Executive Summary

Abstract

Breast cancer remains one of the leading causes of cancer-related mortality worldwide, with survival outcomes heavily influenced by genetic, demographic, and clinical factors. This study explores how these factors impact breast cancer susceptibility, progression, and prognosis, using statistical modeling and survival analysis techniques. Our findings confirm that genetic predisposition plays a crucial role in breast cancer risk, with BRCA1 and BRCA2 mutations significantly increasing the likelihood of developing the disease and influencing overall survival. However, most breast cancer cases occur in individuals without a family history, highlighting the importance of non-genetic risk factors.

Our survival analysis demonstrates that age and race are the strongest predictors of breast cancer outcomes. Black/African American patients experience steeper declines in survival probability, while American Indian or Alaska Native individuals exhibit more favorable long-term survival rates. Age-related trends show that younger patients (<30 and 30-39) have significantly better survival outcomes compared to older populations, reinforcing the importance of early detection and targeted screening efforts. While ethnicity had some impact on survival, its effect was less pronounced than race or age.

Furthermore, our investigation into breast cancer screening methods reveals that while mammograms have contributed to reduced mortality rates, they come with challenges such as overdiagnosis and false positives. Despite improvements in early detection, late-stage breast cancer incidence remains stable, raising concerns about the overall effectiveness of current screening strategies in preventing disease progression.

These findings emphasize the need for targeted interventions, particularly in addressing racial disparities, improving healthcare access, and refining breast cancer screening guidelines. Genetic testing for high-risk individuals, improved demographic-specific screening strategies, and continued research on breast cancer progression are critical next steps in reducing mortality rates and improving patient outcomes.

Introduction

Breast cancer is a leading cause of cancer-related mortality worldwide. In the U.S., approximately 13.1% of women will be diagnosed with breast cancer in their lifetime, and early detection plays a crucial role in improving patient outcomes[1]. The 5-year relative survival rate (compared to the general population) after diagnosis is 99% for localized breast cancer, however, it can drop to as low as 32% if detected in later stages[2]. Despite advancements in screening, many cases are still diagnosed at a regional or distant stage (44.7 per 100,000 females in 2021)[3], leading to poorer prognoses and higher mortality rates. Recent trends indicate a 1% annual increase in breast cancer incidence among women under 50 between 2012 and 2021, with a 1.4% annual increase specifically in women under 50[4].

Given these challenges, this study aims to explore what genetic, clinical, and demographic biomarkers can predict breast cancer and ultimately be used to reduce mortality rates:

- Identify key biological and demographic factors that increase susceptibility to breast cancer.
- Determine how genetic mutations and clinical markers correlate with disease progression and survival rates.
- Assess whether integrating biomarker data can lead to more accurate early detection methods.

Key Findings

Our analysis highlights the multifaceted nature of breast cancer risk, survival disparities, and the role of early detection. We found that both genetic and demographic factors significantly impact survival, but their influence varies. While hereditary mutations in BRCA1 and BRCA2 increase susceptibility to breast cancer, the majority of cases arise in individuals without a family history, emphasizing the importance of non-genetic risk factors. Additionally, age and race emerged as the strongest predictors of survival, with Black/African American patients experiencing worse survival outcomes and older patients facing significantly higher mortality rates. Sex and ethnicity had a moderate impact, but their effects became more apparent after balancing the dataset.

Our survival analysis further demonstrated that early detection plays a crucial role in improving patient outcomes, but current screening methods, particularly mammograms, come with limitations such as overdiagnosis and false positives. Although breast cancer mortality has declined over the years, our findings suggest that this decline is influenced by multiple factors beyond just screening, including advancements in treatment and better healthcare access. Addressing racial and ethnic disparities, optimizing screening protocols, and

---

[1] https://seer.cancer.gov/statfacts/html/breast.html
[2] https://www.komen.org/breast-cancer/facts-statistics/breast-cancer-statistics/survival-rates
[3] https://progressreport.cancer.gov/diagnosis/stage
[4] https://pressroom.cancer.org/breastcancerfactsandfigures2024

incorporating genetic risk assessments could further improve breast cancer survival rates and reduce healthcare inequities.

Genetic Factors and Breast Cancer Susceptibility

Approximately 5-10% of breast cancer cases are hereditary, meaning they result directly from gene mutations inherited from a parent. Among these inherited mutations, the most common causes are BRCA1 (Breast Cancer gene 1) and BRCA2 (BReast CAncer gene 2), which encode proteins responsible for repairing damaged DNA. When these genes mutate they can lead to abnormal cell growth and increase the risk of developing breast cancer. In general, BRCA1 and BRCA2 mutations have been demonstrated to be associated with reduced overall survival (OS) in women diagnosed with breast cancer (BC), especially when investigating European populations.

On average, women carrying a mutation in either gene face up to a 70% chance of developing breast cancer by the age of 80. Men who inherit one of these mutated genes are also at an increased risk of breast cancer, albeit at a much smaller likelihood. Approximately 0.2% to 1.2% of men who inherit a harmful BRCA1 mutation and 1.8% to 7.1% of those with a harmful BRCA2 mutation are at risk of developing breast cancer by the age of 70. Although a family history does increase the risk of breast cancer, especially with close blood relatives, the majority of diagnosed cases are not due to hereditary causes.

The following findings also suggest that family history and genetic predisposition significantly impact breast cancer survival rates. Hazard ratios (HR) and 95% confidence interval (CI) were abstracted and pooled with random-effect modeling. (BRCA1: HR = 1.69, 95% CI, 1.35 to 2.12, $p < 0.001$; BRCA2: HR = 1.50, 95% CI 1.03 to 2.19, $p = 0.034$). Subgroup analysis among individuals carrying a BRCA2 mutation revealed that advanced age (equaling or exceeding 45 years) was associated with a statistically significant reduction in overall survival (OS) when compared with younger age groups. Conversely, an improved survival was observed in BC patients who had a BRCA1 mutation and were treated with endocrine therapy; this phenomenon can be attributed, at least in part, to the heightened sensitivity of BCs with BRCA1 mutations to endocrine therapy. In the close relatives perspective, the 5-year survival rate for daughters whose mothers passed away within 5 years was 87%, whereas it was 91% for those whose mothers survived during that period. Sisters had a more significant difference, with 70% survival if their sibling passed away versus 88% if their sibling survived.

Demographic Disparities in Breast Cancer Survival

Our analysis reveals that demographic factors such as race, ethnicity, sex, and age category play a role in breast cancer survival outcomes, but their impact varies. Initially, our results on unbalanced data suggested that only age showed a statistically significant association with survival. However, after balancing the data, race and age both became stronger predictors of survival outcomes. These findings highlight the importance of addressing demographic disparities in breast cancer prognosis and ensuring fair representation in research.

The Kaplan-Meier survival analysis further reinforces these disparities. The KM curve for race shows that Black/African American patients experience a steeper decline in survival probability compared to other racial groups, while American Indian or Alaska Native patients had better survival rates over time. The KM curve for ethnicity indicates that Hispanic patients tend to have lower survival probabilities earlier, but confidence intervals suggest variability in the dataset. Age also remains a strong predictor, with younger patients (<30 and 30-39) displaying better survival trends compared to older groups. These findings align with our Cox regression and Poisson regression models, where race and age were confirmed as significant predictors of survival after balancing the dataset.

Detailed Analysis of Demographic Factors & Survival Trends

The dataset was highly imbalanced, with a significant disparity between patients who survived breast cancer and those who did not. To address this, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples, ensuring a balanced distribution of both survival outcomes. Following are the results:

Before Balancing the Data:

1. Race and Ethnicity: Showed no significant effect on survival in the unbalanced model.
2. Age: Was the only statistically significant predictor of survival (ANOVA $p = 0.011$).
3. Sex: Had no significant impact on survival ($p = 0.29$ in the t-test and Cox model).
4. Cox Regression Results: Suggested that ethnicity might influence survival ($HR = 6.48$, $p = 0.06$), but the confidence interval was wide, making the result inconclusive.
5. Kaplan-Meier Curves: Before balancing, the KM curves showed large confidence intervals, making it difficult to conclude significant survival differences.

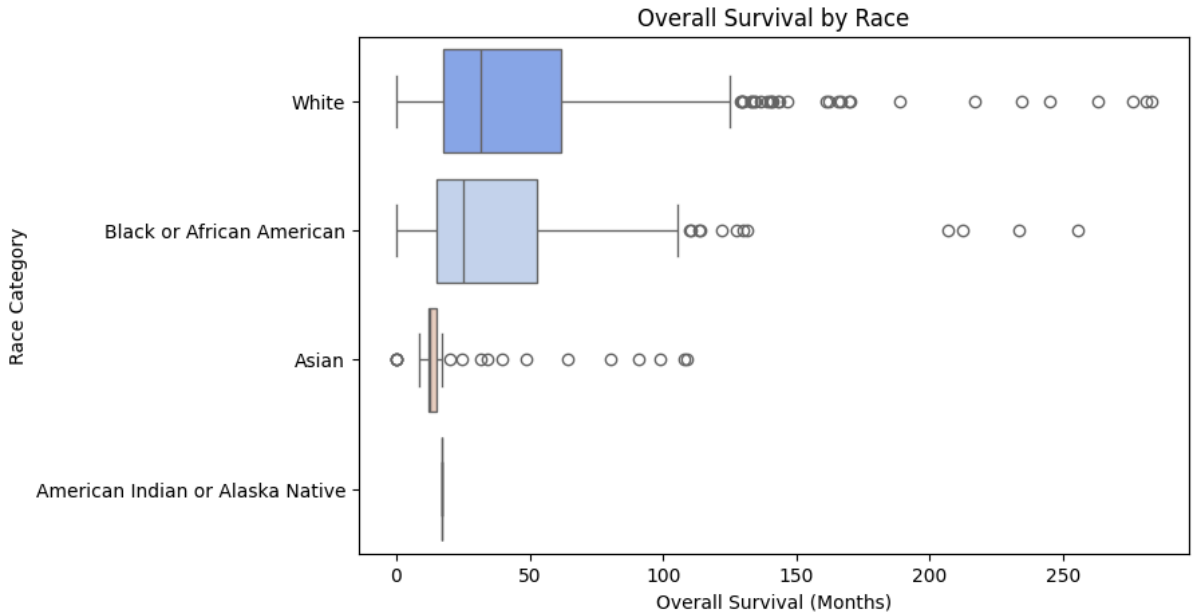

Fig 1: Overall Survival by Race (Boxplot). Compares overall survival distribution across racial groups. Black/African American patients have lower median survival than White and American Indian/Alaska Native individuals.

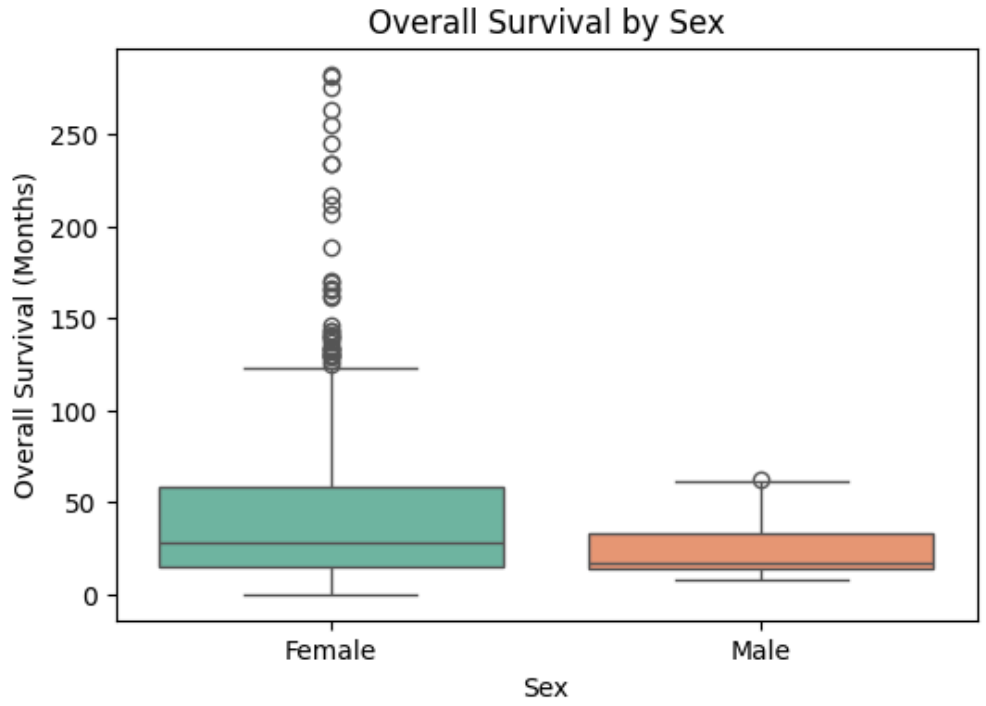

Fig 2: Overall Survival by Sex (Boxplot). Shows the distribution of overall survival months for males vs. females. Females tend to have longer survival times, while males show a lower median survival.

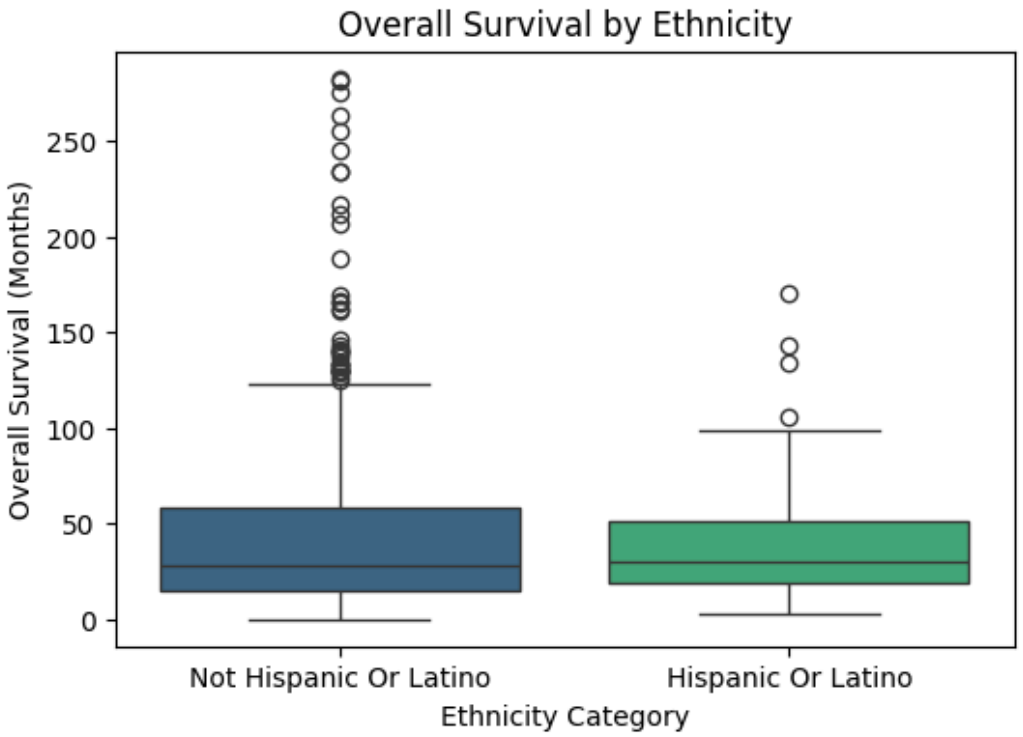

Fig 3: Overall Survival by Ethnicity (Boxplot). Displays survival time differences between Hispanic and Non-Hispanic patients. Hispanic patients show lower median survival times, reinforcing Kaplan-Meier results.

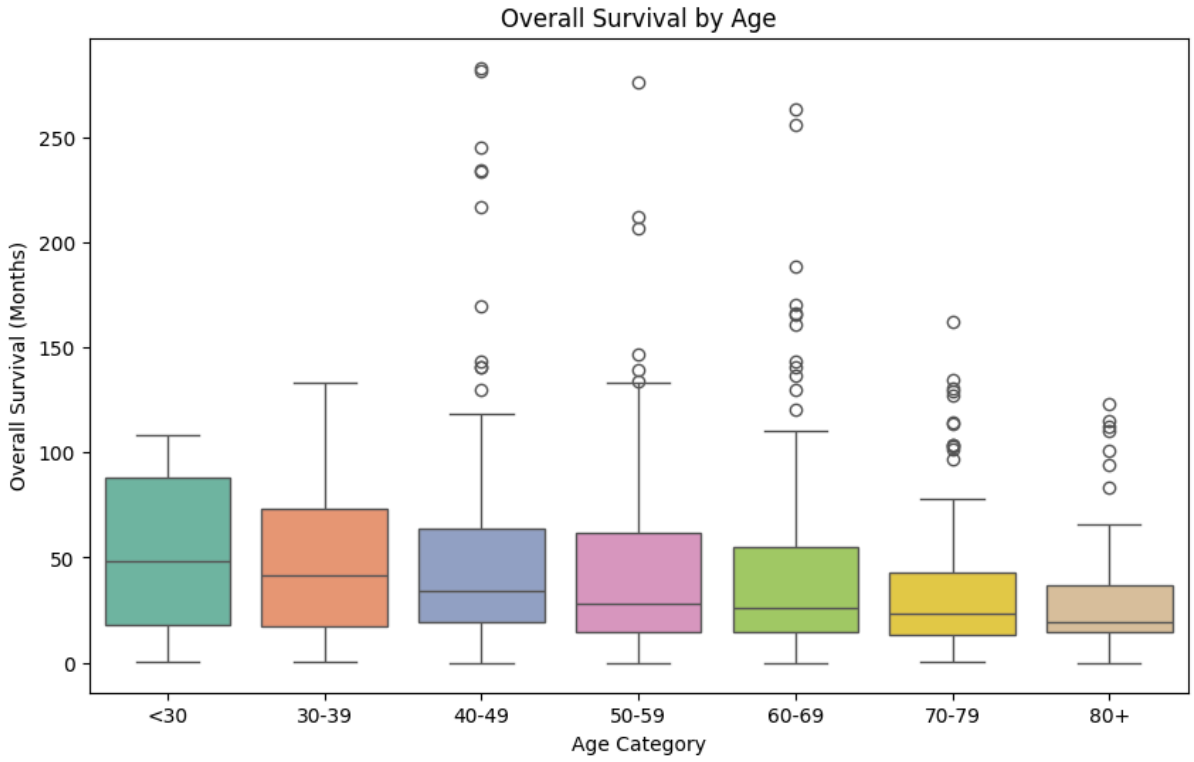

Fig 4: Overall Survival by Age (Boxplot). Highlights the impact of age on survival duration. Older patients (70+) have the shortest survival times, confirming age as a strong predictor of survival.

After Balancing the Data (SMOTE Applied to Survival Status):

1. Race: Became statistically significant in both logistic regression ($p < 0.01$) and Cox regression (HR = 1.22, $p < 0.005$), confirming survival disparities among racial groups.
2. Age: Remained a significant predictor of survival (HR = 1.07, $p = 0.01$), reaffirming that younger patients generally have better outcomes.
3. Ethnicity: Became more relevant in survival models but was still not as strong a predictor as race or age.
4. Sex: Previously insignificant, showed some influence in Poisson regression (IRR = 1.55, $p < 0.005$), meaning males might have shorter survival times, but the Cox model results were still inconclusive.

Kaplan-Meier Curves:

1. Race KM Curve: Showed that Black/African American patients had the worst survival probabilities, while American Indian or Alaska Native patients had better long-term survival.

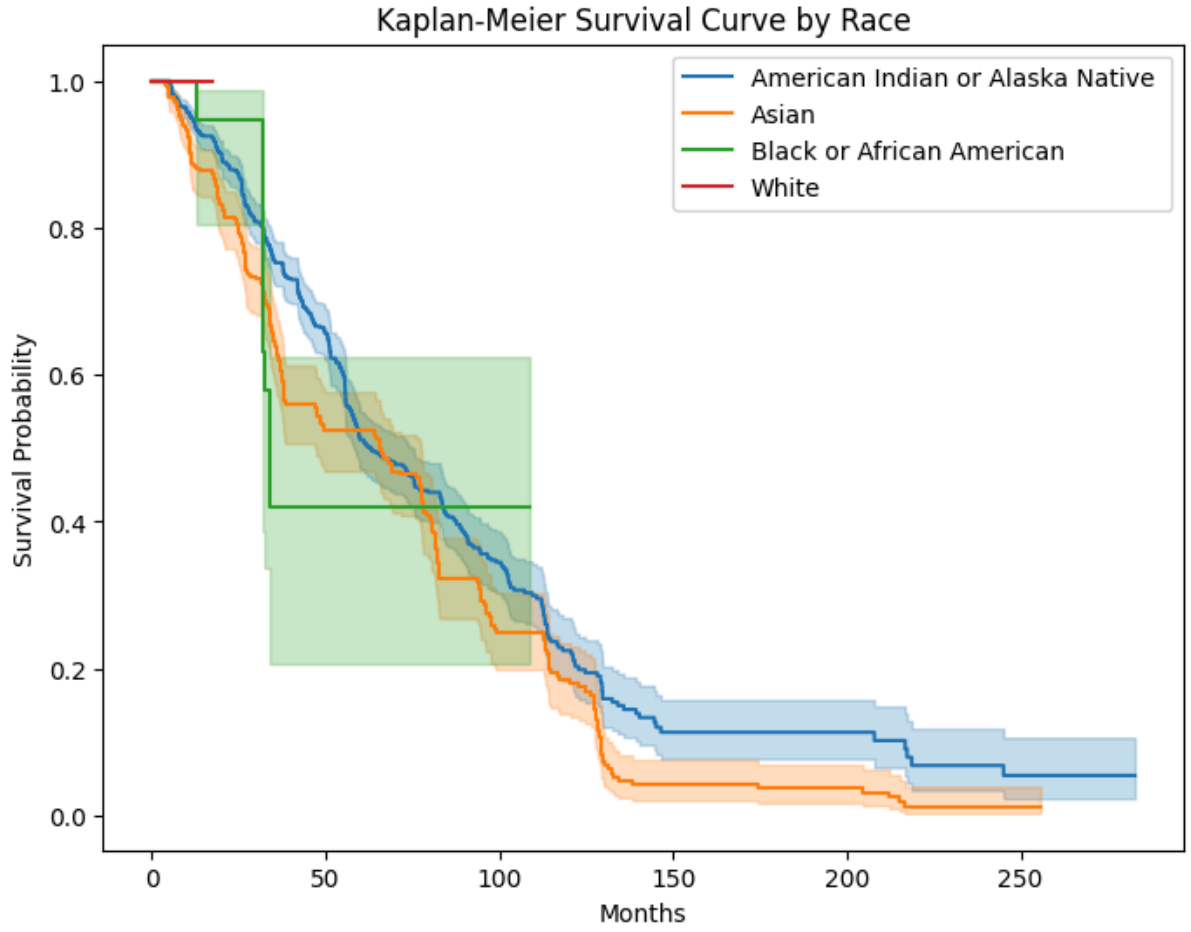

Fig 5: Kaplan-Meier Survival Curve by Race. Illustrates survival probability across different racial groups. Black/African American patients experience steeper declines in survival, while American Indian or Alaska Native individuals have better long-term survival.

2. Ethnicity KM Curve: Hispanic patients had lower survival probability early on, reinforcing the need for targeted screening and intervention programs.

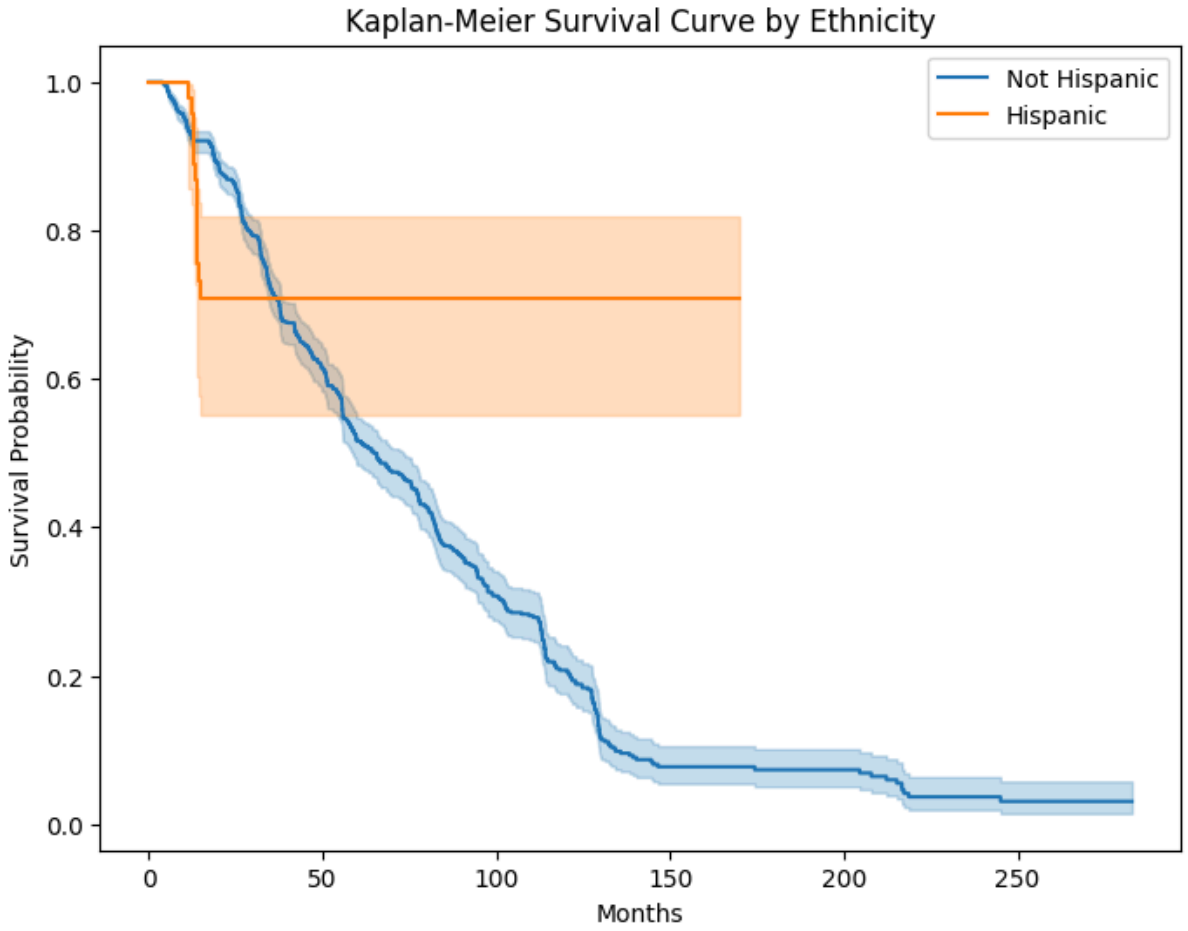

Fig 6: Kaplan-Meier Survival Curve by Ethnicity. Examines survival differences between Hispanic and Non-Hispanic patients. Hispanic patients tend to have lower survival probability early on, but confidence intervals suggest high variability.

3. Age KM Curve: Showed that younger patients (<30 and 30-39) had better survival trends compared to older groups, supporting Cox regression findings.

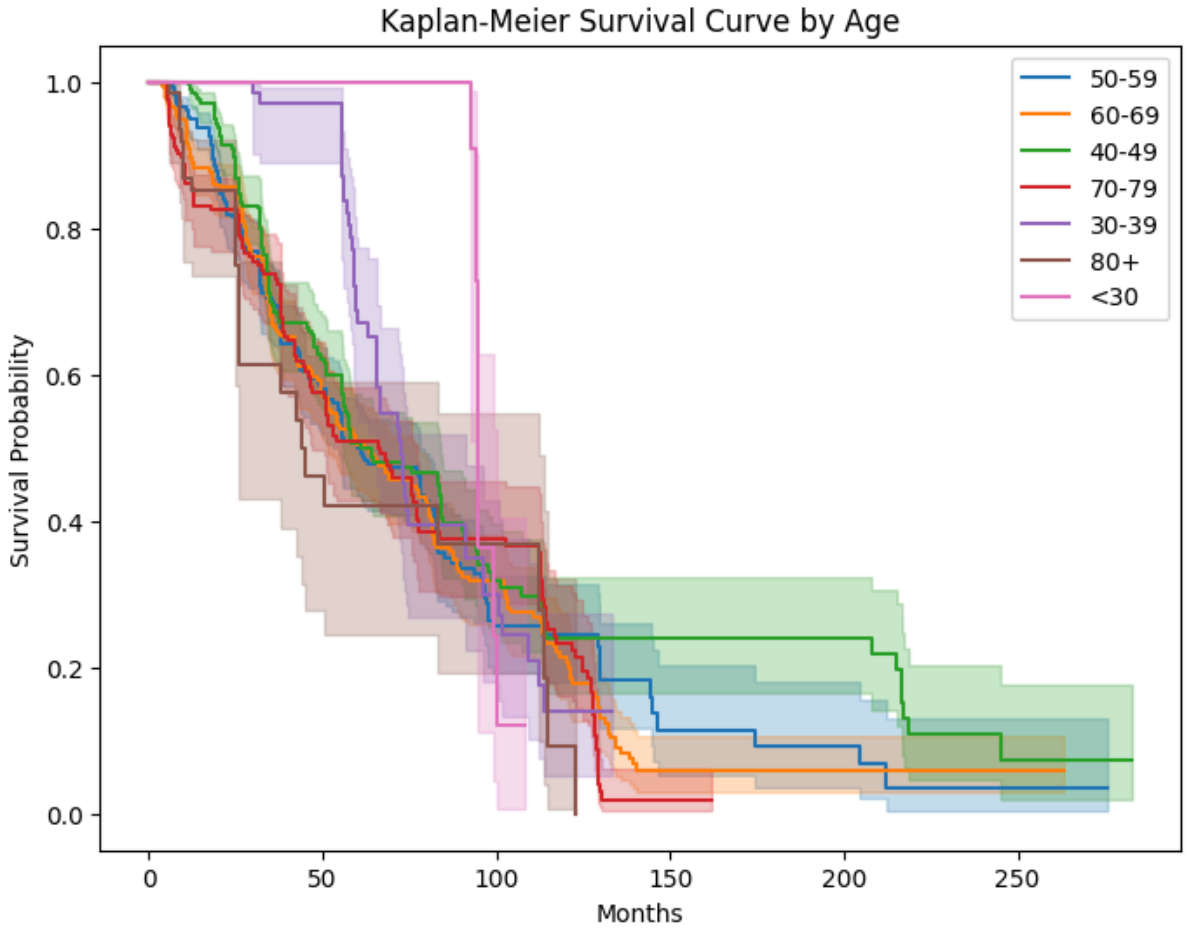

Fig 7: Kaplan-Meier Survival Curve by Age. Visualizes survival probability over time across different age groups. Younger patients (<30 and 30-39) have significantly better survival rates compared to older groups.

4. Sex KM Curve: It shows that females generally have better survival probabilities, but confidence intervals for males are wide, indicating variability

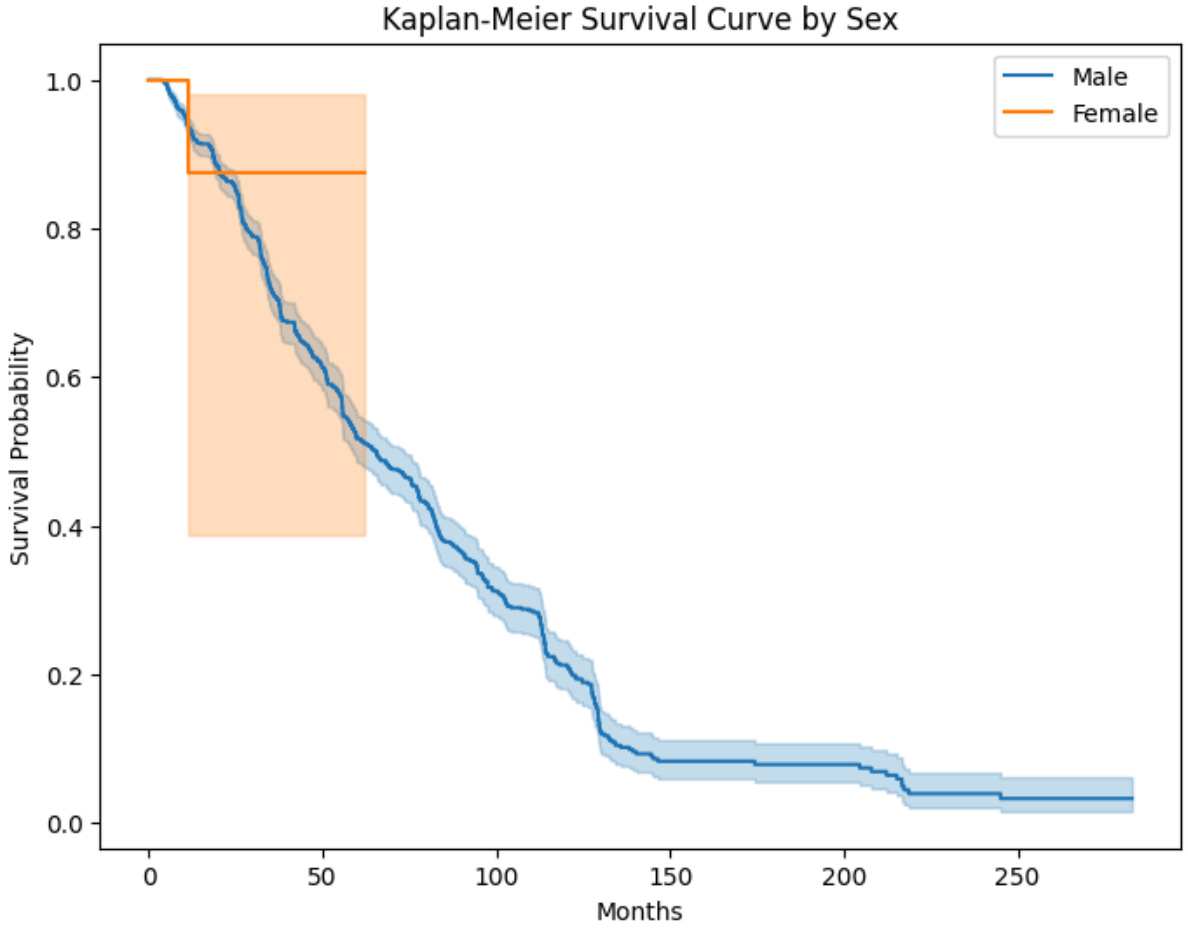

Fig 8: Kaplan-Meier Survival Curve by Sex. Compares survival trends between male and female breast cancer patients.

Cox Regression Hazard Ratios Before and After Balancing:

1. The hazard ratios shifted significantly after balancing, with race and age showing stronger associations with survival.
2. Before balancing, ethnicity showed an elevated risk but lacked statistical significance. After balancing, race had a clearer effect on survival.
3. The concordance index (a measure of model accuracy) improved after balancing, indicating that the model better captured survival patterns.

Our findings align with existing literature, further reinforcing the role of race, ethnicity, sex, and age in breast cancer survival disparities. According to the World Health Organization (WHO, 2022), there are significant global disparities in breast cancer incidence and mortality,

largely influenced by healthcare access, socioeconomic factors, and disease subtypes[5]. Despite similar incidence rates between Black and White women, Black women experience higher mortality rates, partly due to the increased likelihood of being diagnosed with aggressive subtypes such as triple-negative breast cancer (TNBC) and delayed treatment initiation, both of which contribute to worse survival outcomes[6]. This aligns with our Kaplan-Meier survival analysis and Cox regression findings, where race emerged as a significant predictor of survival outcomes.

Additionally, age remains a critical risk factor, with breast cancer incidence rising sharply after age 60 and nearly one-third of cases occurring in individuals aged 70 to 80 (NCI)[7]. Our study supports this, as age was consistently a significant predictor of survival in both unbalanced and balanced datasets, reinforcing the importance of early detection efforts for older populations. While Hispanic individuals tend to have lower incidence rates, breast cancer remains the leading cause of cancer-related deaths among Hispanic women[8], which may explain our observation of early survival declines in Hispanic patients.
These findings emphasize the need for targeted screening programs, equitable healthcare access, and timely interventions, particularly for high-risk racial and ethnic groups and older populations.

Impact of Early Detection and Screening Limitations

Breast cancer is primarily detected with mammograms, which are X-rays of the breast. Other methods used include MRIs, clinical exams, or simply being self-aware of lumps and other changes. However, mammograms are the only screening method that have been found to statistically reduce mortality. There are two types of mammograms, screening and diagnostic mammograms. Diagnostic mammograms are used to confirm the presence of tumors in breasts where lumps or other signs of cancer have been found. Screening mammograms are performed on healthy women at regular intervals as a preventative measure. The logic behind screening mammograms is to catch the cancer at an early stage where treatment will be more effective.

At first glance, this strategy seems to be sound. As the use of screening mammograms dramatically increased since the 1980s, so have breast cancer diagnoses. However, the connection between these increased incidence rates to reduced mortality rates from screening is less direct. While diagnoses of breast cancer in US women have increased by about 1% per year from 2012 to 2021, the death rate has dropped 44% cumulatively since its peak in 1989. This would suggest that early detection of breast cancer has played a large

---

[5] https://www.who.int/news-room/fact-sheets/detail/breast-cancer
[6] Surveillance Research Program, National Cancer Institute. SEER*Explorer. Breast Cancer – 5-year age-adjusted incidence rates, 2017-2021, by race/ethnicity, female, all ages, all stages. Accessed on April 22, 2024. https://seer.cancer.gov/explorer/, 2024.
[7] Surveillance Research Program, National Cancer Institute. SEER*Explorer. Breast Cancer – Breast cancer risk from birth over time, by sex, all races/ethnicities, risk of being diagnosed with cancer (2018-2019,2021). Accessed on July 9, 2024. https://seer.cancer.gov/explorer/, 2024.
[8] Surveillance Research Program, National Cancer Institute. SEER*Explorer. Breast Cancer – Breast cancer risk from birth over time, by sex, all races/ethnicities, risk of dying from cancer (2020-2022,2021). Accessed on July 9, 2024. https://seer.cancer.gov/explorer/, 2024.

role in reducing cancer-related deaths, but the causes of this decline remain unclear. A 2016 meta-analysis found that screening prevents 8 cancer-related deaths per 10,000 women from ages 50-59 and 21 deaths per 10,000 from 60-69. The total risk reduction for these age groups was between 25 and 31%. However, there are some caveats. Overall the effect magnitude is small especially compared to other cancers, and outside of those ages, the results were not statistically significant. Additionally, since screening has become popular the detection of early-stage breast cancer has significantly increased but late-stage cancer diagnosis rates have remained relatively constant. Since the survival rate is much lower for late stage cancer, the question is whether these additional early diagnoses prevented progression to late stage cancer and death. In addition to the increase in early-stage detection not leading to a corresponding decrease in late-stage detection, the studies showed that screening did not reduce all-cause deaths.

These factors indicate a phenomenon called overdiagnosis. Occasionally, a mammogram will identify a slow-growing or non-invasive tumor that would otherwise not affect the patient if left untreated. When this occurs, the patient may suffer more from treatment like surgery or radiation than from the disease itself. Unfortunately these cases are difficult to quantify since it is unknown how the disease would progress if it was not found, but clues like autopsy findings of benign tumors can give an idea of how often they occur. Current estimates put overdiagnosis rates between 10-30%.

Another drawback of mammograms are possible false positives. The overall sensitivity of mammograms is 87%, but can vary depending on age and breast density. False positive rates range from 7 to 12%, and the chance of having at least one false positive after 10 yearly mammograms is 50-60%. False positives are more common in younger women and dense breasts.

Although there is a general scientific consensus that screening mammograms do have a moderate positive effect, there is still some disagreement over implementation. Different organizations have different guidelines on what ages to begin and end screening as well as the frequency that they should be performed. The current US Preventive Services Task Force (USPSTF) recommendations are for screenings every two years between 50-74 with individual patient preferences and risks determining if there should be screening from ages 40-49. Overall it seems that mammograms are an important tool for fighting breast cancer, but more research is required to effectively quantify their benefits and drawbacks.

Conclusion

Our study highlights key factors influencing breast cancer risk, survival outcomes, and early detection strategies. The results provide strong evidence that race, age, and genetic predisposition are significant determinants of survival, reinforcing the need for personalized and equitable healthcare approaches. Our findings highlight the need for targeted interventions in breast cancer detection and treatment. Addressing racial and ethnic disparities, refining screening protocols, and incorporating genetic testing for high-risk individuals could significantly improve survival rates. The role of age in breast cancer prognosis suggests that more aggressive screening and prevention strategies should be

implemented for older patients. Finally, while early detection remains crucial, the limitations of current screening methods must be acknowledged, and efforts should be made to reduce overdiagnosis and improve risk-based screening approaches.

One of the most significant findings of our study is the disparity in breast cancer survival outcomes among racial and ethnic groups. Our survival analysis confirms that Black/African American patients have worse survival probabilities, highlighting the role of healthcare access, socioeconomic status, and biological factors in driving these disparities. To address this issue, policymakers should implement targeted screening programs and work toward improving access to early diagnosis and treatment for high-risk racial and ethnic groups. Ensuring equitable healthcare access is essential to reducing mortality rates and improving survival outcomes.

Another crucial factor in breast cancer risk is genetic predisposition. Given that 5-10% of breast cancer cases are hereditary, genetic screening for BRCA1 and BRCA2 mutations should be strongly recommended for individuals with a family history of breast cancer. Early identification of genetic susceptibility allows for preventative measures such as increased screening frequency, lifestyle modifications, and prophylactic treatments, all of which can significantly improve survival chances.

While mammograms have played a crucial role in reducing breast cancer mortality, our analysis raises concerns about overdiagnosis and false positives. Although early detection is essential, screening programs must carefully balance the benefits of early detection with the risks of unnecessary treatment, particularly for younger women with denser breast tissue who are more prone to false positives. Refining breast cancer screening guidelines based on age, breast density, and risk factors can help maximize the benefits while minimizing potential harm.

Age remains one of the strongest predictors of breast cancer survival, with older patients facing significantly worse outcomes. Our study reaffirms the importance of routine screening for women aged 50-74, but also highlights the need for individualized risk assessments when determining screening recommendations for younger women. Personalized screening strategies can help improve early detection rates while ensuring that resources are directed toward those who need them most.

Finally, future research should focus on understanding the interactions between demographic and genetic factors in breast cancer survival. Investigating the biological and healthcare-related causes of racial disparities can lead to the development of more targeted treatment approaches. By implementing risk-based screening and intervention strategies, healthcare systems can improve early detection, reduce disparities, and ultimately lower breast cancer mortality rates.

Appendix

Model	Key Metric	Significant Predictors	Best Fit?
Logistic Regression (Unbalanced)	Pseudo R^2 = **0.0221**, LL = -383.86	**Age (p = 0.005, OR = 1.22)**	Weak fit, only age significant
Logistic Regression (Balanced)	Pseudo R^2 = **0.0134**, LL = -1053.1	**Race (p < 0.01), Ethnicity (p = 0.002)**	Better fit, race now significant
Poisson Regression (Unbalanced)	Log-Likelihood = **-28007**, Pseudo R^2 = **0.0094**	**Age (p < 0.01)**	Only age significant, weak fit
Poisson Regression (Balanced)	Log-Likelihood = **-27741**, Pseudo R^2 = **0.0095**	**Sex, Race, Ethnicity, Age (all p < 0.01)**	Best fit for survival as count data
Cox Regression (Unbalanced)	Concordance = **0.66**, Log-Likelihood = **-761.88**	**Age (p < 0.005), Ethnicity (borderline)**	Only age was significant
Cox Regression (Balanced)	Concordance = **0.56**, Log-Likelihood = **-4761.56**	**Race (p < 0.005), Age (p = 0.01), Ethnicity (borderline)**	Best for time-to-event survival analysis

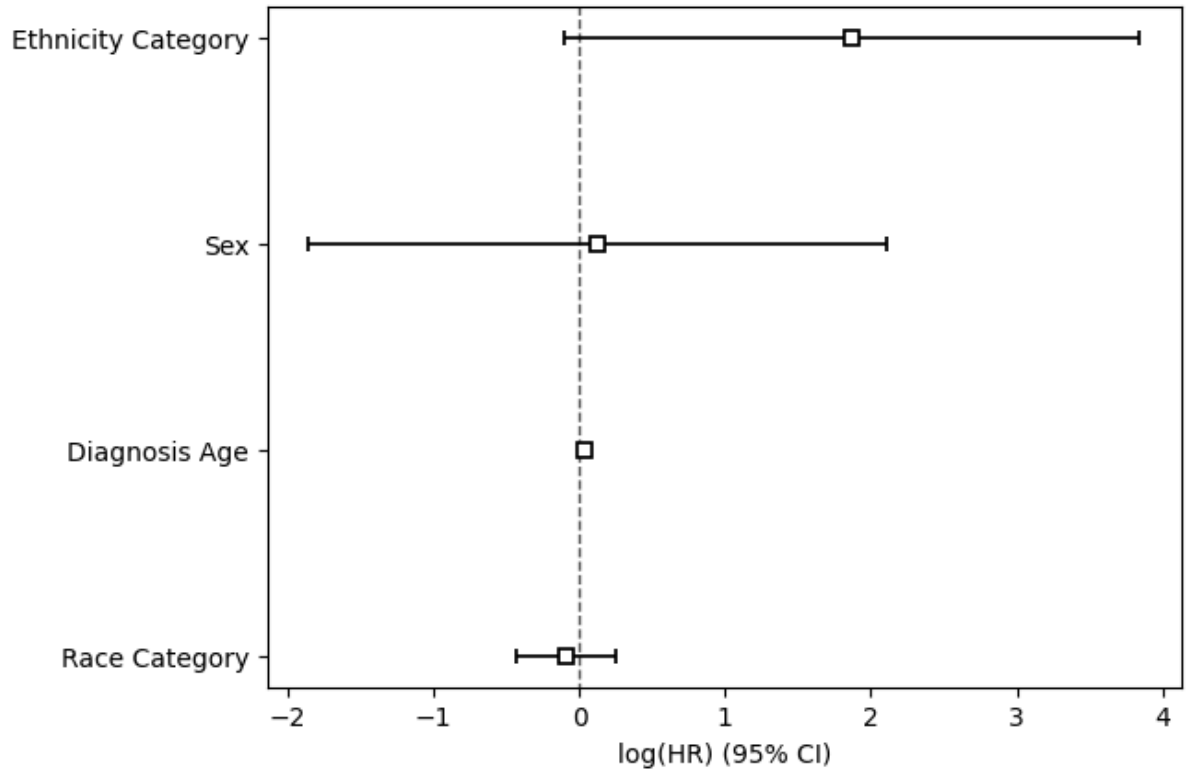

Fig 9: Cox Regression Hazard Ratios (for Unbalanced Data). Displays the impact of race, ethnicity, sex, and age on survival risk before balancing the dataset. Only age showed a significant association with survival, while race and ethnicity had wide confidence intervals.

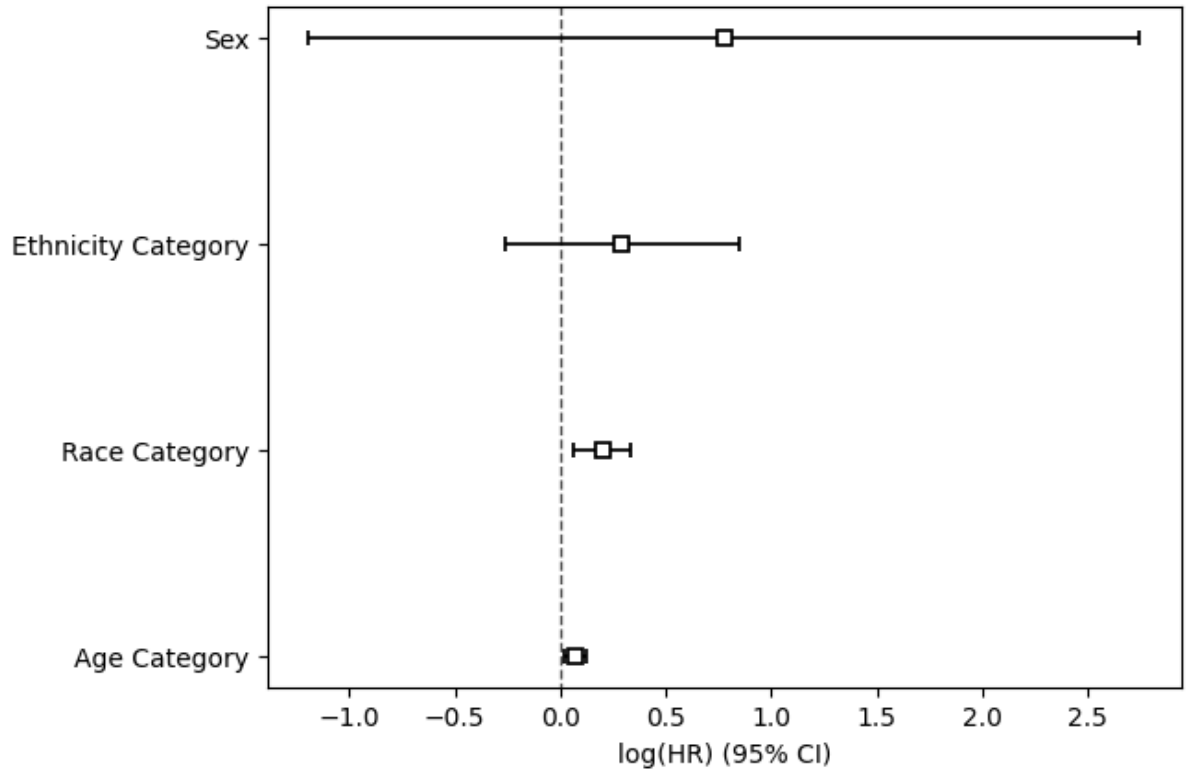

Fig 10: Cox Regression Hazard Ratios (for Balanced Data). Shows survival risk factors after applying SMOTE to balance survival outcomes. Race and age became stronger predictors of survival after balancing, with more reliable confidence intervals.