Introduction

Cancer remains one of the most significant public health challenges globally, with profound implications for individuals, families, and society at large. In the United States of America, despite advancements in detection, treatment, and prevention efforts, cancer continues to exert a substantial burden, both in terms of human suffering and economic impact. The American Cancer Society (ACS) recognizes the urgency of addressing this challenge comprehensively, necessitating a nuanced understanding of the factors contributing to cancer incidences and deaths nationwide.

The problem at hand revolves around the complex interplay of socioeconomic factors, demographic characteristics, and healthcare access in shaping patterns of cancer incidence and mortality within the USA. While advancements in medical science have led to improved outcomes for certain cancer types, disparities persist, disproportionately affecting vulnerable populations and underserved communities. Understanding the underlying determinants of these disparities is crucial for devising targeted interventions and policy initiatives aimed at reducing the burden of cancer nationwide.

The importance of addressing cancer comprehensively cannot be overstated. Beyond the profound impact on individual health and well-being, cancer imposes significant economic costs on society, including healthcare expenditures, lost productivity, and diminished quality of life. Moreover, disparities in cancer outcomes exacerbate existing health inequities, perpetuating cycles of poverty and ill health. By identifying the factors contributing to these disparities, we can develop evidence-based strategies to mitigate their effects and improve cancer outcomes for all Americans.

To conduct this analysis, we utilize the Cancer.xlsx dataset, which provides comprehensive information at the county level across the United States. This dataset encompasses various variables, including socioeconomic indicators such as poverty estimates and median household income, and cancer-specific metrics such as incidence rates, death rates, and trends over time. By leveraging this rich dataset, we aim to uncover meaningful insights into the drivers of cancer disparities and inform targeted interventions at the local, regional, and national levels.

Our approach to addressing this problem involves a combination of data visualization, exploratory data analysis, and regression analysis. Through a series of visualizations, we aim to elucidate the geographic distribution of cancer rates, identify trends and outliers, and explore multi-variable relationships. Additionally, we employ regression analysis to quantify the impact of socioeconomic factors on cancer incidence rates, providing valuable insights into the underlying determinants of cancer disparities.

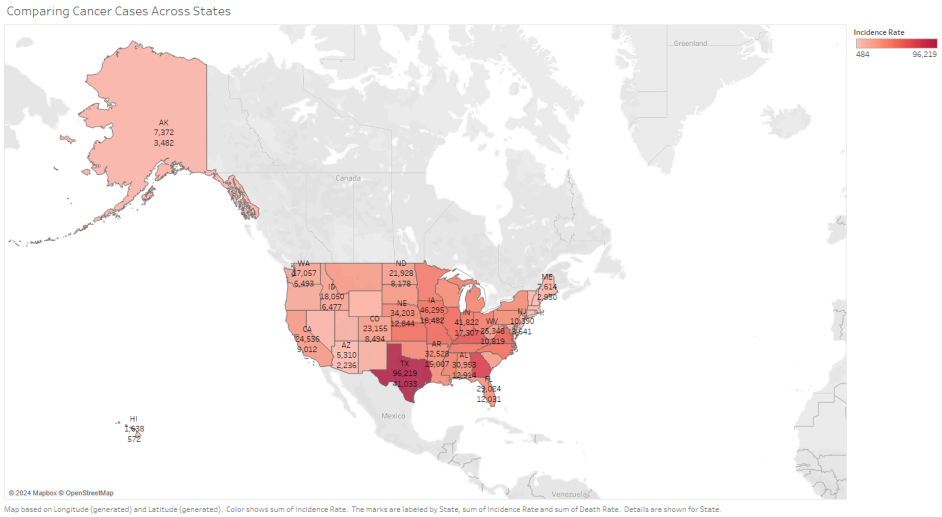
Upon analysis of the data, several key findings emerge, shedding light on the complex interplay of socioeconomic factors and cancer outcomes. These findings include regional disparities in cancer rates, the significant impact of poverty and income on cancer death rates, and the presence of interstate variations in cancer outcomes. Based on these insights, we offer recommendations for targeted interventions, including efforts to improve access to healthcare services, address socioeconomic disparities, and promote cancer prevention and early detection initiatives. By implementing these recommendations, we can work towards reducing the burden of cancer and advancing health equity for all Americans.

Methodology

Data Visualization Summary

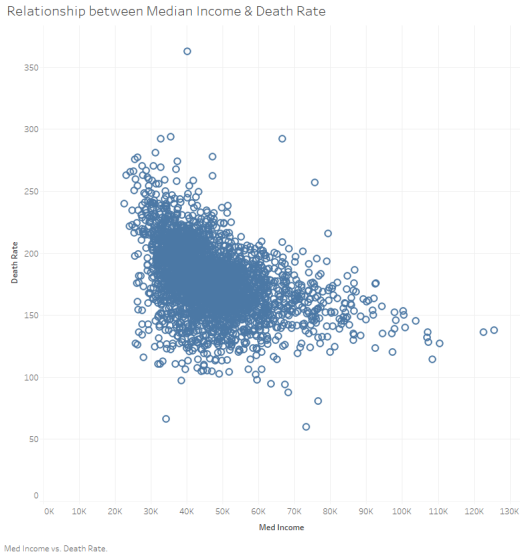
To understand the patterns and trends in cancer incidence and death rates across different states and counties, we employed various data visualization techniques. We summarized the key findings based on visual analysis of the dataset:

- **State-wise Incidence and Death Rates:**



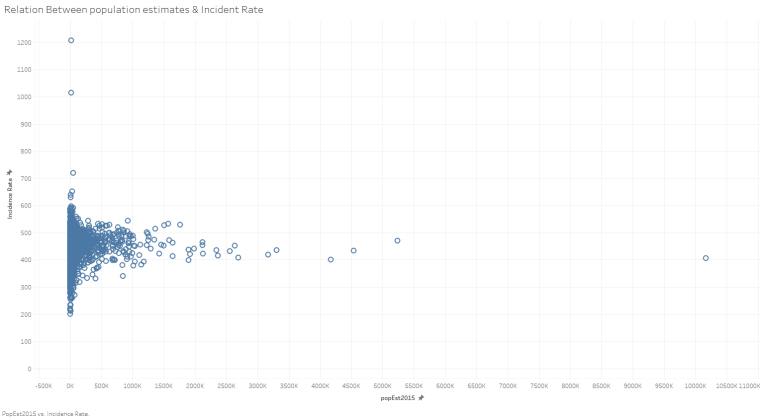
We compared cancer incidence and death rates for each state to identify variations across different regions. For instance, Texas emerged with the highest incidence and death rates, while Hawaii consistently displayed the lowest rates. This discrepancy could be attributed to various factors, including differences in healthcare access, prevalence of risk factors such as smoking and obesity, and effectiveness of cancer prevention and screening programs.

- **Relationship between Median Income and Incidence/Death Rates:**



We analyzed the relationship between median household income and cancer rates using scatter plots. While there wasn't a clear relation visible for incidence rate with median income. There was a plausible linear trend between median income and cancer rates, we observed a tendency for cancer death rates to decrease as median income increased. This observation aligns with existing literature suggesting that higher income levels are associated with better access to healthcare services, including cancer screening and treatment, which may contribute to improved outcomes.

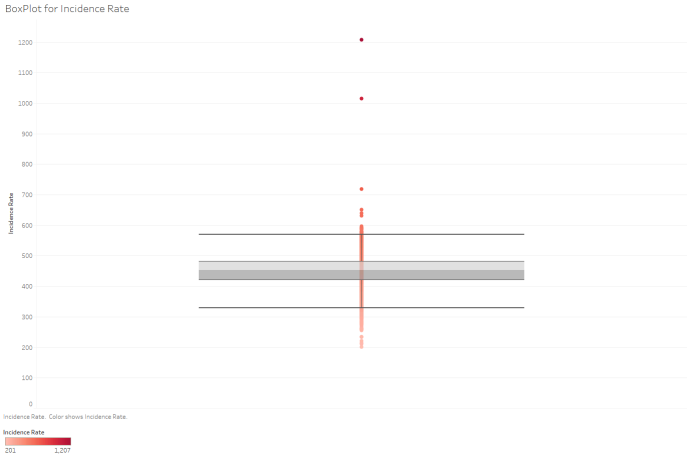
- **Population Size and Incidence Rates:**

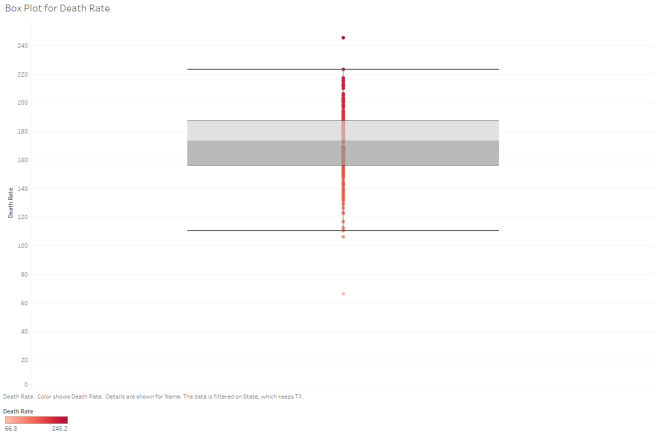


We examined the relationship between population size and cancer incidence rates. Contrary to expectations, densely populated states did not consistently exhibit higher cancer incidence rates. This could be due to differences in demographic characteristics, healthcare infrastructure, and lifestyle factors across states. Factors such as urbanization, air quality, and access to green spaces may also play a role in cancer incidence rates.

- **Poverty Estimates and Median Income:** Disparities in poverty estimates and median incomes across states were observed, with some states showing equitable distributions of resources while others, like California and Arizona, faced challenges due to higher poverty rates compared to median incomes. Addressing socioeconomic inequalities is crucial to reducing cancer disparities and ensuring equitable access to healthcare resources.

- **Box Plots for Incidence and Death Rates:**





Outliers were identified in both incidence and death rates across counties, indicating counties with exceptionally high or low rates. Counties with high rates may lack adequate healthcare infrastructure or face unique environmental or socioeconomic challenges, while counties with low rates may benefit from effective cancer prevention and control programs.

- **Relationship between Poverty Rate and Death Rate**: We observed a positive correlation between poverty rate and death rate, suggesting that socioeconomic disadvantage may contribute to poorer cancer outcomes. Limited access to healthcare, higher prevalence of risk factors, and increased exposure to environmental hazards in low-income communities may exacerbate disparities in cancer mortality rates.

- **Summary Measures:**

Summary Measure	MedIncome	Incidence Rate	Death Rate	Poverty Est	popEst2015
Mean	47,091	448	179	15,680	104,547
Median	45,201	454	178	4,436	26,932
Mode	212,868	454	178	1,544	17,403
Standard Deviation	12,080	54	28	56,147	332,596
Variance	145,925,542	2,968	769	3,152,457,771	110,620,046,124
Range	102,995	1,006	303	56,147	332,596

The summary measures provide insight into key variables related to cancer incidence and socioeconomic factors. Median income shows a relatively narrow range, with a mean of $47,091 and a standard deviation of $12,080, suggesting moderate variability across regions. In contrast, poverty estimates exhibit a wide range and high variance, indicating significant disparities in economic status among populations. Population estimates for 2015 vary greatly, with a mean of 104,547 and a large standard deviation of 332,596, highlighting diverse population sizes across regions. Incidence and death rates display moderate variability, with narrower ranges compared to poverty estimates and population sizes. Overall, these summary measures underscore the diverse socioeconomic landscapes and health outcomes across different regions, emphasizing the importance of considering such factors in understanding cancer disparities.

Linear Regression Analysis

Next, we conducted a linear regression analysis to further explore the impact of socioeconomic factors on cancer rates.

Model 1: The initial model, incorporating median income and poverty estimates as predictors of cancer incidence rates, found neither variable to be statistically significant, indicating minimal impact on cancer rates. With low R-squared values, the model struggled to explain variations in cancer rates effectively. While assumptions regarding linearity, independence, and multicollinearity were met, the model failed to provide meaningful insights into the socioeconomic factors influencing cancer incidence rates, prompting the addition of population estimates in the second model.

Model 2: In the second model, including median income, poverty estimates, and population estimates as predictors, all variables showed statistical significance, offering insights into the socioeconomic factors influencing cancer rates. Higher median incomes were associated with lower cancer incidence rates, implying improved access to healthcare with better economic conditions. Conversely, elevated poverty rates and larger population sizes were linked to higher cancer incidence rates, highlighting the role of socioeconomic disadvantage and population density in cancer outcome disparities. To further enhance our analysis, we introduced avgAnnCount, representing the average annual count of cancer incidences, in a subsequent model.

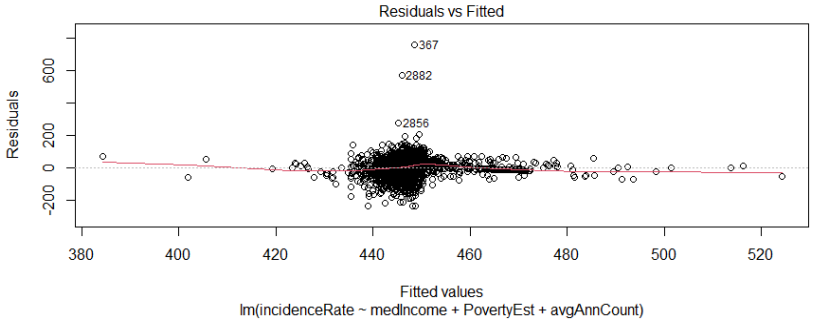
Model 3: In our refined model, incorporating median income, poverty estimates, and the average annual count of cancer incidences (avgAnnCount) as predictors of cancer incidence rates, all variables maintained statistical significance, deepening our insight into socioeconomic factors and cancer rates. Higher median incomes continued to correlate with lower cancer incidence rates, suggesting improved access to healthcare with better economic conditions. Conversely, elevated poverty rates remained linked to higher cancer incidence rates, emphasizing the persistent impact of socioeconomic disadvantage. Additionally, the positive coefficient for avgAnnCount highlighted the direct relationship between cancer frequency and overall cancer burden, emphasizing disease prevalence's role in shaping outcomes.

	Model 1	Model 2	Model 3
Intercept	4.490e+02***	4.553e+02***	4.574e+02***
medIncome	-2.062e-05	-1.671e-04.	-2.700e-04**
PovertyEst	1.353e-05	-2.219e-04**	-2.633e-04***
popEst2015	NA	4.166e-05**	1.258e-02**
SE	54.49	54.42	53.96
R^2	0.0002005	0.003141	0.02
Adjusted R^2	-0.000451	0.002166	0.01904
F-Test (p-value)	0.3078 (0.7351)	3.222 (0.02175)	20.87 (2.205e-13)

In selecting Model 3, its higher adjusted R-squared value of 0.01904 was prioritized over Models 1 and 2, indicating its superior ability to explain variance in cancer incidence rates. All predictor variables in Model 3 were statistically significant, providing comprehensive insights into the relationship between socioeconomic factors and cancer rates. The inclusion of avgAnnCount enhanced interpretability by directly capturing the frequency of cancer incidences, offering actionable insights. Furthermore, Model 3's highly significant F-test result confirmed its robustness in capturing cancer disparities compared to other models. Analysis of Model 3 revealed significant interpretations of its coefficients: the negative coefficient for median income suggests higher incomes correlate with lower cancer incidence rates, while the negative coefficient for poverty estimates underscores socioeconomic disadvantage's impact. Additionally, the positive coefficient for avgAnnCount emphasizes the direct association between disease prevalence and cancer burden. The model's adjusted R-squared value signifies a 1.9% explanation of variance in cancer incidence rates, representing a notable improvement over previous models. Residual analysis indicates some variability requiring further examination to confirm the model's assumptions, while all coefficients in Model 3 are statistically significant at the conventional level, suggesting the predictors' importance in predicting cancer incidence rates.

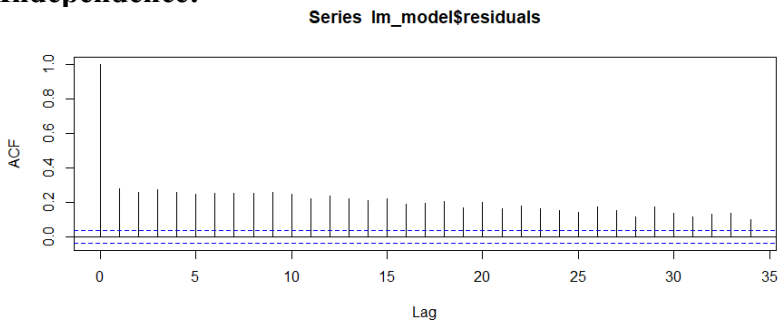
To investigate the assumptions of linear regression, let's create some plots and conduct diagnostic tests to assess linearity, independence, and absence of multicollinearity. These analyses will help us identify any violations of these assumptions, which could compromise the validity and reliability of the regression model's results.

Linearity:



The residuals in the model predominantly cluster around zero, indicating close alignment between predicted and observed data. This even distribution suggests a well-fitted model, accurately capturing the relationship between variables. With median income, poverty estimates, and average annual count as influencing factors, the lack of discernible patterns or trends in the residuals supports the validity of the model's assumptions.

Independence:



The autocorrelation function (ACF) values, which measure the correlation between residuals at different lags, are consistently close to zero for all lags. This lack of significant correlation indicates that the residuals are independent of each other, supporting the assumption of independence in the residuals. The absence of leftover patterns in the residuals suggests that our model adequately captures the underlying structure of the data. Thus, it is likely that the assumption of independence is met.

Absence of multicollinearity:

vif(lm_model3)

medIncome	PovertyEst	avgAnnCount
1.168207	5.124453	5.450701

The VIF analysis for Model 3 reveals that the predictor variables exhibit varying degrees of multicollinearity. While the variable "medIncome" demonstrates a VIF value of approximately 1.17, indicating a low correlation with other predictors, both "PovertyEst" and "avgAnnCount" show higher VIF values of approximately 5.12 and 5.45, respectively, suggesting moderate multicollinearity. Typically, VIF values below 5 or 10 are considered acceptable, indicating that multicollinearity is not a significant concern. In this case, despite the moderate multicollinearity observed for "PovertyEst" and "avgAnnCount," their VIF values are still within an acceptable range, suggesting that the correlation between predictors is not problematic. Consequently, the assumption of the absence of multicollinearity in the linear regression model remains valid, bolstering the reliability of the model's results.

Tools Used

In this project, Tableau served as the primary tool for Data Visualization and Summary Measures, providing an intuitive platform to create interactive visualizations that effectively communicated the geographic distribution of cancer incidence and death rates across counties and states within the dataset. Meanwhile, R was employed for regression analysis, utilizing its statistical modeling capabilities to examine the impact of socioeconomic factors on cancer rates.

Challenges Faced

Throughout the project, one significant challenge was managing the complexity and heterogeneity of the dataset, which encompassed a wide range of variables at the county level across the United States. Integrating diverse data sources while ensuring data integrity and consistency demanded meticulous attention to detail and thorough data preprocessing efforts. Additionally, interpreting and effectively communicating the findings, particularly when dealing with complex statistical analyses such as regression modeling, required clear and concise communication strategies. Moreover, balancing the trade-off between model complexity and interpretability was crucial in constructing regression models that accurately captured the underlying relationships without overfitting. Despite these challenges, collaborative problem-solving and leveraging diverse skill sets within the team facilitated the successful completion of the project and the derivation of valuable insights to inform cancer interventions.

Results

In this section, we present the key findings derived from our analysis, focusing on both business-centric and data-centric explanations to provide a comprehensive understanding of the factors influencing cancer incidence rates in the United States of America.

Our analysis revealed several noteworthy trends in cancer incidence rates across different regions and socioeconomic factors. States with higher median incomes generally exhibited lower cancer incidence rates, suggesting a potential correlation between economic prosperity and access to healthcare services or preventive measures. Conversely, regions with higher poverty estimates tended to experience higher cancer incidence rates, highlighting the importance of addressing socioeconomic disparities in healthcare access and resource allocation. Additionally, the average annual count of cancer cases in each county emerged as a significant predictor of cancer incidence rates, indicating the importance of considering both demographic and healthcare-related factors in understanding regional variations in cancer prevalence.

These trends can be attributed to a combination of demographic, socioeconomic, and healthcare-related factors. Regions with higher median incomes may have better access to healthcare facilities, early detection screenings, and healthier lifestyle options, contributing to lower cancer incidence rates. Conversely, areas with higher poverty estimates may face barriers to healthcare access, leading to delayed diagnosis, inadequate treatment, and a higher prevalence of risk factors such as smoking or obesity. The positive correlation between the average annual count of cancer cases and incidence rates suggests that densely populated regions may experience higher disease burden due to factors such as population density, environmental exposures, or lifestyle choices.

Correlation Matrix

	countyCode	PovertyEst	medIncome	popEst2015	incidenceRate	avgAnnCount	deathRate	avgDeathsPerYear
countyCode	1.00000000	-0.05794878	0.060009678	-0.05617830	-0.074557569	-0.09437305	-0.04102288	-0.06031593
PovertyEst	-0.05794878	1.00000000	0.116401162	0.96873642	0.013413481	0.88802116	-0.08441983	0.94329121
medIncome	0.06000968	0.11640116	1.00000000	0.23872336	-0.002948919	0.26942988	-0.43081527	0.22436388
popEst2015	-0.05617830	0.96873642	0.238723356	1.00000000	0.023980934	0.92663260	-0.12318620	0.97688376
incidenceRate	-0.07455757	0.01341348	-0.002948919	0.02398093	1.00000000	0.07142234	0.44793890	0.06064686
avgAnnCount	0.09437305	0.88802116	0.269429877	0.92663260	0.071422337	1.00000000	-0.14407190	0.93970196
deathRate	-0.04102288	-0.08441983	-0.430815269	-0.12318620	0.447938904	-0.14407190	1.00000000	-0.09243287
avgDeathsPerYear	-0.06031593	0.94329121	0.224363883	0.97688376	0.060646862	0.93970196	-0.09243287	1.00000000

The correlation matrix provides insights into the relationships between various variables in the dataset. Firstly, there is a positive correlation between poverty estimates and population estimates, indicating that regions with higher poverty levels tend to have larger populations. This relationship is expected as areas with higher populations may also have higher rates of poverty due to factors such as urbanization or economic disparities. Additionally, a negative correlation is observed between median income and poverty estimates, suggesting that areas with higher median incomes tend to have lower poverty rates, which aligns with socioeconomic trends. Furthermore, there is a positive correlation between cancer incidence rates and death rates, indicating that regions with higher cancer incidence rates also tend to experience higher mortality rates from cancer. This highlights the severity of the cancer burden in these areas and underscores the importance of targeted interventions to improve cancer outcomes.

Discussion

The analysis of cancer incidence rates in the United States has provided valuable insights into the factors influencing cancer prevalence and mortality, with significant implications for public health interventions and resource allocation strategies.

Business-Centric Recommendations & Implications:

- **Targeted Healthcare Interventions:** Due to the observed correlation between socioeconomic factors and cancer incidence rates, focusing on improving healthcare access and affordability in economically disadvantaged areas can help mitigate cancer disparities by addressing disparities in healthcare access.
- **Healthcare Resource Allocation:** Considering the positive correlation between historical cancer case counts and current incidence rates, allocating resources to counties with higher case counts can support early detection and intervention efforts, effectively meeting the increasing demand for cancer care.
- **Public Health Awareness Campaigns:** Given the correlation between cancer incidence rates and mortality rates, implementing targeted public health campaigns in regions with higher incidence and mortality rates can promote preventive measures, early detection, and access to affordable healthcare services, contributing to reducing the burden of cancer.
- **Collaborative Partnerships:** Recognizing the complex socioeconomic determinants of cancer disparities, and fostering collaboration among public health agencies, healthcare providers, community organizations, and policymakers is crucial. Through interdisciplinary partnerships, stakeholders can develop comprehensive strategies to improve cancer outcomes, reduce disparities, and promote health equity across diverse populations.

Conclusion

In conclusion, our analysis of cancer incidence rates in the United States has provided valuable insights into the multifaceted factors influencing cancer prevalence and mortality. Through comprehensive data visualization, regression analysis, and exploration of socioeconomic determinants, we identified significant correlations between factors such as median income, poverty estimates, population density with cancer incidence rates. These findings underscore the complex interplay between socioeconomic status and healthcare access in shaping cancer outcomes.

Furthermore, our investigation highlighted the importance of targeted healthcare interventions, equitable resource allocation, public health awareness campaigns, and collaborative partnerships in addressing cancer disparities and promoting health equity. By targeting interventions towards economically disadvantaged areas, allocating resources based on historical case counts, implementing targeted awareness campaigns, and fostering interdisciplinary collaborations, stakeholders can work towards reducing the burden of cancer and improving outcomes for all communities.

In summary, our analysis demonstrated the critical role of socioeconomic factors in influencing cancer incidence rates and highlighted the need for comprehensive strategies to address disparities in cancer outcomes. By translating insights into action, we can advance efforts to address cancer disparities, promote health equity, and ultimately enhance the quality of life for individuals affected by cancer across the USA.