7. Main Results Summary

The analyses conducted in this study provided a comprehensive characterization of prognostic factors in breast cancer survival and established the foundation for the development of a novel interactive application. Classical survival approaches confirmed the strong prognostic influence of stage at diagnosis, with regional and distant disease associated with markedly higher hazards of death compared to localized disease, while molecular subtype and HER2 status further differentiated prognosis in the modern treatment era. Machine learning extended these insights by offering improved discriminatory performance under class imbalance, with the XGBoost model achieving the highest predictive accuracy and identifying variable importance patterns consistent with the Cox regression findings.

The significance of these analytical results lies not only in the confirmation of known prognostic factors but also in their integration into an accessible and interactive platform. The Breast Cancer Data Explorer was developed to translate statistical analyses and predictive models into a resource that is user-friendly and capable of supporting clinicians, researchers, and policymakers. Unlike static tables or complex code-based analyses, the application allows real-time exploration of descriptive statistics, survival curves, hazard ratios, and individualized predictions through a graphical interface. By embedding both

classical survival modeling and modern machine-learning predictions within the same tool, the application bridges inferential and predictive perspectives, while also addressing a key gap in the literature: the absence of interactive, publicly available platforms that democratize the use of registry-based cancer survival data.

The results confirmed known survival patterns and highlighted persistent socioeconomic disparities. More importantly, they show how we implemented these findings in an interactive tool that improves accessibility, transparency, and interpretability. Therefore, the Breast Cancer Data Explorer is a key outcome of this thesis, combining rigorous analysis with a practical tool for disseminating results.

## 5. Conclusion

This thesis used population-based survival analysis and modern predictive modeling to study breast cancer prognostic factors. Its central contribution is the Breast Cancer Data Explorer, an interactive Shiny application that makes these analyses accessible to users without programming expertise. While survival differences by stage, tumor size, subtype, and sociodemographic factors were robustly confirmed through Cox regression and Kaplan-Meier methods, and while machine-learning models such as XGBoost provided enhanced predictive discrimination, the innovative step was to embed these outputs into a platform that allows non-technical users to interact with the data and models directly.

The first objective, to characterize the study population and assess data quality, was achieved through a detailed descriptive analysis of over two decades of SEER data. Age distributions, stage at diagnosis, tumor size, and molecular subtype patterns were systematically examined, while missingness and outliers were carefully assessed to ensure a sound analytical foundation. This process not only highlighted the heterogeneity of the patient population but also clarified data limitations, such as incomplete molecular subtype availability before 2010, that shaped subsequent analyses.

Secondly, the identification of prognostic variables using statistical analyses was equally met using Kaplan-Meier estimators and Cox regression models. These analyses confirmed well-established survival differences across tumor stage, tumor size, subtype, and sociodemographic variables, while also revealing violations of the proportional hazards assumption that underscored the time-varying effects of certain covariates. These findings both validated prior knowledge and pointed toward methodological opportunities for more flexible modeling frameworks in the future.

The third objective, to develop and compare machine learning models for 5-year survival prediction, was addressed by implementing algorithms such as XGBoost, which outperformed traditional Cox regression in terms of discrimination metrics, including ROC AUC and PR AUC. Despite challenges with class imbalance and calibration, these models demonstrated the potential of machine learning to provide enhanced individualized risk prediction. The comparison with statistical methods highlighted the complementary strengths of interpretability and predictive accuracy, offering a nuanced perspective on how both approaches can be integrated in prognostic modeling.

The development of an interactive Shiny application that consolidates these analyses, the fourth and last objective met, was realized through the creation of the Breast Cancer Data Explorer. This application integrates descriptive epidemiology, classical survival methods, and modern predictive models into a single accessible platform. By providing interactive plots, individualized forecasts, and exportable results without requiring programming expertise, the tool lowers the technical barriers that typically limit the dissemination and use of registry-based research. The app is significant both for its capabilities and its practical value. It lets clinicians visualize differences between patient subgroups, allows researchers to explore hypotheses interactively, and helps policymakers examine survival disparities clearly.

First, the app converts complex statistical and machine-learning results into a format that clinicians, researchers, and other stakeholders can use right away. In this way, it helps bring technical advances into practical use. Second, it shows how population-based registry data can be used dynamically in oncology. Instead of being fixed findings, the data become adaptable resources for new research questions, educational purposes, and policy discussions. The methodological contributions of this work, while robust in their own right, serve ultimately as the analytical backbone of the application, ensuring that the user experience is underpinned by rigorous and validated models.

By achieving these objectives, the thesis shows that registry-based cancer research can go beyond static reports. It can become an interactive, evolving resource. The methodological contributions, such as robust survival analyses, validated prognostic modeling, and machine-learning applications, form the analytical backbone of the application. Still, the lasting impact of the work lies in the way these methods have been operationalized for diverse audiences.

Nevertheless, several limitations temper the interpretation of the findings and highlight directions for future development. The availability of molecular data was restricted to the period after 2010, necessitating separate models for the full cohort and the more recent era. This division ensured analytical completeness but limited the long-term characterization of subtypes. The Cox models also revealed significant violations of the proportional hazards assumption, suggesting that some covariate effects evolve over time and that time-dependent or flexible survival models may provide more nuanced insights. On the predictive side, class imbalance in the dataset constrained raw accuracy, even though discrimination metrics such as ROC AUC and PR AUC were high. Finally, SEER lacks

detailed treatment regimens, comorbidity information, and adherence data, all of which would enhance prognostic modeling, and external validation remains a necessary step before clinical deployment.

These limitations, however, do not undermine the value of the study and rather indicate where the Breast Cancer Data Explorer can be further refined and expanded. From a methodological perspective, the integration of time-dependent Cox models, accelerated failure time models, or flexible survival frameworks would allow more accurate capture of evolving covariate effects. From a predictive standpoint, class imbalance could be addressed with advanced resampling strategies or cost-sensitive learning, while calibration and decision-curve analyses would provide clearer evidence of clinical utility. Richer datasets, ideally linking SEER to comorbidity and treatment records, would improve prognostic accuracy and allow for finer-grained predictions.

Most importantly, the application itself could evolve into a publishable and open-source tool for the research and clinical community. By hosting it online or releasing it through a public repository, the app could be continually updated with new SEER releases, extended to other cancer sites, or adapted for different geographic registries. Future iterations could incorporate features such as patient-specific dashboards, interactive explanations of model predictions using explainable AI techniques, or integration with external datasets for validation. In this way, the Breast Cancer Data Explorer could become not only a product of this project but also a living platform with enduring impact in cancer outcomes research.

In conclusion, this thesis demonstrates that the potential of registry-based cancer research extends beyond statistical publications to the creation of resources that empower diverse audiences. By delivering an interactive and user-friendly application that integrates descriptive, inferential, and predictive analyses, the study contributes not only to the understanding of breast cancer prognosis but also to the way such knowledge can be disseminated and applied. The Breast Cancer Data Explorer stands as a proof of concept for how methodological rigor can be combined with usability, offering a model for future translational research that seeks to bridge the gap between data, analysis, and real-world impact. Through further development, dissemination, and validation, the application has the potential to become a valuable open-access resource for clinicians, researchers, educators, and policymakers alike. Beyond its methodological contributions, the ultimate value of this work lies in its potential to democratize access to cancer survival data and

predictive models. By lowering technical barriers and offering a transparent, interactive, and user-friendly platform, the Breast Cancer Data Explorer contributes to making complex population-level insights accessible to clinicians, researchers, policymakers, and educators alike. In doing so, it not only advances breast cancer outcomes research but also illustrates how interactive tools can foster equity in knowledge dissemination and support more informed, data-driven decisions in oncology and public health.

## References

- [1] A. Tfayli, S. Temraz, R. A. Mrad, and A. Shamseddine, “Breast cancer in low- and middle-income countries: An emerging and challenging epidemic,” *Journal of Oncology*, vol. 2010, no. 1, p. 490631, 2010. [Online]. Available: <https://doi.org/10.1155/2010/490631> [Cited on page 1.]
- [2] M. Fu, Z. Peng, M. Wu *et al.*, “Global, regional, and national time trends in mortality for breast cancer, 1992-2021: an age-period-cohort analysis for the global burden of disease 2021 study,” *BMC Public Health*, vol. 25, p. 1599, 2025. [Online]. Available: <https://doi.org/10.1186/s12889-025-22855-5> [Cited on page 1.]
- [3] International Agency for Research on Cancer (IARC), “Breast cancer cases and deaths are projected to rise globally (press release no. 361),” Press release, IARC / World Health Organization, Feb. 2025, published 24 February 2025. Available from: [https://www.iarc.who.int/wp-content/uploads/2025/02/pr361\\_E.pdf](https://www.iarc.who.int/wp-content/uploads/2025/02/pr361_E.pdf). [Cited on page 1.]
- [4] Q. Li, C. Xia, H. Li *et al.*, “Disparities in 36 cancers across 185 countries: secondary analysis of global cancer statistics,” *Frontiers of Medicine*, vol. 18, pp. 911–920, 2024. [Online]. Available: <https://doi.org/10.1007/s11684-024-1058-6> [Cited on pages 1 and 2.]
- [5] R. Sha, X. M. Kong, X. Y. Li *et al.*, “Global burden of breast cancer and attributable risk factors in 204 countries and territories, from 1990 to 2021: results from the global burden of disease study 2021,” *Biomarker Research*, vol. 12, p. 87, 2024. [Online]. Available: <https://doi.org/10.1186/s40364-024-00631-8> [Cited on page 1.]
- [6] L. Guo, D. Kong, J. Liu *et al.*, “Breast cancer heterogeneity and its implication in personalized precision therapy,” *Experimental Hematology & Oncology*, vol. 12, p. 3, 2023. [Online]. Available: <https://doi.org/10.1186/s40164-022-00363-1> [Cited on page 2.]
- [7] O. Ginsburg, H. Yip, A. Brooks, A. Cabanes, M. Caleffi, J. A. Dunstan Yataco, B. Gyawali, V. McCormack, R. Mehrotra, A. Mohar, R. Murillo, L. E. Pace, E. D. Paskett, A. Romanoff, A. F. Rositch, J. R. Scheel, M. Schneidman, K. Unger-Saldaña, V. Vanderpuye, and B. O. Anderson, “Breast cancer early detection:

- A phased approach to implementation,” *Cancer*, vol. 126, pp. 2379–2393, 2020. [Online]. Available: <https://doi.org/10.1002/cncr.32887> [Cited on page 2.]
- [8] B. Feleke, L. M. Tesfaw, and A. A. Mitku, “Survival analysis of women breast cancer patients in northwest amhara, ethiopia,” *Frontiers in Oncology*, vol. 12, p. 1041245, Dec. 2022. [Cited on page 2.]
- [9] I. Soerjomataram, M. W. Louwman, J. G. Ribot, J. A. Roukema, and J. W. Coebergh, “An overview of prognostic factors for long-term survivors of breast cancer,” *Breast Cancer Research and Treatment*, vol. 107, no. 3, pp. 309–330, Feb. 2008. [Cited on page 2.]
- [10] A. B. Mariotto, A. M. Noone, N. Howlader, H. Cho, G. E. Keel, J. Garshell, S. Woloshin, and L. M. Schwartz, “Cancer survival: an overview of measures, uses, and interpretation,” *J Natl Cancer Inst Monogr*, vol. 2014, no. 49, pp. 145–186, Nov. 2014, erratum in: *J Natl Cancer Inst Monogr*. 2015 May;2015(51):97. doi:10.1093/jncimonographs/lgv001. [Cited on page 2.]
- [11] “Definition of cause-specific survival,” <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cause-specific-survival>, n.d., accessed August 27, 2025. [Cited on page 2.]
- [12] National Breast Cancer Foundation, “Breast cancer facts & stats,” <http://nationalbreastcancer.org/breast-cancer-facts/>, n.d., accessed August 27, 2025. [Cited on page 3.]
- [13] L. Chen, H. M. Linden, B. O. Anderson, and C. I. Li, “Trends in 5-year survival rates among breast cancer patients by hormone receptor status and stage,” *Breast Cancer Research and Treatment*, vol. 147, no. 3, pp. 609–616, Oct. 2014. [Cited on page 3.]
- [14] American Cancer Society, “Survival rates for breast cancer,” <https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html>, 2025, accessed: 2025-08-27. [Cited on page 3.]
- [15] A. Sadighian *et al.*, “Demographic characteristics and survival outcomes of pediatric clear cell sarcoma of the kidney: A national cancer database retrospective study,” *JCO*, vol. 43, p. 10033, 2025. [Cited on page 3.]

- [16] W.-Q. Che, Y.-J. Li, C.-K. Tsang *et al.*, “How to use the surveillance, epidemiology, and end results (seer) data: Research design and methodology,” *Military Medical Research*, vol. 10, no. 1, p. 50, 2023. [Online]. Available: <https://doi.org/10.1186/s40779-023-00488-2> [Cited on page 3.]
- [17] E. K. Roberts, L. Luo, A. M. Mondul, M. Banerjee, C. M. Veenstra, A. B. Mariotto, M. J. Schipper, K. He, J. M. G. Taylor, and A. F. Brouwer, “Time-varying associations of patient and tumor characteristics with cancer survival: an analysis of seer data across 14 cancer sites, 2004-2017,” *Cancer Causes & Control*, vol. 35, no. 10, pp. 1393–1405, 2024. [Cited on page 3.]
- [18] C. S. Santos and M. Amorim-Lopes, “Externally validated and clinically useful machine learning algorithms to support patient-related decision-making in oncology: a scoping review,” *BMC Medical Research Methodology*, vol. 25, p. 45, 2025. [Online]. Available: <https://doi.org/10.1186/s12874-025-02463-y> [Cited on page 3.]
- [19] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany *et al.*, “Revolutionizing healthcare: the role of artificial intelligence in clinical practice,” *BMC Medical Education*, vol. 23, p. 689, 2023. [Online]. Available: <https://doi.org/10.1186/s12909-023-04698-z> [Cited on page 3.]
- [20] P. E. Bretos-Azcona, E. Sánchez-Iriso, and J. M. Cabasés Hita, “Tailoring integrated care services for high-risk patients with multiple chronic conditions: a risk stratification approach using cluster analysis,” *BMC Health Services Research*, vol. 20, p. 806, 2020. [Online]. Available: <https://doi.org/10.1186/s12913-020-05668-7> [Cited on page 4.]
- [21] J. Rueff, B. Weixler, C. T. Viehl, A. Ochsner, R. Warschkow, U. Gueller, W. Mingrone, and M. Zuber, “Improved quality of colon cancer surveillance after implementation of a personalized surveillance schedule,” *Journal of Surgical Oncology*, vol. 122, no. 3, pp. 529–537, 2020. [Online]. Available: <https://doi.org/10.1002/jso.25973>
- [22] C. K. Osterman, H. K. Sanoff, W. A. Wood, M. Fasold, and J. E. Lafata, “Predictive modeling for adverse events and risk stratification programs for people receiving cancer treatment,” *JCO Oncology Practice*, vol. 18, no. 2, pp. 127–136, Feb. 2022, epub 2021 Sep 1. [Cited on page 4.]

- [23] Z. Chen, N. Liang, H. Zhang, H. Li, Y. Yang, X. Zong, Y. Chen, Y. Wang, and N. Shi, "Harnessing the power of clinical decision support systems: challenges and opportunities," *Open Heart*, vol. 10, no. 2, p. e002432, Nov. 2023. [Cited on page 4.]
- [24] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, no. 2, pp. 94–98, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2514664524010592> [Cited on page 4.]
- [25] D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," *First publ. in: Lecture notes in computer science, No. 4404 (2008), pp. 76-90*, vol. 1, 04 2008. [Cited on page 4.]
- [26] W. Cui, "Visual analytics: A comprehensive overview," *IEEE Access*, vol. 7, pp. 1–1, 06 2019. [Cited on page 5.]
- [27] H. Abdul Rahman, S. N. N. Zaim, U. S. Suhaimei, and A. A. Jamain, "Prognostic factors associated with breast cancer-specific survival from 1995 to 2022: A systematic review and meta-analysis of 1,386,663 cases from 30 countries," *Diseases*, vol. 12, no. 6, 2024. [Online]. Available: <https://doi.org/10.3390/diseases12060111> [Cited on page 9.]
- [28] F. Tonello, A. Bergmann, K. de Souza Abrahão, S. S. de Aguiar, M. A. Bello, and L. C. S. Thuler, "Impact of number of positive lymph nodes and lymph node ratio on survival of women with node-positive breast cancer," *European Journal of Breast Health*, vol. 15, no. 2, pp. 76–84, apr 2019. [Online]. Available: <https://doi.org/10.5152/ejbh.2019.4414> [Cited on page 9.]
- [29] G. Dong, D. Wang, X. Liang, H. Gao, L. Wang, X. Yu, and J. Liu, "Factors related to survival rates for breast cancer patients," *International Journal of Clinical and Experimental Medicine*, vol. 7, no. 10, pp. 3719–3724, oct 2014. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4238527/> [Cited on page 9.]
- [30] L. Vaz-Gonçalves, L. Marquart-Wilson, M. M. Protani, M. T. Stephensen, J. Moore, M. F. Morris, J. M. Saunus, and M. M. Reeves, "Capturing breast cancer subtypes in cancer registries: Insights into real-world incidence and survival," *Journal of Cancer Policy*, vol. 44, p. 100567, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2213538325000116> [Cited on page 9.]

- [31] M. Darida, G. Rubovszky, Z. Kiss, B. Székely, B. Madaras, Z. Horváth, J. Kocsis, I. Sipőcz, M. Várnai, É. Balogh, K. Kovács, V. Buga, E. Karamousouli, T. G. Szabó, G. Rokszin, I. Fábián, I. Kenessey, A. Wéber, P. Nagy, and C. Polgár, “Improvement in breast cancer survival across molecular subtypes in hungary between 2011 and 2020: A nationwide, retrospective study,” *Frontiers in Oncology*, vol. 15, p. 1465511, 2025. [Online]. Available: <https://doi.org/10.3389/fonc.2025.1465511> [Cited on page 9.]
- [32] H. Huang, T. Wei, A. Zhang *et al.*, “Trends in the incidence and survival of women with hormone receptor-positive breast cancer from 1990 to 2019: a large population-based analysis,” *Scientific Reports*, vol. 14, p. 23690, 2024. [Online]. Available: <https://doi.org/10.1038/s41598-024-74746-1> [Cited on page 9.]
- [33] L. K. Dunnwald, M. A. Rossing, and C. I. Li, “Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients,” *Breast Cancer Research*, vol. 9, no. 1, p. R6, 2007. [Online]. Available: <https://doi.org/10.1186/bcr1639> [Cited on page 9.]
- [34] M. Chen, J. Wu, and L. Zhu, “Correlation of hormonal receptor level and survival outcome in her2-positive nonmetastatic breast cancer.” *Journal of Clinical Oncology*, vol. 38, no. 15\_suppl, pp. e12568–e12568, 2020, pMID:. [Online]. Available: [https://ascopubs.org/doi/abs/10.1200/JCO.2020.38.15\\_suppl.e12568](https://ascopubs.org/doi/abs/10.1200/JCO.2020.38.15_suppl.e12568) [Cited on page 9.]
- [35] J. Fu, L. Wu, T. Xu, D. Li, M. Ying, M. Jiang, T. Jiang, W. Fu, F. Wang, and J. Du, “Young-onset breast cancer: a poor prognosis only exists in low-risk patients,” *J Cancer*, vol. 10, no. 14, pp. 3124–3132, Jun. 2019. [Online]. Available: <https://doi.org/10.7150/jca.30432> [Cited on page 9.]
- [36] J. Fernandes, C. Cardoso-Filho, M. Kraft, A. Detoni, B. Duarte, J. Shinzato, and D. Vale, “Differences in breast cancer survival and stage by age in off-target screening groups: a population-based retrospective study,” *AJOG Global Reports*, vol. 3, no. 2, p. 100208, Apr. 2023. [Online]. Available: <https://doi.org/10.1016/j.xagr.2023.100208> [Cited on page 9.]
- [37] E. Hormones, B. C. C. Group, T. Key, P. Appleby, G. Reeves, A. Roddam, K. Helzlsouer, A. Alberg, D. Rollison, J. Dorgan, L. Brinton, K. Overvad, R. Kaaks, A. Trichopoulou, F. Clavel-Chapelon, S. Panico, E. Duell, P. Peeters, S. Rinaldi,

- I. Fentiman, M. Dowsett, and H. Strickler, "Circulating sex hormones and breast cancer risk factors in postmenopausal women: reanalysis of 13 studies," *British Journal of Cancer*, vol. 105, no. 5, pp. 709–722, 2011. [Online]. Available: <https://doi.org/10.1038/bjc.2011.254> [Cited on page 10.]
- [38] S. Scomersi, F. Giudici, G. Cacciatoe *et al.*, "Comparison between male and female breast cancer survival using propensity score matching analysis," *Scientific Reports*, vol. 11, p. 11639, 2021. [Online]. Available: <https://doi.org/10.1038/s41598-021-91131-4> [Cited on page 10.]
- [39] S. Chatterji, J. Niehues, M. van Treeck *et al.*, "Prediction models for hormone receptor status in female breast cancer do not extend to males: further evidence of sex-based disparity in breast cancer," *npj Breast Cancer*, vol. 9, p. 91, 2023. [Online]. Available: <https://doi.org/10.1038/s41523-023-00599-y> [Cited on page 10.]
- [40] R. Bleicher, "Timing and delays in breast cancer evaluation and treatment," *Annals of Surgical Oncology*, vol. 25, no. 10, pp. 2829–2838, Oct. 2018. [Online]. Available: <https://doi.org/10.1245/s10434-018-6615-2> [Cited on page 10.]
- [41] S. Jeong, S. Chun, H. Lee *et al.*, "Impact of diagnosis-to-treatment interval on mortality in patients with early-stage breast cancer: a retrospective nationwide korean cohort," *BMC Women's Health*, vol. 25, p. 247, 2025. [Online]. Available: <https://doi.org/10.1186/s12905-025-03780-6> [Cited on page 10.]
- [42] G. Wright, J. Wong, J. Morgan, S. Roy-Chowdhury, K. Kazanjian, and S. Lum, "Time from diagnosis to surgical treatment of breast cancer: Factors influencing delays in initiating treatment," *The American Surgeon*, vol. 76, no. 10, pp. 1119–1122, 2010. [Online]. Available: <https://doi.org/10.1177/000313481007601022> [Cited on page 10.]
- [43] I. Kolářová, B. Melichar, I. Sirák, J. Vaňásek, J. Petera, K. Horáčková, D. Pohanková, F. Ďatelinka, Z. Šinkorová, and M. Vošmik, "The role of adjuvant radiotherapy in the treatment of breast cancer," *Current Oncology*, vol. 31, no. 3, pp. 1207–1220, 2024. [Online]. Available: <https://doi.org/10.3390/currncol31030090> [Cited on page 10.]
- [44] N. Nakajima, M. Oguchi, Y. Kumai, M. Yoshida, H. Inoda, Y. Yoshioka, T. Iwase, Y. Ito, F. Akiyama, and S. Ohno, "Clinical outcomes and prognostic factors in

- patients with stage ii-iii breast cancer treated with neoadjuvant chemotherapy followed by surgery and postmastectomy radiation therapy in the modern treatment era,” *Advances in Radiation Oncology*, vol. 3, no. 3, pp. 271–279, Apr. 2018. [Online]. Available: <https://doi.org/10.1016/j.adro.2018.04.006> [Cited on page 10.]
- [45] X. Li, M. Wang, M. Wang, X. Yu, J. Guo, T. Sun, L. Yao, Q. Zhang, and Y. Xu, “Predictive and prognostic roles of pathological indicators for patients with breast cancer on neoadjuvant chemotherapy,” *Journal of Breast Cancer*, vol. 22, no. 4, pp. 497–521, Dec. 2019. [Online]. Available: <https://doi.org/10.4048/jbc.2019.22.e49> [Cited on page 10.]
- [46] M. Zhang, J. Yuan, M. Wang, M. Zhang, and H. Chen, “Chemotherapy is of prognostic significance to metaplastic breast cancer,” *Scientific Reports*, vol. 14, no. 1, p. 1210, 2024. [Online]. Available: <https://doi.org/10.1038/s41598-024-51627-1> [Cited on page 10.]
- [47] S. Lehrer, S. Green, and K. Rosenzweig, “Affluence and breast cancer,” *The Breast Journal*, vol. 22, no. 5, pp. 564–567, Sep. 2016. [Online]. Available: <https://doi.org/10.1111/tbj.12630> [Cited on page 11.]
- [48] T. Pudrovska and B. Anikputa, “The role of early-life socioeconomic status in breast cancer incidence and mortality: unraveling life course mechanisms,” *Journal of Aging and Health*, vol. 24, no. 2, pp. 323–344, Mar. 2012. [Online]. Available: <https://doi.org/10.1177/0898264311422744> [Cited on page 11.]
- [49] B. Sprague, A. Trentham-Dietz, and E. Burnside, “Socioeconomic disparities in the decline in invasive breast cancer incidence,” *Breast Cancer Research and Treatment*, vol. 122, no. 3, pp. 873–878, Aug. 2010. [Online]. Available: <https://doi.org/10.1007/s10549-010-0737-y> [Cited on page 11.]
- [50] C. G. Yedjou, J. N. Sims, L. Miele, F. Noubissi, L. Lowe, D. D. Fonseca, R. A. Alo, M. Payton, and P. B. Tchounwou, “Health and racial disparity in breast cancer,” in *Advances in Experimental Medicine and Biology*. Springer, 2019, vol. 1152, pp. 31–49. [Online]. Available: [https://doi.org/10.1007/978-3-030-20301-6\\_3](https://doi.org/10.1007/978-3-030-20301-6_3) [Cited on page 11.]
- [51] C. Ku, A. Desai, A. Hickey, B. Valcarcel, C. Wareham, A. Hernandez, E. E. Arias-Rivera, D. Chambergo-Michilot, D. Linshaw, G. Ziegler-Rodriguez, S. M. Persing, C. Homsy, A. Chatterjee, and S. M. Nardello, “Disparities in triple negative

- breast cancer among hispanic population living in latin america versus the united states,” *Journal of Surgical Oncology*, vol. 131, no. 7, pp. 1258–1271, 2025. [Online]. Available: <https://doi.org/10.1002/jso.28088> [Cited on page 11.]
- [52] E. Roldan-Vasquez, S. Mitri, K. D. Brantley, T. A. James, and I. Schlam, “Breast cancer treatment delay and the hispanic paradox: A seer database analysis,” *The American Journal of Surgery*, vol. 244, p. 116135, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0002961024006871> [Cited on page 11.]
- [53] G. LeBlanc, I. Lee, H. Carretta, Y. Luo, D. Sinha, and G. Rust, “Rural-urban differences in breast cancer stage at diagnosis,” *Women’s Health Reports (New Rochelle)*, vol. 3, no. 1, pp. 207–214, Feb. 2022. [Online]. Available: <https://doi.org/10.1089/whr.2021.0082> [Cited on page 11.]
- [54] D. Riley, E. Chrischilles, I. Lizarraga, M. Charlton, B. Smith, and C. Lynch, “Rural-urban differences in secular trends of locoregional treatment for ductal carcinoma in situ: A patterns of care analysis,” *Cancer Medicine*, vol. 11, no. 11, pp. 2284–2295, 2022. [Online]. Available: <https://doi.org/10.1002/cam4.4605> [Cited on page 11.]
- [55] F. Williams, S. Jeanetta, D. O’Brien, and J. Fresen, “Rural-urban difference in female breast cancer diagnosis in missouri,” *Rural and Remote Health*, vol. 15, p. 3063, 2015. [Online]. Available: <https://doi.org/10.22605/RRH3063> [Cited on page 11.]
- [56] D. Jiao, Y. Ma, J. Zhu, H. Dai, Y. Yang, Y. Zhao, X. Guo, and Z. Liu, “Impact of marital status on prognosis of patients with invasive breast cancer: A population-based study using seer database,” *Frontiers in Oncology*, vol. 12, p. 913929, 2022. [Online]. Available: <https://doi.org/10.3389/fonc.2022.913929> [Cited on page 12.]
- [57] W. Ding, G. Ruan, Y. Lin *et al.*, “Dynamic changes in marital status and survival in women with breast cancer: a population-based study,” *Scientific Reports*, vol. 11, p. 5421, 2021. [Online]. Available: <https://doi.org/10.1038/s41598-021-84996-y> [Cited on page 12.]
- [58] M. Martínez, J. Unkart, L. Tao, C. Kroenke, R. Schwab, I. Komenaka, and S. Gomez, “Prognostic significance of marital status in breast cancer survival:

- A population-based study,” *PLOS ONE*, vol. 12, no. 5, p. e0175515, May 2017. [Online]. Available: <https://doi.org/10.1371/journal.pone.0175515> [Cited on page 12.]
- [59] S. Cai, W. Zuo, X. Lu, Z. Gou, Y. Zhou, P. Liu, Y. Pan, and S. Chen, “The prognostic impact of age at diagnosis upon breast cancer of different immunohistochemical subtypes: A surveillance, epidemiology, and end results (seer) population-based analysis,” *Frontiers in Oncology*, vol. 10, p. 559229, 2020. [Online]. Available: <https://doi.org/10.3389/fonc.2020.01729> [Cited on page 12.]
- [60] J. Moore, S. Andrzejak, S. Jones, and Y. Han, “Exploring the intersectionality of race/ethnicity with rurality on breast cancer outcomes: Seer analysis, 2000–2016,” *Breast Cancer Research and Treatment*, vol. 197, no. 3, pp. 633–645, 2023. [Online]. Available: <https://doi.org/10.1007/s10549-022-06830-x> [Cited on page 12.]
- [61] Y. Wei, P. Liu, X. Liu, M. Wang, D. Liu, H. Cui, S. Lin, H. Wu, X. Ma, and H. Kang, “Analysis of the relationship between primary tumor site and clinicopathological characteristics and survival prognosis of breast cancer patients based on seer database,” *Clinical Breast Cancer*, vol. 24, no. 8, pp. 728–745, 2024. [Online]. Available: <https://doi.org/10.1016/j.clbc.2024.08.021> [Cited on page 13.]
- [62] Y. Liu, D. Wang, Z. Yang, R. Ma, Z. Li, W. Suo, Z. Zhao, and Z. Li, “Marital status is an independent prognostic factor in inflammatory breast cancer patients: an analysis of the surveillance, epidemiology, and end results database,” *Breast Cancer Research and Treatment*, vol. 178, no. 2, pp. 379–388, 2019. [Online]. Available: <https://doi.org/10.1007/s10549-019-05385-8> [Cited on pages 12 and 13.]
- [63] G. Alsaffaf and S. Serttaş, “Performance of machine learning methods on breast cancer prediction,” *Journal of Scientific Reports-B*, no. 012, p. 1–9, 2025. [Cited on page 14.]
- [64] X. Song, J. Chu, Z. Guo, Q. Wei, Q. Wang, W. Hu, L. Wang, W. Zhao, H. Zheng, X. Lu, and J. Zhou, “Prognostic prediction of breast cancer patients using machine learning models: a retrospective analysis,” *Gland Surgery*, vol. 13, no. 9, pp. 1575–1587, 2024. [Online]. Available: <https://doi.org/10.21037/gs-24-106> [Cited on pages 14, 16, and 17.]

- [65] J. J. Kim, G. R. Nahass, Y. Dai, and T. Tulabandhula, "Prediction of lung metastasis from hepatocellular carcinoma using the seer database," *arXiv preprint arXiv:2501.11720*, 2025. [Online]. Available: <https://arxiv.org/abs/2501.11720> [Cited on pages 15, 16, and 17.]
  
- [66] W. Li, Q. Zhou, W. Liu, C. Xu, Z. Tang, S. Dong, H. Wang, W. Li, K. Zhang, R. Li, W. Zhang, Z. Hu, S. Shibin, Q. Liu, S. Kuang, and C. Yin, "A machine learning-based predictive model for predicting lymph node metastasis in patients with ewing's sarcoma," *Frontiers in Medicine*, vol. 9, p. 832108, 2022. [Online]. Available: <https://doi.org/10.3389/fmed.2022.832108> [Cited on pages 14, 15, 16, and 17.]
  
- [67] M. S. A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, A. T. Azar, and A. Shaikh, "Enhancing breast cancer detection and classification using advanced multi-model features and ensemble machine learning techniques," *Life*, vol. 13, no. 10, p. 2093, 2023. [Online]. Available: <https://doi.org/10.3390/life13102093> [Cited on pages 15, 16, and 17.]
  
- [68] P. Karatza, K. Dalakleidi, M. Athanasiou, and K. S. Nikita, "Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 2310–2313. [Online]. Available: <https://doi.org/10.1109/EMBC46164.2021.9630556> [Cited on page 16.]
  
- [69] V. LeBaron, N. Crimp, N. Homdee, K. Reed, V. Petermann, W. Ashe, L. Blackhall, and B. Lewis, "'Less words, more pictures': Creating and sharing data visualizations from a remote health monitoring system with clinicians to improve cancer pain management," *Frontiers in Digital Health*, vol. 7, p. 1520990, 2025. [Online]. Available: <https://doi.org/10.3389/fdgth.2025.1520990> [Cited on pages 18 and 19.]
  
- [70] P. L. Chia, C. Gedye, P. C. Boutros, P. Wheatley-Price, and T. John, "Current and evolving methods to visualize biological data in cancer research," *Journal of the National Cancer Institute*, vol. 108, no. 8, p. djw031, 2016. [Online]. Available: <https://doi.org/10.1093/jnci/djw031> [Cited on page 18.]

- [71] National Cancer Institute, “Visualizing data: The basics,” <https://www.cancer.gov/about-nci/organization/cbiit/training/learn/visualize>, 2025, accessed: 2025-09-14. [Cited on page 18.]
- [72] J. Stellamanns, D. Ruetters, K. Dahal, Z. Schillmoeller, and J. Huebner, “Visualizing risks in cancer communication: A systematic review of computer-supported visual aids,” *Patient Education and Counseling*, vol. 100, no. 8, pp. 1421–1431, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0738399117300678> [Cited on page 18.]
- [73] R. Qureshi, X. Chen, C. Goerg, E. Mayo-Wilson, S. Dickinson, L. Golzarri-Arroyo, H. Hong, R. Phillips, V. Cornelius, M. McAdams DeMarco, E. Guallar, and T. Li, “Comparing the value of data visualization methods for communicating harms in clinical trials,” *Epidemiologic Reviews*, vol. 44, no. 1, pp. 55–66, Dec. 2022. [Online]. Available: <https://doi.org/10.1093/epirev/mxac005> [Cited on page 18.]
- [74] S. Prabhakaran, C. Yapp, G. Baker, J. Beyer, Y. Chang, A. Creason, R. Krueger, J. Muhlich, N. Patterson, K. Sidak, D. Sudar, A. Taylor, L. Ternes, J. Troidl, X. Yubin, A. Sokolov, and D. Tyson, “Addressing persistent challenges in digital image analysis of cancer tissue: resources developed from a hackathon,” *Molecular Oncology*, vol. 19, no. 6, pp. 1565–1581, Jun. 2025. [Online]. Available: <https://doi.org/10.1002/1878-0261.13783> [Cited on page 19.]
- [75] W. Jia, H. Li, S. Li, L. Chen, and S. Li, “Oviz-bio: a web-based platform for interactive cancer genomics data visualization,” *Nucleic Acids Research*, vol. 48, no. W1, pp. W415–W426, 2020. [Online]. Available: <https://doi.org/10.1093/nar/gkaa371> [Cited on page 19.]
- [76] Y. E. Wang, L. Kutnetsov, A. Partensky, J. Farid, and J. Quackenbush, “Webmev: A cloud platform for analyzing and visualizing cancer genomic data,” *Cancer Research*, vol. 77, no. 21, pp. e11–e14, nov 2017. [Online]. Available: <https://doi.org/10.1158/0008-5472.CAN-17-0802> [Cited on page 19.]
- [77] D. Boehm, C. Strantz, J. Christoph, H. Busch, T. Ganslandt, and P. Unberath, “Data visualization support for tumor boards and clinical oncology: Protocol for a scoping review,” *JMIR Res Protoc*, vol. 13, p. e53627, Mar 2024. [Online]. Available: <https://www.researchprotocols.org/2024/1/e53627> [Cited on page 19.]

- [78] D. G. Goldberg, T. G. Soylu, C. F. Hoffman, R. E. Kishton, and P. F. Cronholm, “Clinicians’ perspectives on the adoption and implementation of emr-integrated clinical decision support tools in primary care,” *DIGITAL HEALTH*, vol. 11, 2025. [Cited on page 19.]
- [79] Q. Dai, M. Li, M. Yang, S. Shi, Z. Wang, J. Liao, Z. Li, W. E. L. Tao, and Y. D. Tang, “Attitudes, perceptions, and factors influencing the adoption of ai in health care among medical staff: Nationwide cross-sectional survey study,” *Journal of Medical Internet Research*, vol. 27, p. e75343, 2025. [Cited on page 19.]
- [80] F. Coelho, D. C. P. Câmara, E. C. Araújo, L. M. Bianchi, I. Ogasawara, J. Dalal, A. James, J. L. Abbate, A. Merzouki, I. C. Dos Reis, K. D. Nwosu, and O. Keiser, “A platform for data-centric, continuous epidemiological analyses (epigraphhub): Descriptive analysis,” *Journal of Medical Internet Research*, vol. 25, p. e40554, 2023. [Cited on page 19.]
- [81] P. Soares, R. J. Alves, A. B. Abecasis *et al.*, “intb - a data integration platform for molecular and clinical epidemiological analysis of tuberculosis,” *BMC Bioinformatics*, vol. 14, p. 264, 2013.
- [82] Z. Valero-Ramon, C. Fernandez-Llatas, G. Collantes, B. Valdivieso, A. Billis, P. Bamidis, and V. Traver, “Analytical exploratory tool for healthcare professionals to monitor cancer patients’ progress,” *Frontiers in Oncology*, vol. 12, p. 1043411, 2023. [Cited on page 19.]
- [83] Surveillance, Epidemiology, and End Results (SEER) Program, National Cancer Institute, “Registry groupings in seer data and statistics,” <https://seer.cancer.gov/registries/terms.html>, 2024. [Cited on page 24.]
- [84] —, “Variable & recode definitions,” <https://seer.cancer.gov/analysis/>, 2024, accessed: 2025-09-15. [Cited on page 24.]
- [85] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, “Big data preprocessing: methods and prospects,” *Big Data Analytics*, vol. 1, no. 1, p. 9, 2016. [Online]. Available: <https://doi.org/10.1186/s41044-016-0014-0> [Cited on page 27.]

- [86] N. Howlader, A.-M. Noone, M. Yu, and K. A. Cronin, "Use of imputed population-based cancer registry data as a method of accounting for missing information: application to estrogen receptor status for breast cancer," *American Journal of Epidemiology*, vol. 176, no. 4, pp. 347–356, 2012. [Cited on pages 30, 50, and 55.]
- [87] A. Alruhaymi and C. Kim, "Why can multiple imputations and how (mice) algorithm work?" *Open Journal of Statistics*, vol. 11, pp. 759–777, 2021. [Cited on page 33.]
- [88] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. Chapman and Hall/CRC, 2018. [Online]. Available: <https://doi.org/10.1201/9780429492259> [Cited on page 33.]
- [89] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, vol. 30, no. 4, pp. 377–399, 2011, epub 2010 Nov 30. [Cited on page 33.]
- [90] Y. Zhang and A. M. Molinaro, "Categorizing continuous biomarkers: More cons than pros," *Neuro-oncology Practice*, vol. 9, no. 2, pp. 81–82, 2022. [Cited on page 34.]
- [91] C. Andrade, "A student's guide to the classification and operationalization of variables in the conceptualization and design of a clinical study: Part 2," *Indian Journal of Psychological Medicine*, vol. 43, no. 3, pp. 265–268, 2021. [Cited on page 34.]
- [92] National Cancer Institute. (2025) Age standards for survival (standard populations). Accessed August 6 2025. [Online]. Available: <https://seer.cancer.gov/stdpopulations/survival.html> [Cited on page 34.]
- [93] R. D. Rosen and A. Sapra, *TNM Classification*, internet ed. Treasure Island (FL): StatPearls Publishing, 2025, updated 2023 Feb 13; accessed 2025 Aug 06. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK553187/> [Cited on page 34.]
- [94] American Cancer Society. (2024) Stages of breast cancer. Updated September 10 2024; accessed August 6 2025. [Online]. Available: <https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html> [Cited on page 34.]
- [95] P. J. Ho, A. R. Cook, N. K. Binte Mohamed Ri, J. Liu, J. Li, and M. Hartman, "Impact of delayed treatment in women diagnosed with breast cancer: A population-based study," *Cancer Medicine*, vol. 9, no. 7, pp. 2435–2444, 2020. [Cited on page 35.]

- [96] Z. Ungvari, M. Fekete, A. Buda, A. Lehoczki, G. Munkácsy, P. Scaffidi, T. Bonaldi, J. T. Fekete, G. Bianchini, P. Varga, A. Ungvari, and B. Györfy, “Quantifying the impact of treatment delays on breast cancer survival outcomes: a comprehensive meta-analysis,” *GeroScience*, 2025, published online June 10, 2025; Received April 3, 2025; Accepted May 25, 2025. [Cited on page 35.]
- [97] M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*. New York, Heidelberg, Dordrecht, London: Springer Science+Business Media, 2014, library of Congress Control Number: 2014934697. [Online]. Available: <https://doi.org/10.1007/978-0-8176-4811-4> [Cited on page 37.]
- [98] M. L. McHugh, “The chi-square test of independence,” *Biochemia Medica*, vol. 23, no. 2, pp. 143–149, 2013. [Online]. Available: <https://doi.org/10.11613/BM.2013.018> [Cited on page 37.]
- [99] M. Kumar Goel, P. Khanna, and J. Kishore, “Ayurveda and scientific publication,” *International Journal of Ayurveda Research*, vol. 1, p. 274–278, 2010. [Cited on page 38.]
- [100] J. M. Bland and D. G. Altman, “The logrank test,” *BMJ*, vol. 328, no. 7447, p. 1073, 2004. [Cited on page 38.]
- [101] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson, *shiny: Web Application Framework for R*, 2024, r package version 1.8.0. [Online]. Available: <https://shiny.rstudio.com/> [Cited on page 42.]
- [102] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: <https://www.R-project.org/> [Cited on page 42.]
- [103] N. C. I. S. Program). (2025) Breast subtype (2010+) - seer\*stat variable description. [Online]. Available: <https://seer.cancer.gov/seerstat/databases/ssf/breast-subtype.html> [Cited on page 44.]
- [104] ——. (2023) Seer treatment data limitations (november 2023 submission). [Online]. Available: <https://seer.cancer.gov/data-software/documentation/seerstat/nov2023/treatment-limitations-nov2023.html> [Cited on page 45.]
- [105] National Cancer Institute, Surveillance, Epidemiology, and End Results Program (SEER), *Summary Stage 2018: Breast*, National Cancer Institute, Bethesda, MD,

2018, accessed: 2025-08-10. [Online]. Available: <https://seer.cancer.gov/tools/ssm/SSM2018-BREAST.pdf> [Cited on page 47.]

- [106] American Joint Committee on Cancer, “Breast,” in *AJCC Cancer Staging Manual*, 8th ed. Cham, Switzerland: Springer International Publishing, 2017, ch. 48, used with permission of the American College of Surgeons, Chicago, Illinois. [Cited on page 47.]
- [107] B. L. Sprague and A. Trentham-Dietz, “Prevalence of breast carcinoma in situ in the united states,” *JAMA*, vol. 302, no. 8, pp. 846–848, 2009. [Cited on page 48.]
- [108] H. Sackey, M. Hui, K. Czene, H. Verkooijen, G. Edgren, J. Frisell, and M. Hartman, “The impact of \*in situ\* breast cancer and family history on risk of subsequent breast cancer events and mortality: a population-based study from sweden,” *Breast Cancer Research*, vol. 18, no. 1, p. 105, 2016. [Cited on page 48.]
- [109] M.-C. Hsieh, Q. Yu, X.-C. Wu, B. Wohler, Y. Fan, B. Qiao, A. Jemal, and U. A. Ajani, “Evaluating factors associated with unknown seer summary stage 2000 derived from collaborative stage at central registry level,” *Journal of Registry Management*, vol. 39, no. 3, pp. 101–106, Fall 2012, pMCID: PMC5861716. [Cited on page 48.]
- [110] National Cancer Institute, Surveillance, Epidemiology, and End Results (SEER) Program, “Seer\*stat frequency exercise 1b: User-defined variables,” Online tutorial, n.d., accessed: 2025-08-11. [Online]. Available: <https://seer.cancer.gov/seerstat/tutorials/freq1b/webprint/> [Cited on page 48.]
- [111] —, “Derived her2 recode (2010+),” Online documentation, n.d., accessed: 2025-08-11. [Online]. Available: <https://seer.cancer.gov/seerstat/databases/ssf/her2-derived.html> [Cited on page 49.]
- [112] Surveillance, Epidemiology, and End Results (SEER) Program, “Seer solid tumor rules: Breast,” [https://seer.cancer.gov/tools/solidtumor/Breast\\_STM.pdf](https://seer.cancer.gov/tools/solidtumor/Breast_STM.pdf), 2024, accessed: 2025-08-11. [Cited on page 55.]
- [113] A. H. Lee, “Why is carcinoma of the breast more frequent in the upper outer quadrant? a case series based on needle core biopsy diagnoses,” *The Breast*, vol. 14, no. 2, pp. 151–152, 2005. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960977604001043> [Cited on page 55.]

- [114] S. Haghpanah, M. Hosseini-Bensenjan, M. Ramzi, Z. Khosravizadegan, and A. Rezaianzadeh, "Investigating the trends of incidence rates of breast cancer in southern iran: a population based survey," *BMC Women's Health*, vol. 23, no. 1, p. 589, Nov 2023. [Cited on page 56.]
  
- [115] K. E. Salerno, "Nccn guidelines update: Evolving radiation therapy recommendations for breast cancer," *Journal of the National Comprehensive Cancer Network*, vol. 15, no. 5S, pp. 682–684, 2017. [Online]. Available: <https://doi.org/10.6004/jnccn.2017.0072> [Cited on page 58.]
  
- [116] M. Adamo, C. Groves, L. Dickie, and J. Ruhl, *SEER Program Coding and Staging Manual 2023*, U.S. Department of Health and Human Services, National Institutes of Health, National Cancer Institute, Bethesda, MD 20892, Sep. 2022. [Online]. Available: [https://seer.cancer.gov/archive/manuals/2023/SPCSM\\_2023\\_MainDoc.pdf](https://seer.cancer.gov/archive/manuals/2023/SPCSM_2023_MainDoc.pdf) [Cited on page 66.]

# Contingency Tables Appendix

Table 1: Five-Year Survival Counts and Percentages by Age Group

| Age Group   | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|-------------|-----------------------------|----------------------|
| 0-19 years  | 16 (25.81%)                 | 46 (74.19%)          |
| 20-24 years | 186 (27.11%)                | 500 (72.89%)         |
| 25-29 years | 1,090 (26.73%)              | 2,988 (73.27%)       |
| 30-34 years | 3,028 (23.92%)              | 9,631 (76.08%)       |
| 35-39 years | 6,298 (21.80%)              | 22,587 (78.20%)      |
| 40-44 years | 10,262 (17.10%)             | 49,735 (82.90%)      |
| 45-49 years | 14,135 (15.32%)             | 78,139 (84.68%)      |
| 50-54 years | 17,343 (16.12%)             | 90,246 (83.88%)      |
| 55-59 years | 19,042 (16.66%)             | 95,263 (83.34%)      |
| 60-64 years | 18,277 (15.59%)             | 98,936 (84.41%)      |
| 65-69 years | 16,865 (15.11%)             | 94,732 (84.89%)      |
| 70-74 years | 15,185 (16.27%)             | 78,141 (83.73%)      |
| 75-79 years | 14,527 (18.53%)             | 63,883 (81.47%)      |
| 80-84 years | 11,936 (20.85%)             | 45,301 (79.15%)      |
| 85-89 years | 8,049 (24.51%)              | 24,796 (75.49%)      |
| 90+ years   | 4,335 (29.37%)              | 10,425 (70.63%)      |

Table 2: Five-Year Survival Counts and Percentages by Breast Cancer Subtype

| Breast Subtype       | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|----------------------|-----------------------------|----------------------|
| HR-/HER2-            | 10,359 (24.37%)             | 32,153 (75.63%)      |
| HR-/HER2+            | 3,292 (18.79%)              | 14,224 (81.21%)      |
| HR+/HER2-            | 31,023 (10.88%)             | 254,164 (89.12%)     |
| HR+/HER2+            | 5,957 (14.41%)              | 35,370 (85.59%)      |
| Recode not available | 101,665 (20.13%)            | 403,436 (79.87%)     |
| Unknown              | 8,278 (24.15%)              | 26,002 (75.85%)      |

Table 3: Five-Year Survival Counts and Percentages by Chemotherapy Status

| Chemotherapy Status | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|---------------------|-----------------------------|----------------------|
| No / Unknown        | 77,601 (13.84%)             | 482,967 (86.16%)     |
| Yes                 | 82,973 (22.71%)             | 282,382 (77.29%)     |

Table 4: Five-Year Survival Counts and Percentages by Combined Summary Stage

| Summary Stage    | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|------------------|-----------------------------|----------------------|
| Localized        | 34,420 (7.38%)              | 431,700 (92.62%)     |
| Regional         | 49,887 (22.92%)             | 167,754 (77.08%)     |
| Distant          | 31,227 (75.92%)             | 9,907 (24.08%)       |
| Unknown/Unstaged | 3,230 (37.95%)              | 5,282 (62.05%)       |
| Blank(s)         | 41,810 (21.72%)             | 150,706 (78.28%)     |

Table 5: Five-Year Survival Counts and Percentages by Derived HER2 Status

| Derived HER2 Status  | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|----------------------|-----------------------------|----------------------|
| Borderline/Unknown   | 8,007 (24.00%)              | 25,349 (76.00%)      |
| Negative             | 41,581 (12.66%)             | 286,791 (87.34%)     |
| Positive             | 9,321 (15.77%)              | 49,773 (84.23%)      |
| Recode not available | 101,665 (20.13%)            | 403,436 (79.87%)     |

Table 6: Five-Year Survival Counts and Percentages by ER Status

| ER Status            | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|----------------------|-----------------------------|----------------------|
| Borderline/Unknown   | 19,576 (27.78%)             | 50,901 (72.22%)      |
| Negative             | 41,354 (25.58%)             | 120,306 (74.42%)     |
| Positive             | 99,301 (14.35%)             | 592,735 (85.65%)     |
| Recode not available | 343 (19.60%)                | 1,407 (80.40%)       |

Table 7: Five-Year Survival Counts and Percentages by First Malignant Primary Indicator

| First Malignant Primary Indicator | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|-----------------------------------|-----------------------------|----------------------|
| No                                | 28,464 (20.07%)             | 113,336 (79.93%)     |
| Yes                               | 132,110 (16.85%)            | 652,013 (83.15%)     |

Table 8: Five-Year Survival Counts and Percentages by Marital Status at Diagnosis

| Marital Status                 | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|--------------------------------|-----------------------------|----------------------|
| Divorced                       | 18,245 (18.86%)             | 78,483 (81.14%)      |
| Married (including common law) | 74,427 (14.75%)             | 430,263 (85.25%)     |
| Separated                      | 2,063 (21.72%)              | 7,434 (78.28%)       |
| Single (never married)         | 26,670 (21.20%)             | 99,135 (78.80%)      |
| Unknown                        | 7,309 (18.59%)              | 32,007 (81.41%)      |
| Unmarried or Domestic Partner  | 158 (11.53%)                | 1,212 (88.47%)       |
| Widowed                        | 31,702 (21.35%)             | 116,815 (78.65%)     |

Table 9: Five-Year Survival Counts and Percentages by Median Household Income (Inflation Adjusted to 2022)

| Median Household Income                | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|----------------------------------------|-----------------------------|----------------------|
| < \$40,000                             | 2,320 (22.26%)              | 8,102 (77.74%)       |
| \$40,000 - \$44,999                    | 3,049 (20.65%)              | 11,718 (79.35%)      |
| \$45,000 - \$49,999                    | 5,451 (20.34%)              | 21,345 (79.66%)      |
| \$50,000 - \$54,999                    | 7,213 (19.56%)              | 29,671 (80.44%)      |
| \$55,000 - \$59,999                    | 8,212 (17.97%)              | 37,483 (82.03%)      |
| \$60,000 - \$64,999                    | 12,200 (18.17%)             | 54,940 (81.83%)      |
| \$65,000 - \$69,999                    | 13,852 (17.50%)             | 65,320 (82.50%)      |
| \$70,000 - \$74,999                    | 29,498 (18.46%)             | 130,272 (81.54%)     |
| \$75,000 - \$79,999                    | 15,228 (16.99%)             | 74,411 (83.01%)      |
| \$80,000 - \$84,999                    | 14,239 (17.55%)             | 66,891 (82.45%)      |
| \$85,000 - \$89,999                    | 7,250 (16.93%)              | 35,575 (83.07%)      |
| \$90,000 - \$94,999                    | 9,843 (15.00%)              | 55,781 (85.00%)      |
| \$95,000 - \$99,999                    | 9,795 (15.29%)              | 54,281 (84.71%)      |
| \$100,000 - \$109,999                  | 11,154 (15.94%)             | 58,804 (84.06%)      |
| \$110,000 - \$119,999                  | 6,415 (16.44%)              | 32,595 (83.56%)      |
| \$120,000+                             | 4,841 (14.70%)              | 28,082 (85.30%)      |
| Unknown/missing/no match/Not 1990-2022 | 14 (15.22%)                 | 78 (84.78%)          |

Table 10: Five-Year Survival Counts and Percentages by PR Status

| PR Status            | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|----------------------|-----------------------------|----------------------|
| Borderline/Unknown   | 21,723 (26.81%)             | 59,288 (73.19%)      |
| Negative             | 60,517 (23.66%)             | 195,314 (76.34%)     |
| Positive             | 77,991 (13.28%)             | 509,340 (86.72%)     |
| Recode not available | 343 (19.60%)                | 1,407 (80.40%)       |

Table 11: Five-Year Survival Counts and Percentages by Primary Site

| Primary Site                         | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|--------------------------------------|-----------------------------|----------------------|
| C50.0-Nipple                         | 976 (16.68%)                | 4,874 (83.32%)       |
| C50.1-Central portion of breast      | 9,777 (19.49%)              | 40,394 (80.51%)      |
| C50.2-Upper-inner quadrant of breast | 13,218 (12.93%)             | 89,044 (87.07%)      |
| C50.3-Lower-inner quadrant of breast | 7,331 (14.61%)              | 42,839 (85.39%)      |
| C50.4-Upper-outer quadrant of breast | 44,329 (14.51%)             | 261,217 (85.49%)     |
| C50.5-Lower-outer quadrant of breast | 9,729 (15.15%)              | 54,496 (84.85%)      |
| C50.6-Axillary tail of breast        | 1,187 (20.45%)              | 4,617 (79.55%)       |
| C50.8-Overlapping lesion of breast   | 32,430 (16.29%)             | 166,640 (83.71%)     |
| C50.9-Breast, NOS                    | 41,597 (29.12%)             | 101,228 (70.88%)     |

Table 12: Five-Year Survival Counts and Percentages by Race/Ethnicity

| Race/Ethnicity                                | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|-----------------------------------------------|-----------------------------|----------------------|
| Hispanic (All Races)                          | 17,007 (17.86%)             | 78,230 (82.14%)      |
| Non-Hispanic<br>American Indian/Alaska Native | 932 (19.45%)                | 3,861 (80.55%)       |
| Non-Hispanic<br>Asian or Pacific Islander     | 9,676 (13.61%)              | 61,409 (86.39%)      |
| Non-Hispanic Black                            | 23,525 (25.95%)             | 67,143 (74.05%)      |
| Non-Hispanic Unknown Race                     | 131 (4.80%)                 | 2,598 (95.20%)       |
| Non-Hispanic White                            | 109,303 (16.53%)            | 552,108 (83.47%)     |

Table 13: Five-Year Survival Counts and Percentages by Radiation Therapy Type

| Radiation Therapy Type               | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|--------------------------------------|-----------------------------|----------------------|
| Beam radiation                       | 60,325 (14.53%)             | 354,868 (85.47%)     |
| Combination beam + implants/isotopes | 218 (12.68%)                | 1,501 (87.32%)       |
| None/Unknown                         | 91,103 (20.10%)             | 362,065 (79.90%)     |
| Radiation NOS (unspecified)          | 1,088 (20.89%)              | 4,120 (79.11%)       |
| Radioactive implants (brachytherapy) | 751 (4.79%)                 | 14,943 (95.21%)      |
| Radioisotopes                        | 44 (12.19%)                 | 317 (87.81%)         |
| Recommended, unknown if administered | 3,446 (16.68%)              | 17,219 (83.32%)      |
| Refused (1988+)                      | 3,599 (25.86%)              | 10,316 (74.14%)      |

Table 14: Five-Year Survival Counts and Percentages by Rural-Urban Continuum Code

| Rural-Urban Continuum Code                                   | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|--------------------------------------------------------------|-----------------------------|----------------------|
| Counties in metropolitan areas $\geq$ 1 million pop          | 95,820 (17.10%)             | 464,633 (82.90%)     |
| Counties in metropolitan areas of 250,000 to 1 million pop   | 32,171 (16.80%)             | 159,334 (83.20%)     |
| Counties in metropolitan areas of < 250,000 pop              | 12,955 (18.14%)             | 58,465 (81.86%)      |
| Nonmetropolitan counties adjacent to a metropolitan area     | 11,622 (19.20%)             | 48,901 (80.80%)      |
| Nonmetropolitan counties not adjacent to a metropolitan area | 7,822 (19.09%)              | 33,159 (80.91%)      |
| Unknown/missing/no match (Alaska or Hawaii - Entire State)   | 170 (17.91%)                | 779 (82.09%)         |
| Unknown/missing/no match/Not 1990-2022                       | 14 (15.22%)                 | 78 (84.78%)          |

Table 15: Five-Year Survival Counts and Percentages by Sex

| Sex    | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|--------|-----------------------------|----------------------|
| Female | 159,057 (17.30%)            | 760,149 (82.70%)     |
| Male   | 1,517 (22.58%)              | 5,200 (77.42%)       |

Table 16: Five-Year Survival Counts and Percentages by Time From Diagnosis to Treatment

| Time to Treatment   | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|---------------------|-----------------------------|----------------------|
| 0-30 days           | 89,855 (17.29%)             | 429,738 (82.71%)     |
| 31-60 days          | 29,722 (12.99%)             | 199,048 (87.01%)     |
| 61-90 days          | 7,084 (13.28%)              | 46,250 (86.72%)      |
| 91-120 days         | 2,265 (16.27%)              | 11,660 (83.73%)      |
| 120 days            | 2,790 (23.52%)              | 9,070 (76.48%)       |
| Unable to calculate | 28,858 (29.32%)             | 69,583 (70.68%)      |

Table 17: Five-Year Survival Counts and Percentages by Total Number of Benign/Borderline Tumors

| Total Number of Benign/Borderline Tumors | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|------------------------------------------|-----------------------------|----------------------|
| 0                                        | 159,573 (17.36%)            | 759,539 (82.64%)     |
| 1                                        | 961 (14.68%)                | 5,584 (85.32%)       |
| 2 or more                                | 40 (15.09%)                 | 225 (84.91%)         |

Table 18: Five-Year Survival Counts and Percentages by Total Number of In Situ Malignant Tumors

| Total Number of In Situ Malignant Tumors | Did Not Survive 5 Years (%) | Survived 5 Years (%) |
|------------------------------------------|-----------------------------|----------------------|
| 01                                       | 111,244 (17.45%)            | 526,352 (82.55%)     |
| 02                                       | 38,979 (17.36%)             | 185,540 (82.64%)     |
| 03                                       | 8,406 (16.53%)              | 42,454 (83.47%)      |
| 04                                       | 1,552 (14.95%)              | 8,826 (85.05%)       |
| 05                                       | 313 (16.21%)                | 1,618 (83.79%)       |
| 6 or more                                | 80 (12.52%)                 | 559 (87.48%)         |

Table 19: Five-Year Survival Counts and Percentages by Tumor Size

| <b>Tumor Size</b>                                               | <b>Did Not Survive 5 Years (%)</b> | <b>Survived 5 Years (%)</b> |
|-----------------------------------------------------------------|------------------------------------|-----------------------------|
| ≤ 1 mm                                                          | 1,261 (15.25%)                     | 7,006 (84.75%)              |
| 1 mm but ≤ 5 mm                                                 | 3,068 (5.09%)                      | 57,152 (94.91%)             |
| 5 mm but ≤ 10 mm                                                | 9,360 (6.10%)                      | 144,143 (93.90%)            |
| 10 mm but ≤ 20 mm                                               | 33,811 (10.89%)                    | 276,689 (89.11%)            |
| 20-50 mm                                                        | 63,876 (23.55%)                    | 207,313 (76.45%)            |
| 50 mm                                                           | 24,387 (42.28%)                    | 33,289 (57.72%)             |
| 990 (microscopic focus)                                         | 669 (5.69%)                        | 11,093 (94.31%)             |
| 998 (site-specific code)                                        | 4,016 (70.11%)                     | 1,712 (29.89%)              |
| Unknown or unreasonable size (includes any tumor sizes 401-989) | 20,121 (42.75%)                    | 26,948 (57.25%)             |

Table 20: Five-Year Survival Counts and Percentages by Year of Diagnosis

| <b>Year of Diagnosis</b> | <b>Did Not Survive 5 Years (%)</b> | <b>Survived 5 Years (%)</b> |
|--------------------------|------------------------------------|-----------------------------|
| 2000-2002                | 31,828 (21.87%)                    | 113,736 (78.13%)            |
| 2003-2005                | 30,477 (20.89%)                    | 115,400 (79.11%)            |
| 2006-2008                | 29,599 (18.83%)                    | 127,631 (81.17%)            |
| 2009-2011                | 28,682 (16.94%)                    | 140,593 (83.06%)            |
| 2012-2014                | 25,730 (14.23%)                    | 155,066 (85.77%)            |
| 2015-2016                | 14,258 (11.21%)                    | 112,923 (88.79%)            |

# Kaplan-Meier Survival Curves Appendix

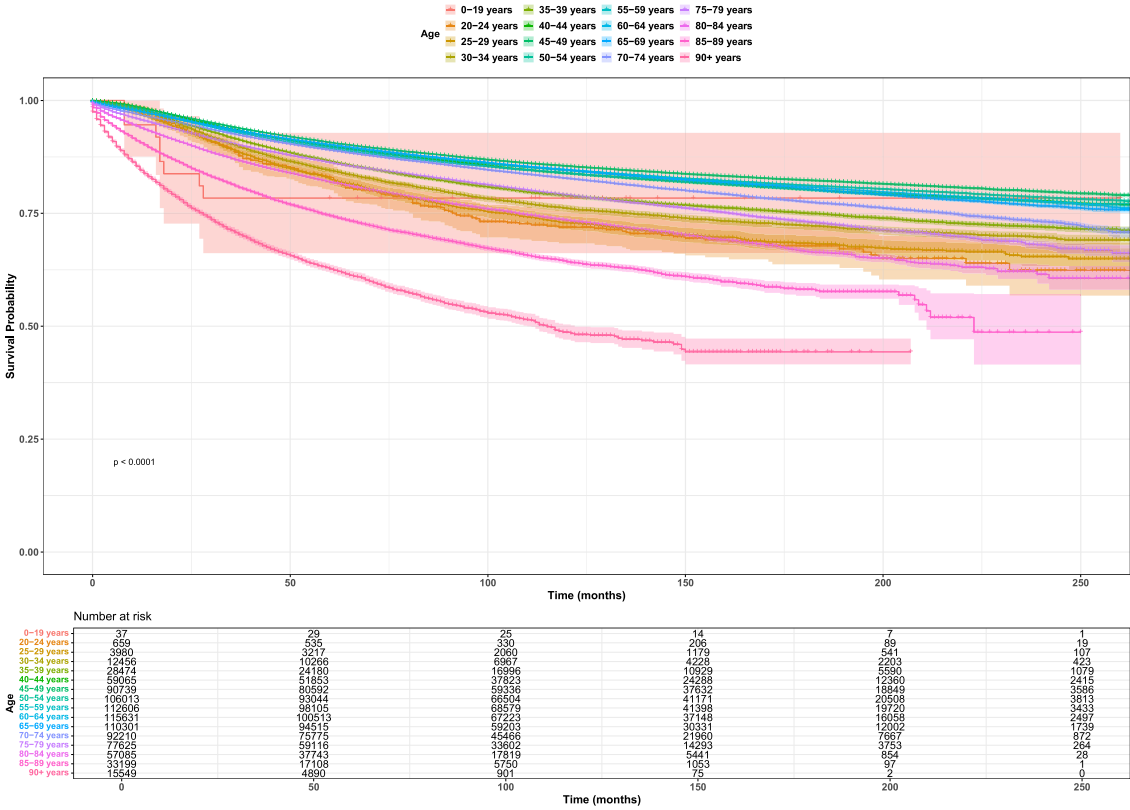


Figure 1: Kaplan-Meier Survival Curve for Age Categorical.

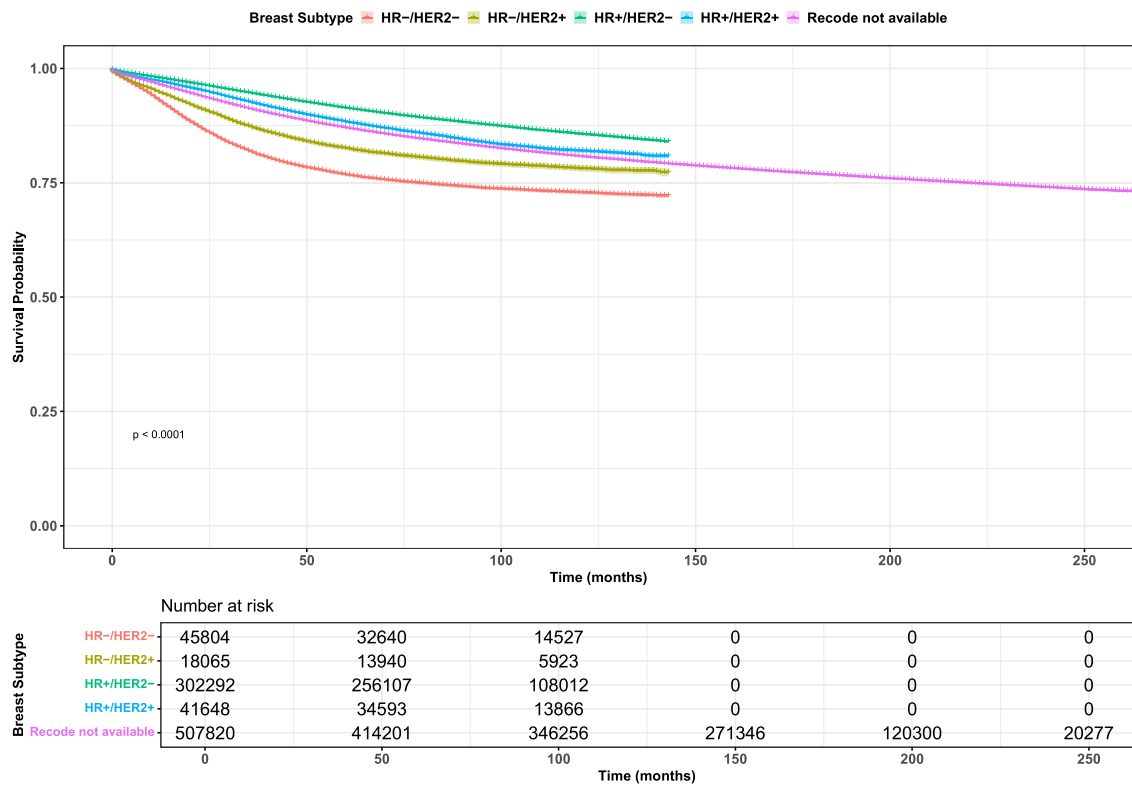


Figure 2: Kaplan-Meier Survival Curve for Breast Cancer Subtype.

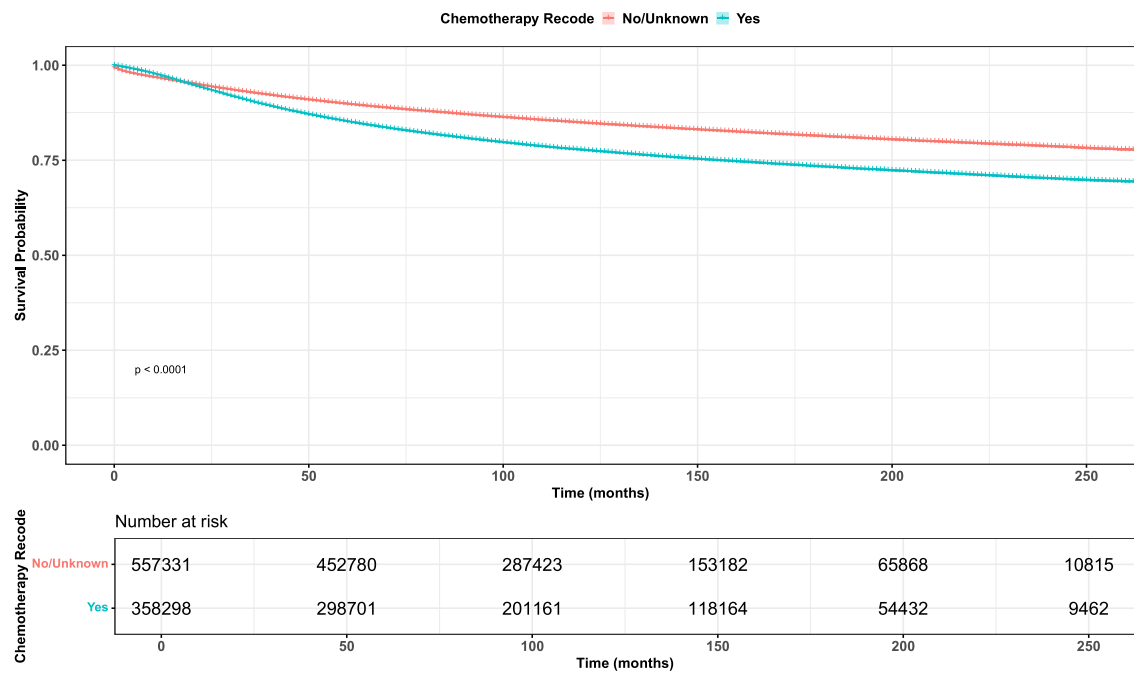


Figure 3: Kaplan-Meier Survival Curve for Chemotherapy Status.

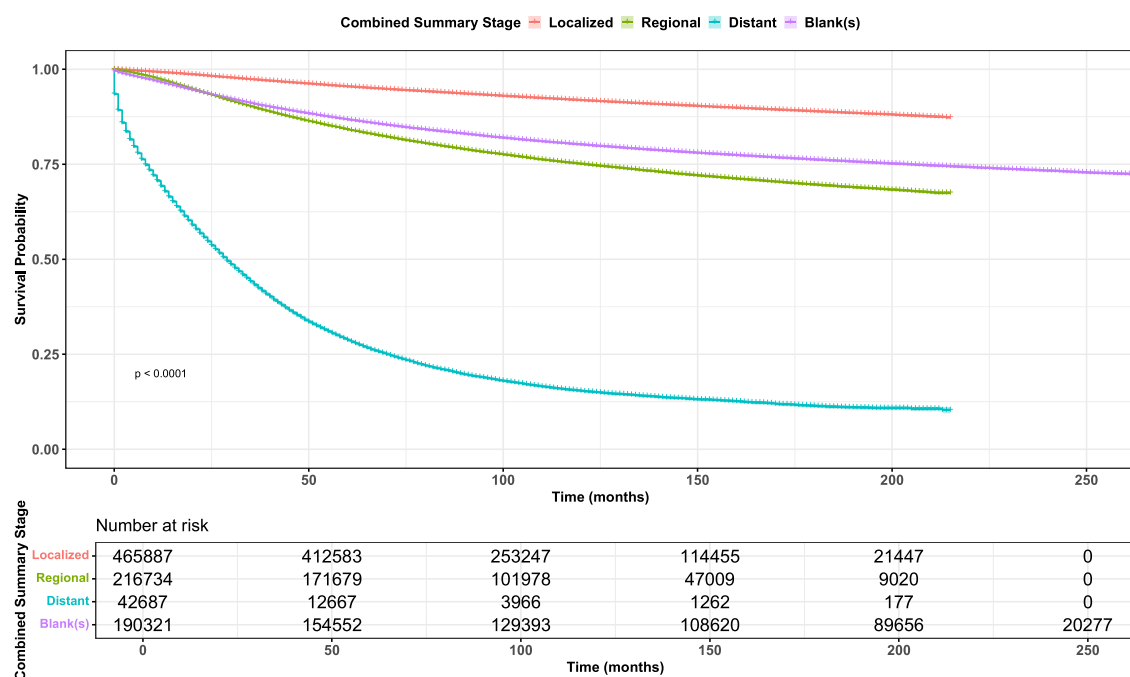


Figure 4: Kaplan-Meier Survival Curve for Combined Summary Stage.

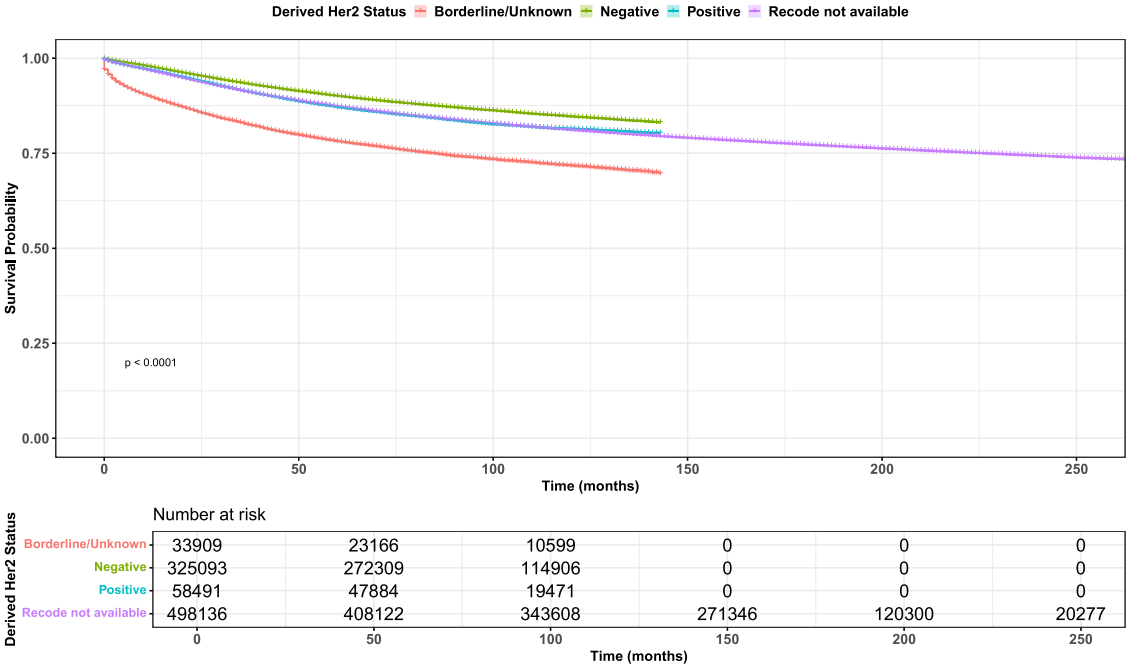


Figure 5: Kaplan-Meier Survival Curve for Derived HER2 Status.

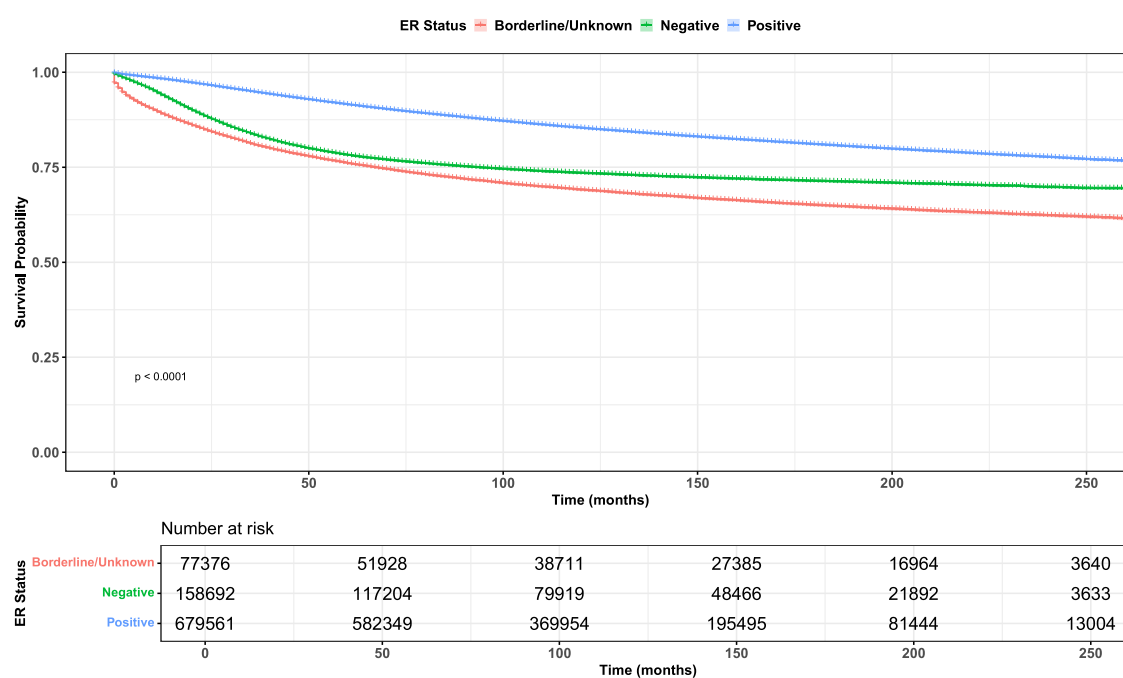


Figure 6: Kaplan-Meier Survival Curve for ER Status.

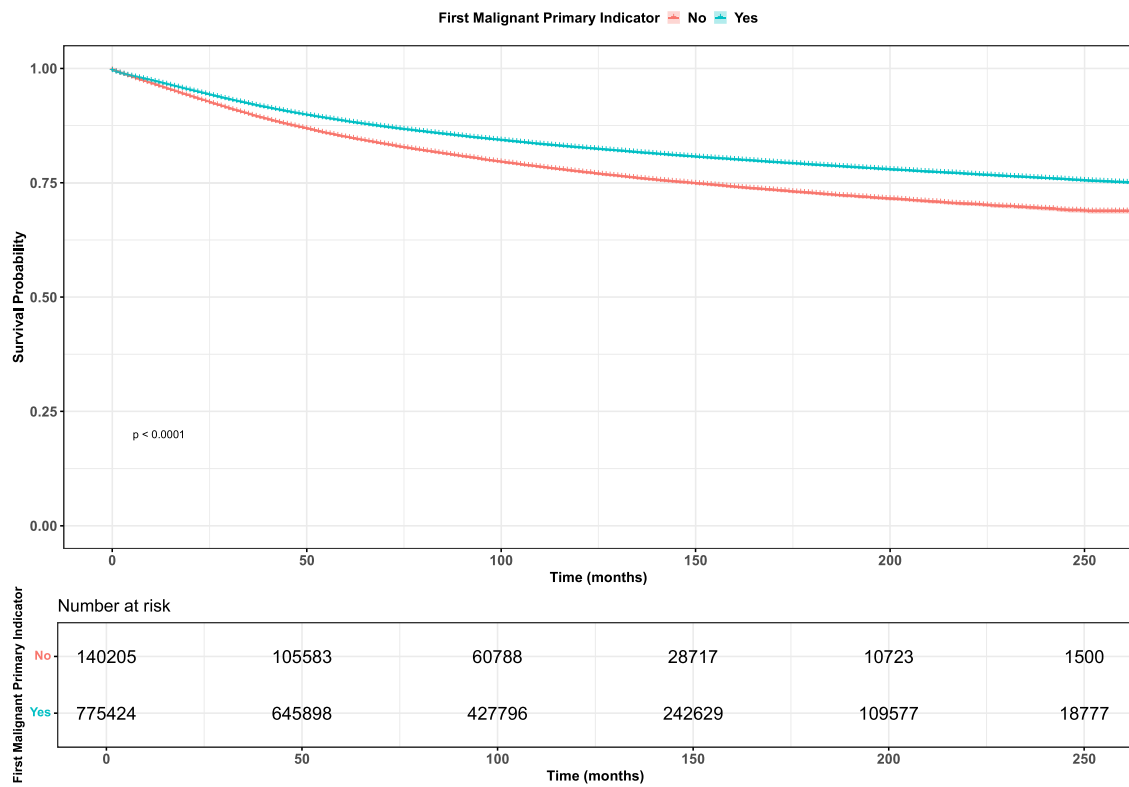


Figure 7: Kaplan-Meier Survival Curve for First Malignant Primary Indicator.

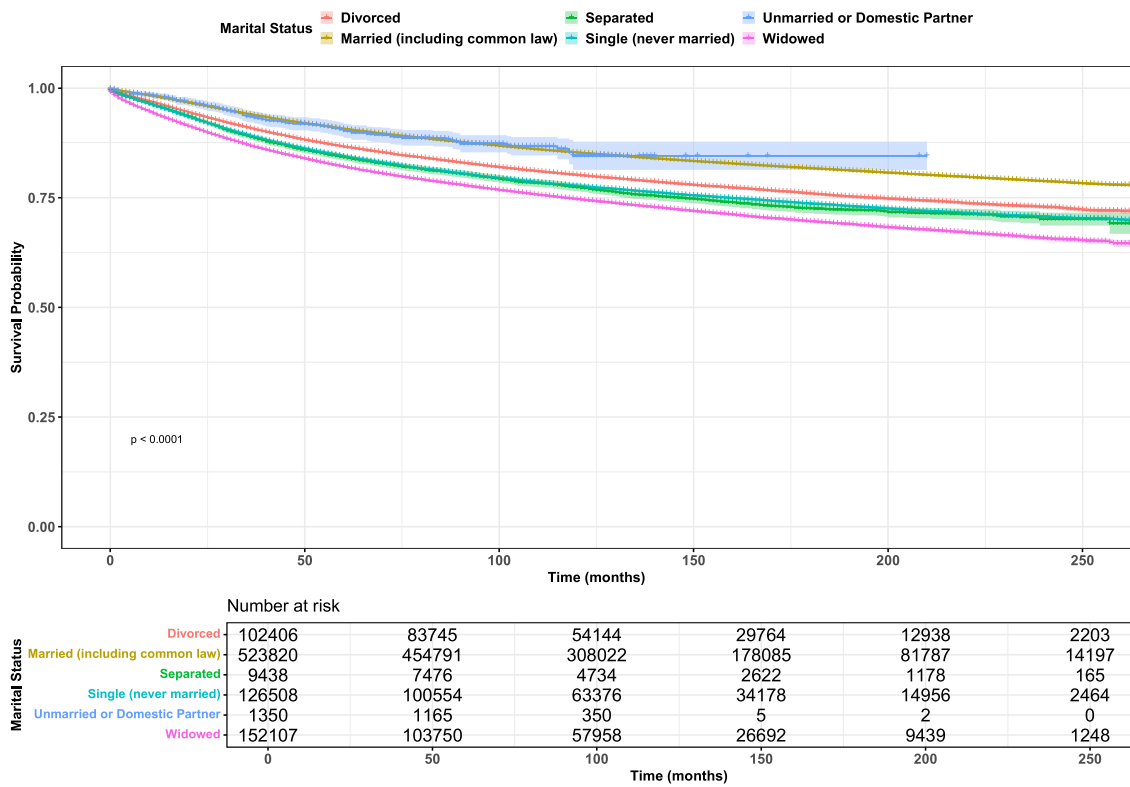


Figure 8: Kaplan-Meier Survival Curve for Marital Status at Diagnosis.

Determinants of Breast Cancer Survival: A Registry-Based Analysis with Classical and Visual Analytical Approaches

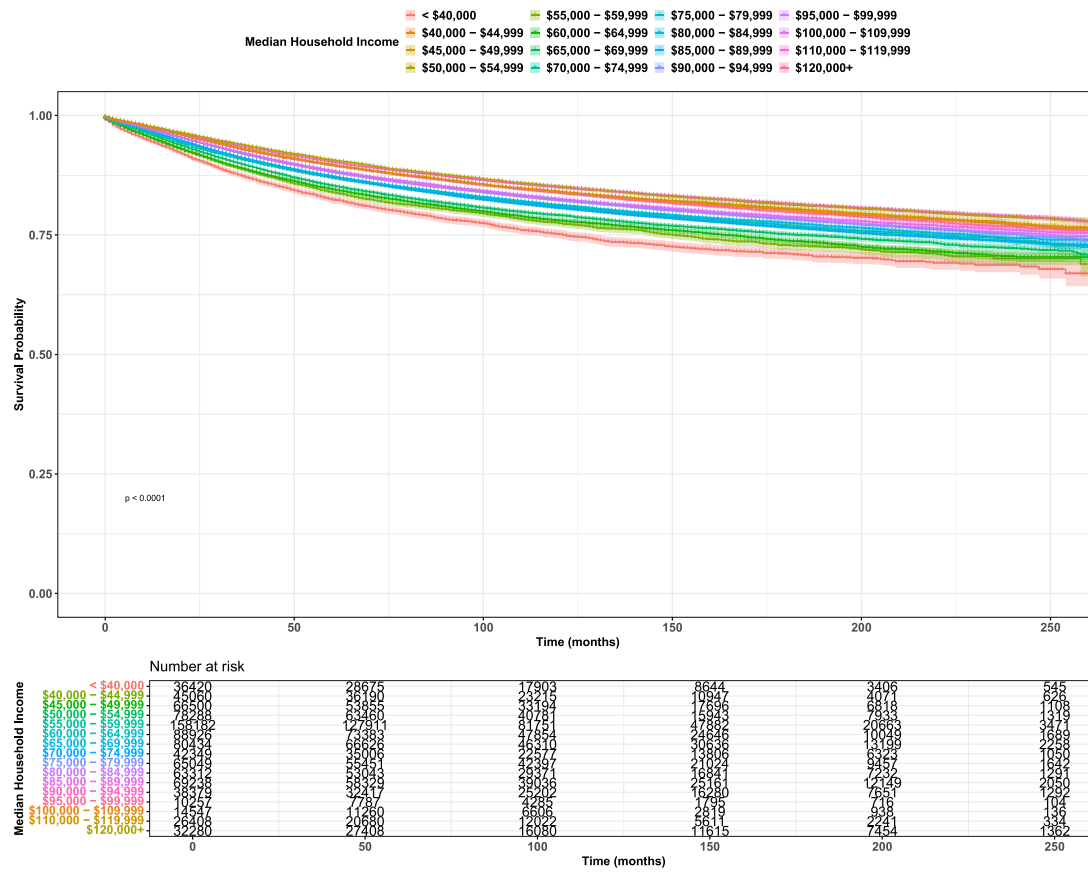


Figure 9: Kaplan-Meier Survival Curve for Median Household Income (Inflation Adjusted to 2022).

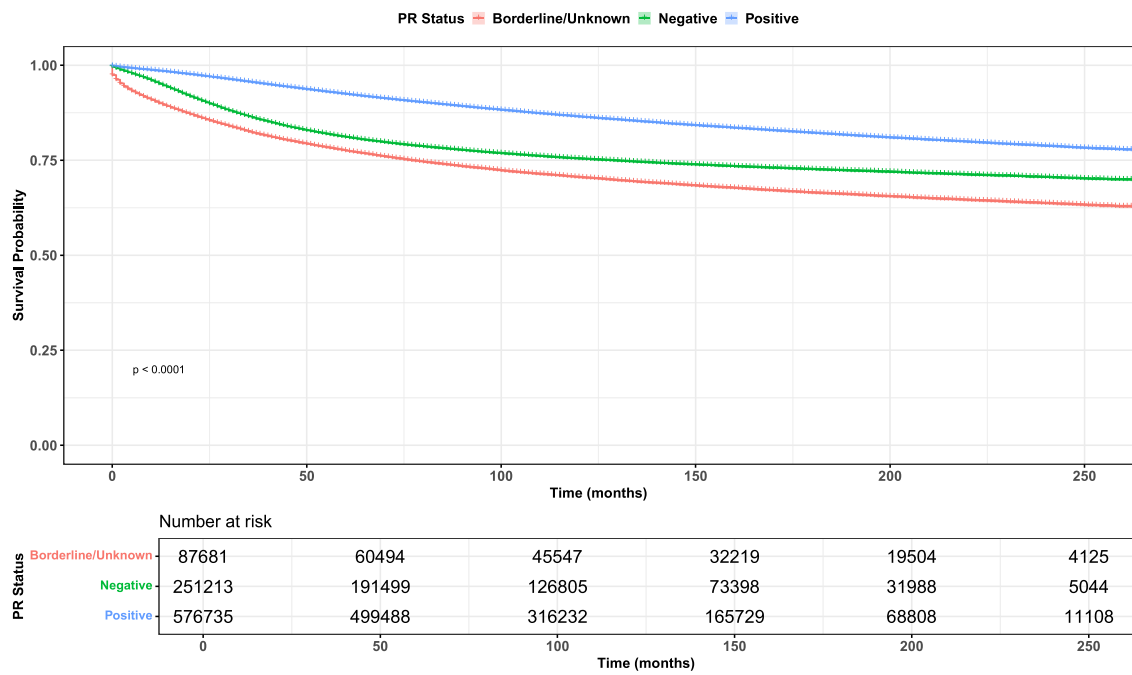


Figure 10: Kaplan-Meier Survival Curve for PR Status.

# Determinants of Breast Cancer Survival: A Registry-Based Analysis with Classical and Visual Analytical Approaches

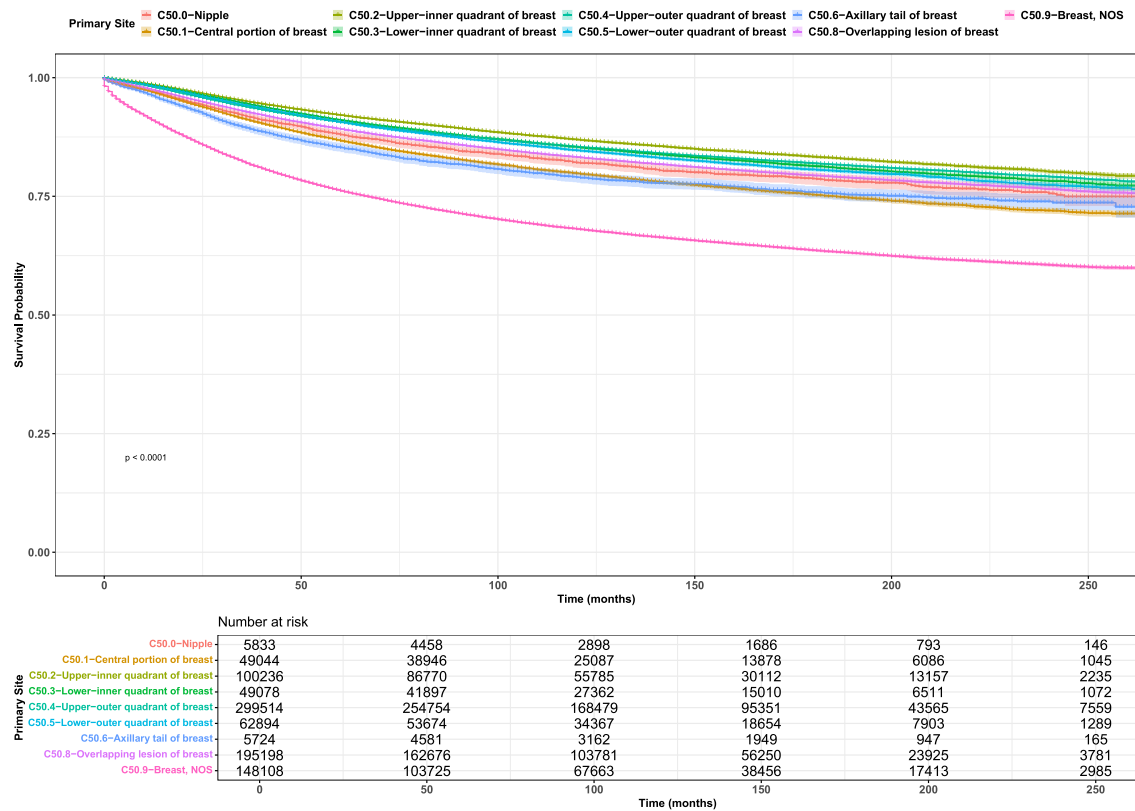


Figure 11: Kaplan-Meier Survival Curve for Primary Site.

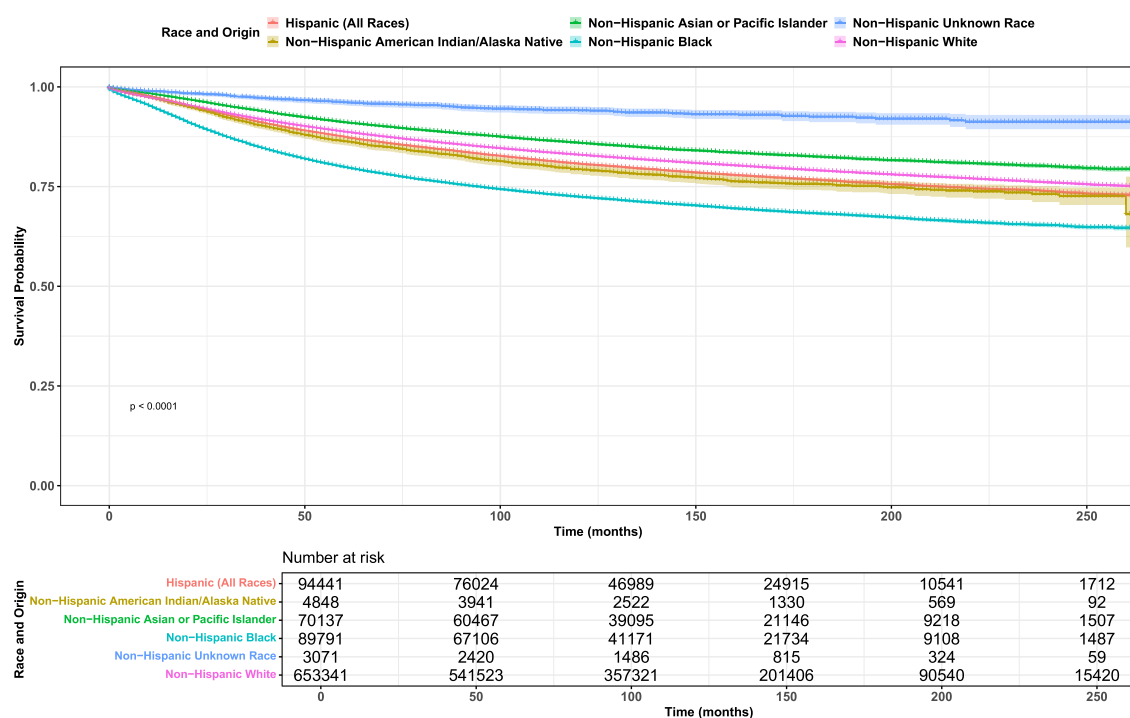


Figure 12: Kaplan-Meier Survival Curve for Race/Ethnicity.

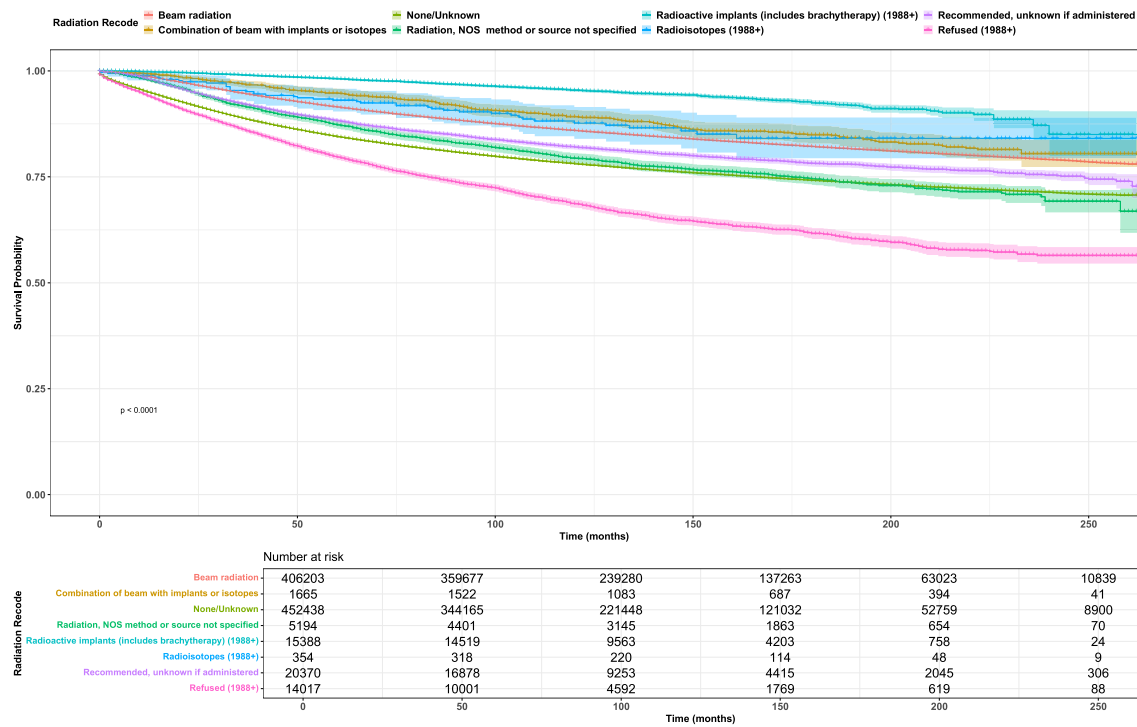


Figure 13: Kaplan-Meier Survival Curve for Radiation Therapy Recode.

# Determinants of Breast Cancer Survival: A Registry-Based Analysis with Classical and Visual Analytical Approaches

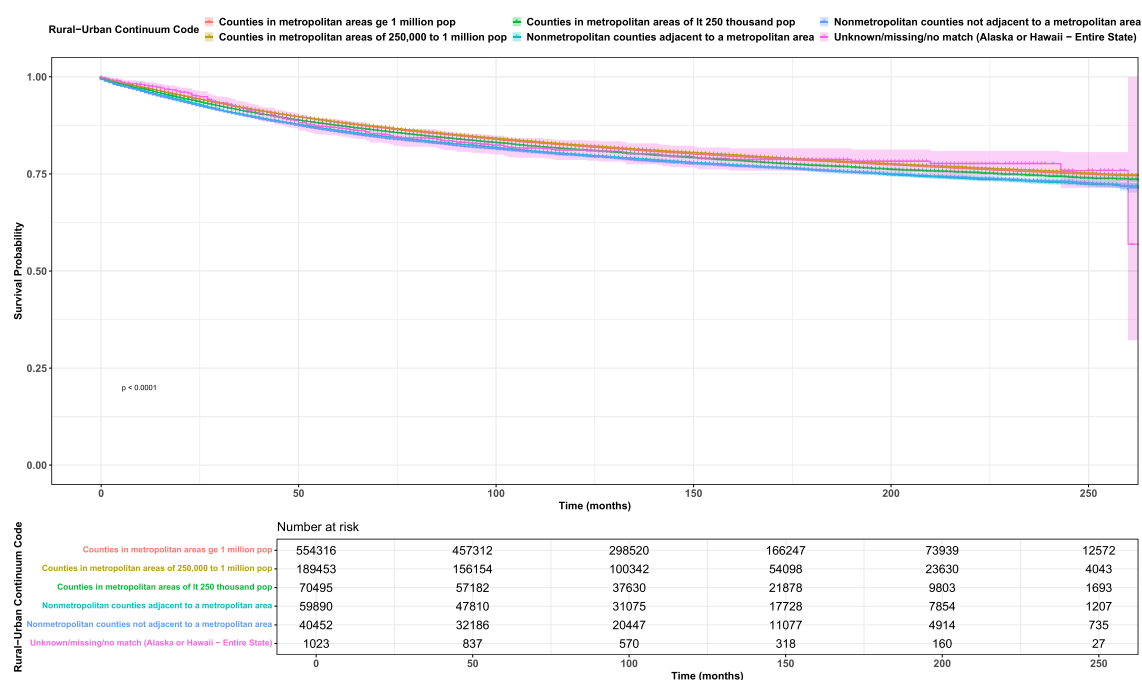


Figure 14: Kaplan-Meier Survival Curve for PR Status.

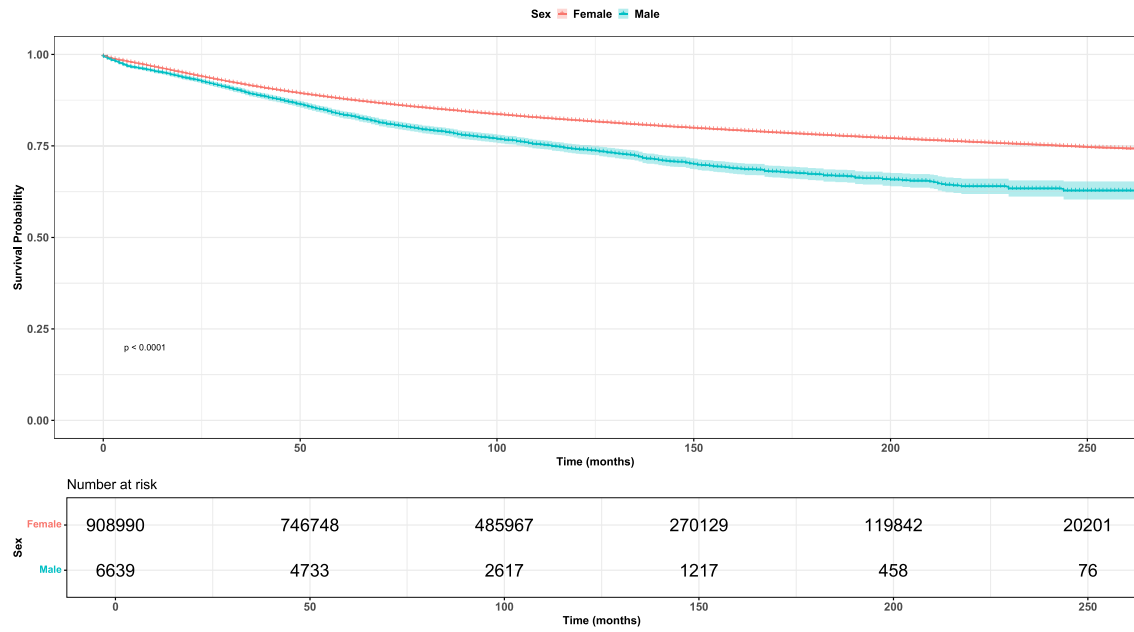


Figure 15: Kaplan-Meier Survival Curve for Patients' Sex.

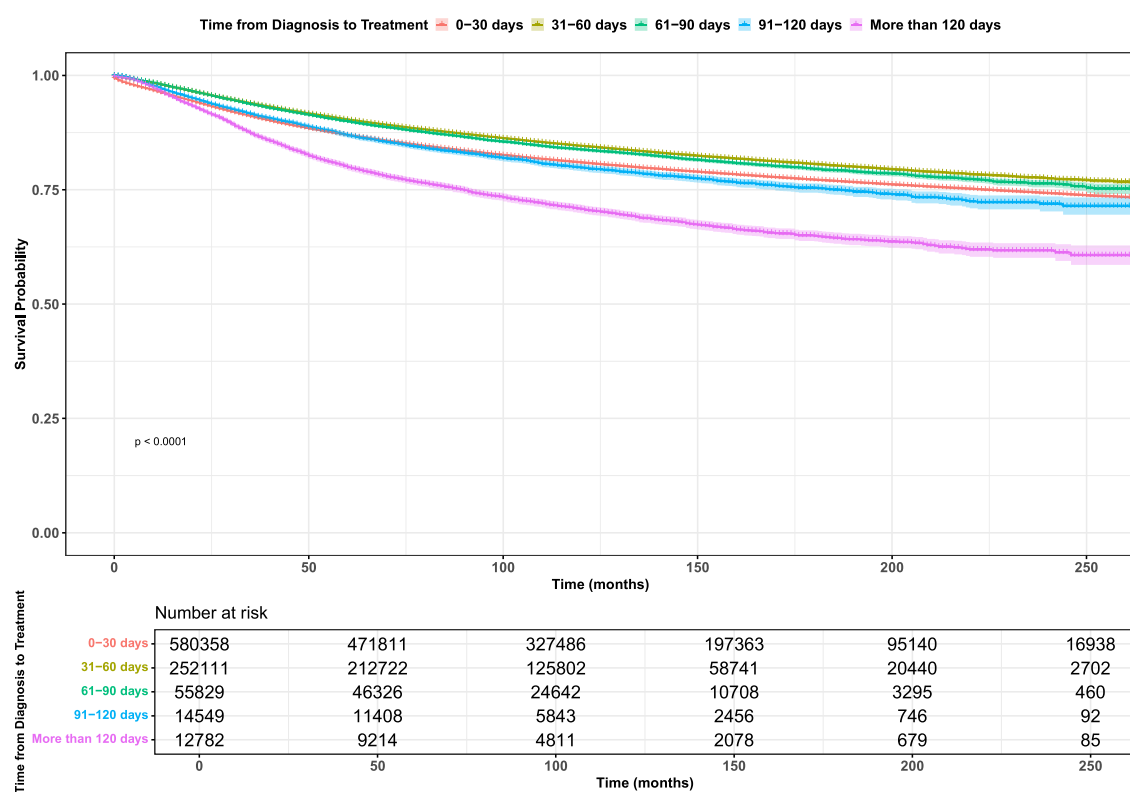


Figure 16: Kaplan-Meier Survival Curve for Time From Diagnosis to Treatment.

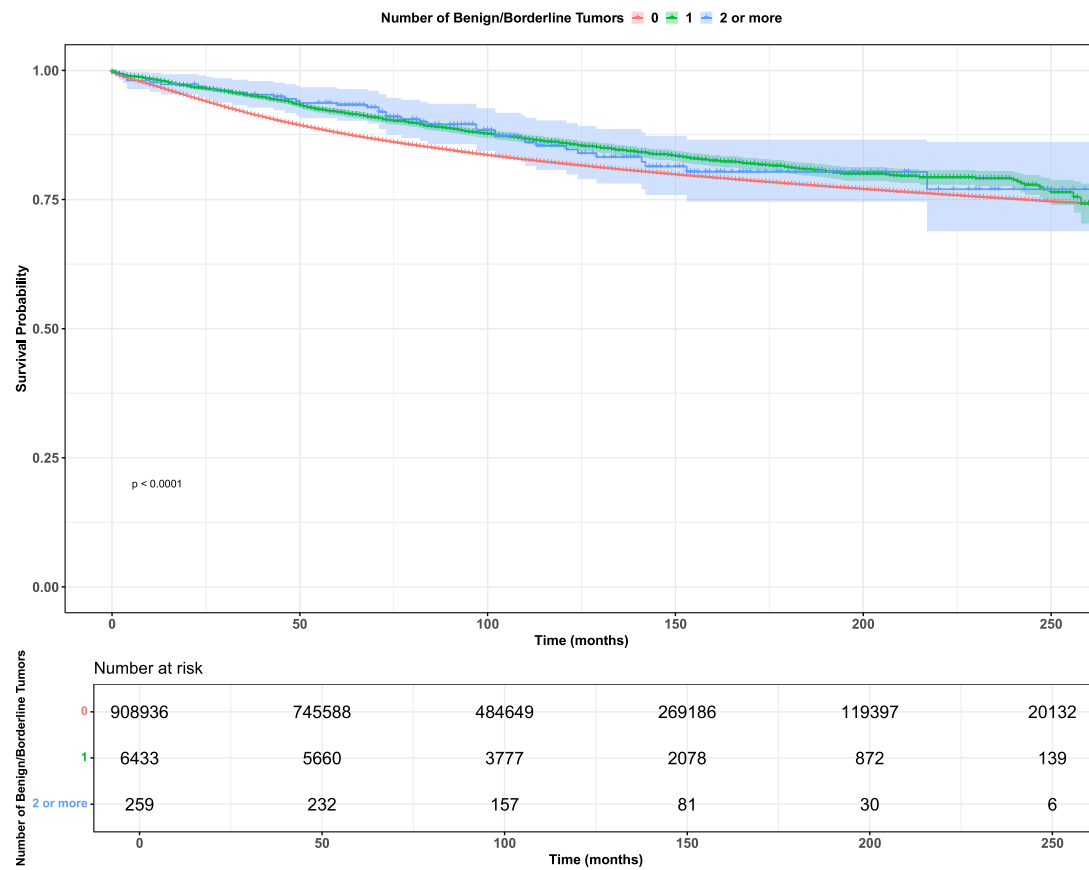


Figure 17: Kaplan-Meier Survival Curve for Total Number of Benign/Borderline Tumors.

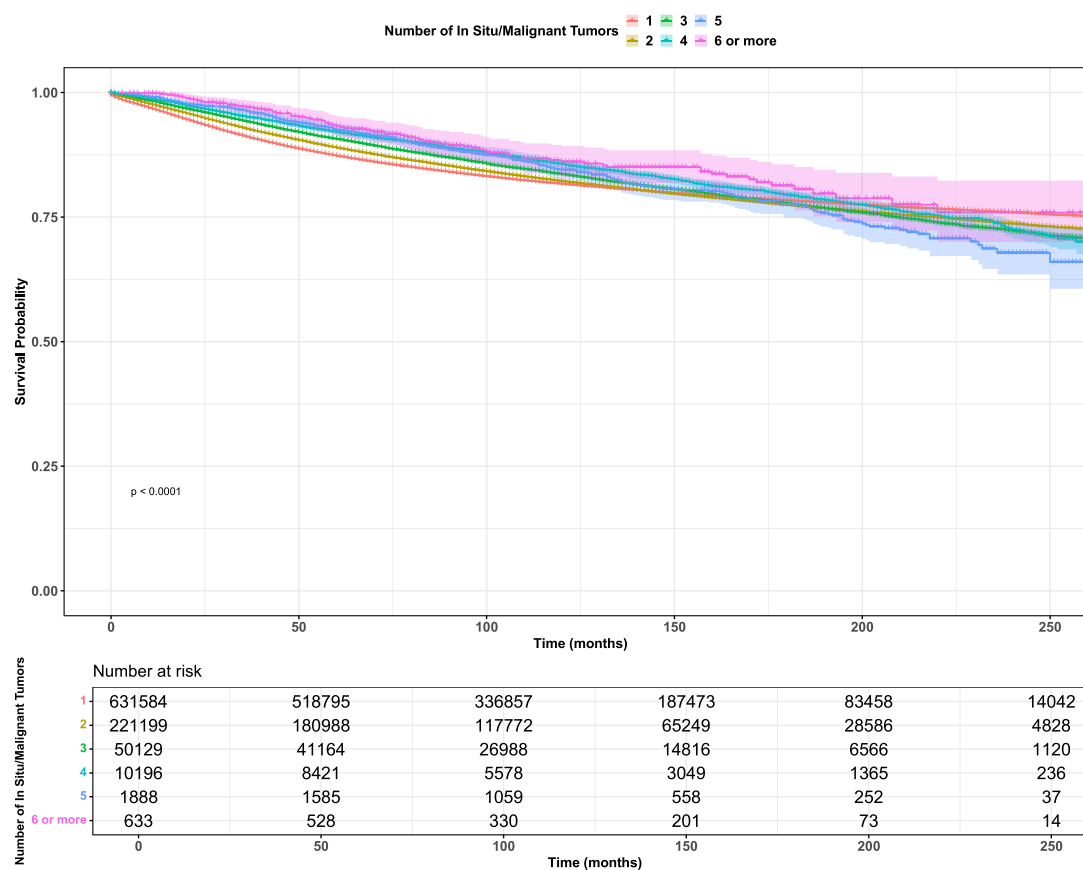


Figure 18: Kaplan-Meier Survival Curve for Total Number of In Situ Malignant Tumors.

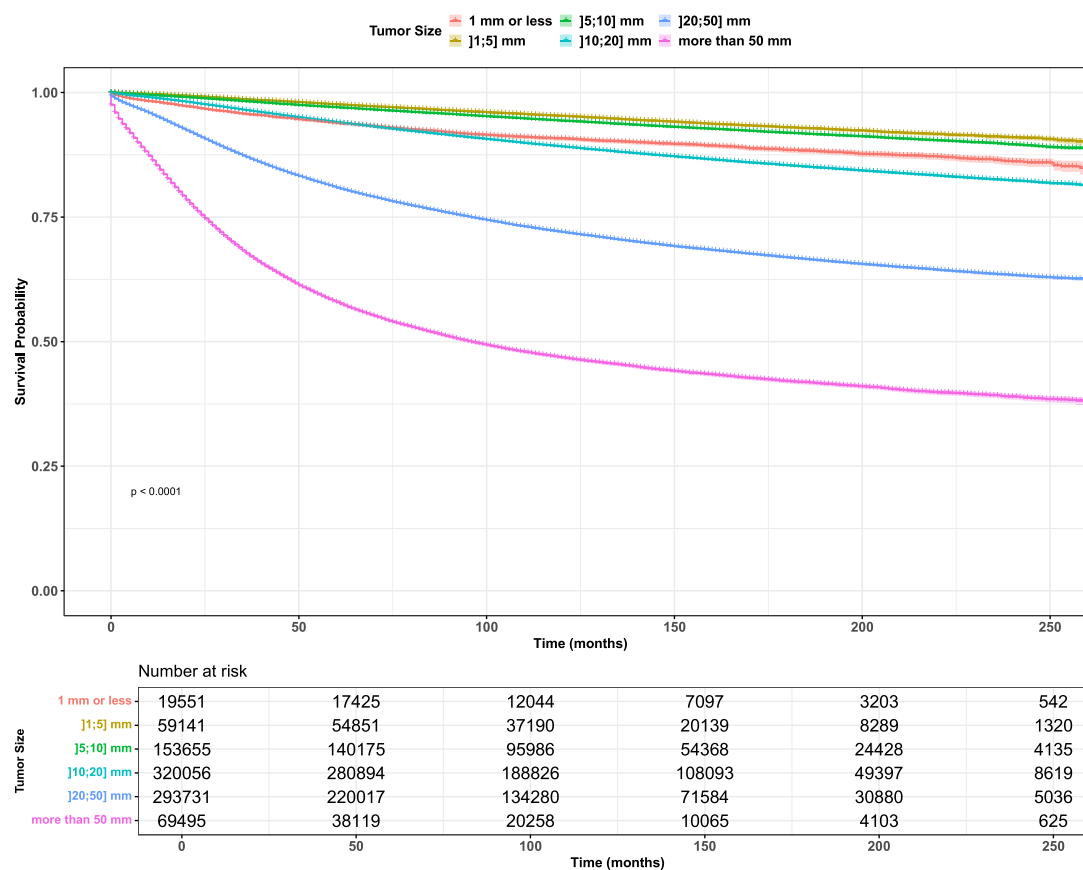


Figure 19: Kaplan-Meier Survival Curve for Tumor Size.

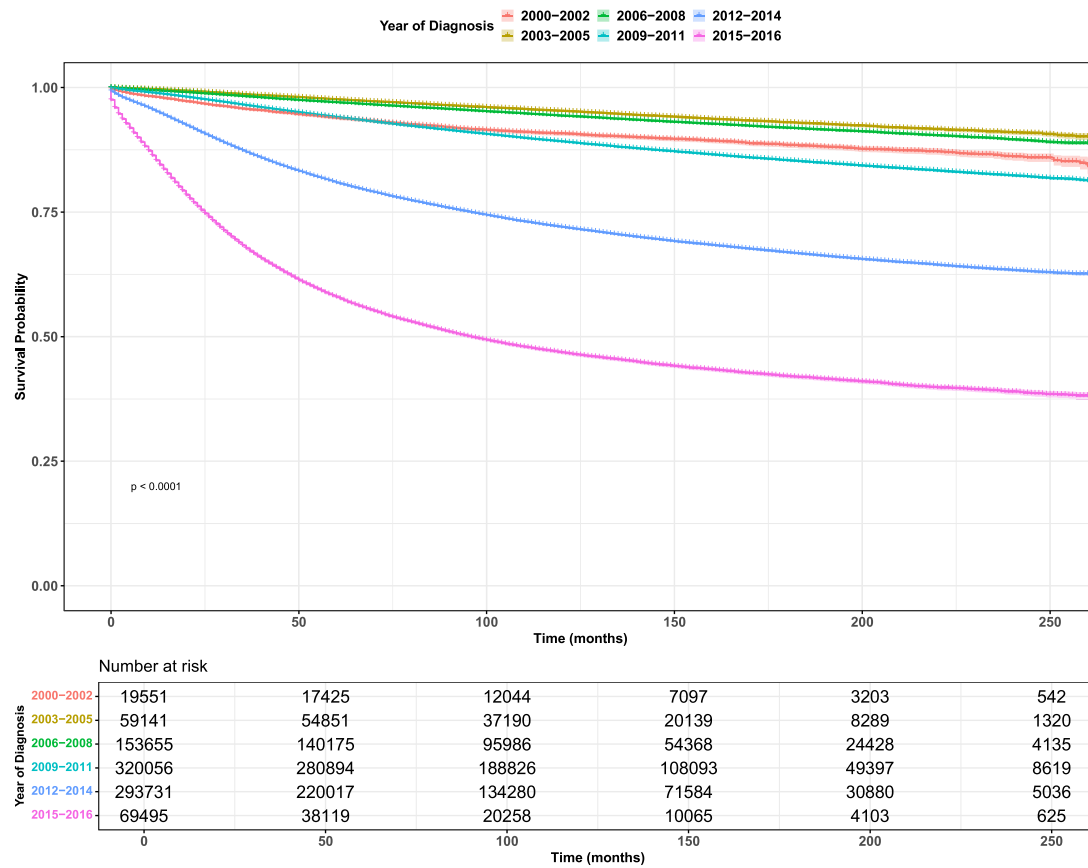


Figure 20: Kaplan-Meier Survival Curve for Year of Diagnosis.