PM-566 Final Project

Carissa Feliciano

Racial/Ethnic Disparities in Survival among Young Patients with Non-Small Cell Lung Cancer (2011-2021): A SEER Analysis

Introduction

The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) collects cancer incidence and survival data from population-based cancer registries across the US. The SEER database includes information on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, first course of treatment, and vital statistics. The SEER Research Plus and NCCR Database includes data from 1975-2021.

Lung cancer is the leading cause of cancer deaths for both men and women in the US, with non-small cell lung cancer (NSCLC) accounting for 80-85% of cases (1). Lung cancer in young adults, defined here as age 50 years, is relatively rare. From 2014-2018, 7.2% of incident lung cancer cases were in adults less than 55 years of age (2). Studies have suggested that young patients with non-small cell lung cancer have different clinical and pathologic characteristics compared to older patients, such as a higher proportion of adenocarcinoma, higher prevalence of targetable driver mutations, and better prognosis (3,4). While several studies have investigated racial/ethnic disparities among all lung cancer patients, there is limited data characterizing racial/ethnic disparities among young NSCLC patients (aged 50 years) (5).

The primary questions of interest: Is survival time associated with race/ethnicity among young adults (age 18-50) with non-small cell lung cancer (NSCLC)? If so, can differences in the frequency distributions of sex, histologic subtype, and stage at diagnosis explain any differences in survival across the racial/ethnic groups?

Methods

Data Acquisition and Wrangling

The data was extracted from the SEER-17 database, which covers approximately 26.5% of the US population (6). The SEER Stat program was used to access the SEER-17 dataset. The SEER Stat program was used to filter for cases that met the following criteria: incident diagnosis of non-small cell lung cancer between January 1, 2011 and December 31, 2021, aged 18 to 50 years at the time of diagnosis, first primary malignancy, and residence in California.

The SEER 17 dataset includes cancer cases diagnosed between 2000 and 2021. Our analysis was restricted to cases diagnosed between 2011 and 2021 to ensure the evidence was relatively recent and relevant. Cases of non-small cell lung cancer (NSCLC) were identified using primary site codes and histology ICD-O-3 codes, as described by Hansen et al. and Ganti et al., respectively (7, 8). The primary site codes were C34.0 (Main bronchus), C34.1 (Upper lobe, lung), C34.2 (Middle lobe, lung), C34.3 (Lower lobe, lung), C34.8 (Overlapping lesion of lung), and C34.9 (Lung, NOS). The histology ICD-O-3 codes are listed below.

Histology ICD-O-3 codes included by category:\ - Squamous cell carcinoma: 8051–8052, 8070-8076, 8078, 8083-8084, 8090, 8094, 8123 \ - Adenocarcinoma: 8015, 8050, 8140-8141, 8143-8145, 8147, 8190, 8201, 8211, 8250-8255, 8260, 8290, 8310, 8320, 8323, 8333, 8401, 8440, 8470-8471, 8480-8481, 8490, 8503, 8507, 8550, 8570-8572, 8574, 8576 \ - Large cell carcinoma: 8012–8014, 8021, 8034, 8082 \ - Not otherwise specified: 8046, 8003–8004, 8022, 8030, 8031-8033, 8035, 8120, 8200, 8240–8241, 8243–8246, 8249, 8430, 8525, 8560, 8562, 8575

Using the SEER Stat program, 40 variables were extracted, including demographics, staging, treatment, and survival. The resulting dataset was exported from the SEER Stat program as a csv file and then uploaded into R. Once in R, seven variables relevant to the primary question were selected: "Age recode with single ages and 90+", "Year of diagnosis", Sex", "Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)", "Histologic Type ICD-O-3", "Combined Summary Stage (2004+)", and "Survival months".

The key variables were renamed to streamline coding. The survival months, age, and ICD-O-3 variables were converted from characters to integers. Prior to converting the age variable, the text "years" was removed from each value. The "race and origin" variable was converted from characters to factors. A new variable called "histology" was created to group the histologies based on ICD-O-3 codes into the following four categories: Squamous cell carcinoma, Adenocarcinoma, Large cell carcinoma, and Not otherwise specified. The ICD-O-3 codes were grouped based on the categories cited by Ganti et al (8). To ensure the variable was correctly coded, a summary table was created that contained the total number of observations per ICD code grouped by histology.

Exploratory Data Analysis

The dim() function was used to determine the dimensions of the dataset. This dataset includes 4,427 cases of NSCLC. The dataset has 7 variables. The head() and tail() functions were used to view the top and bottom of the dataset. Based on the top and bottom of the dataset, there appears to be no irregularities. The str() function was used to view the variables and identify any abnormal variables.

The key variables of interest (race/ethnicity, survival months, sex, histology, and stage at diagnosis) were closely examined. The table() function was used to check the categorical variables. To check numerical variables, the summary() function was used and a histogram was plotted. The proportions of missing values were calculated.

To clean the data, observations with missing survival data and observations with race/ethnicity listed as "Non-Hispanic Unknown Race" were excluded.

Data Exploration

The R software packages gtsummary and kable were used to generate a descriptive table showing distributions of the key variables (sex, cancer stage at diagnosis, histology) stratified by race/ethnicity. The median, minimum, maximum, and interquartile range of survival time were calculated for each racial/ethnic group and summarized in a table using kable. The dplyr and boot packages were used to calculate the median survival times and 95% confidence intervals, stratifying by key variables. The median survival times were stratified by race/ethnicity and sex, race/ethnicity and stage at diagnosis, and race/ethnicity and histological diagnosis. The dplyr package was also used to calculate the frequency distributions of sex, stage at diagnosis, and histological diagnosis within each racial/ethnic group. The ggplot2 and RColorBrewer packages were used to generate the boxplots and barcharts. The plotly package was used to create the interactive visualizations on the website.

Results

The analyses included 4,384 persons aged 18 to 50 years who were diagnosed with first-primary, non-small cell lung cancer and resided in California. The characteristics of each racial/ethnic group are described in Table 1. Among the patients, 1719 identified as non-Hispanic White (NHW), 438 as non-Hispanic Black (NHB), 1060 as Hispanic, 1145 as non-Hispanic Asian or Pacific Islander (NHAPI), and 22 as non-Hispanic American Indian/Alaska Native (NHAIAN). The median age of the groups ranged from 45 to 47 years. Across all the groups, the most common cancer stage at diagnosis was distant, and the most common histological diagnosis was adenocarcinoma.

Table 1: Patient characteristics by race/ethnicity

Characteristic	**Non-Hispanic White** N = 1,719[1]	**Non-Hispanic Black** N = 438[1]	**Hispanic (All Races)** N = 1,060[1]	**Non-Hispanic Asian or Pacific Islander** N = 1,145[1]	**Non-Hispanic American Indian/Alaska Native** N = 22[1]
Age (Years)	46 (42, 49)	47 (42, 49)	45 (38, 48)	46 (41, 49)	47 (42, 49)
Sex					
Female	866 (50%)	228 (52%)	608 (57%)	624 (54%)	13 (59%)
Male	853 (50%)	210 (48%)	452 (43%)	521 (46%)	9 (41%)
Cancer Stage at Diagnosis					
Localized	350 (20%)	64 (15%)	202 (19%)	127 (11%)	2 (9.1%)
Regional	316 (18%)	79 (18%)	143 (13%)	129 (11%)	6 (27%)
Distant	1,036 (60%)	289 (66%)	693 (65%)	868 (76%)	14 (64%)
Unknown/unstaged	17 (1.0%)	6 (1.4%)	22 (2.1%)	21 (1.8%)	0 (0%)
Histology					
Adenocarcinoma	1,062 (62%)	282 (64%)	716 (68%)	956 (83%)	16 (73%)
Squamous Cell Carcinoma	196 (11%)	43 (9.8%)	69 (6.5%)	64 (5.6%)	4 (18%)
Large Cell Carcinoma	27 (1.6%)	13 (3.0%)	10 (0.9%)	11 (1.0%)	0 (0%)
Not Otherwise Specified	434 (25%)	100 (23%)	265 (25%)	114 (10.0%)	2 (9.1%)

[1] Median (Q1, Q3); n (%)

Table 2: Survival time by race/ethnicity

Race/Ethnicity	Number of Observations	Median	Min	Max	1st Quartile	3rd Quartile
Non-Hispanic White	1719	19.0	0	131	6	50.00
Non-Hispanic Black	438	11.0	0	130	4	29.00
Hispanic (All Races)	1060	18.0	0	131	6	41.00
Non-Hispanic Asian or Pacific Islander	1145	19.0	0	130	7	43.00
Non-Hispanic American Indian/Alaska Native	22	21.5	0	82	5	35.75

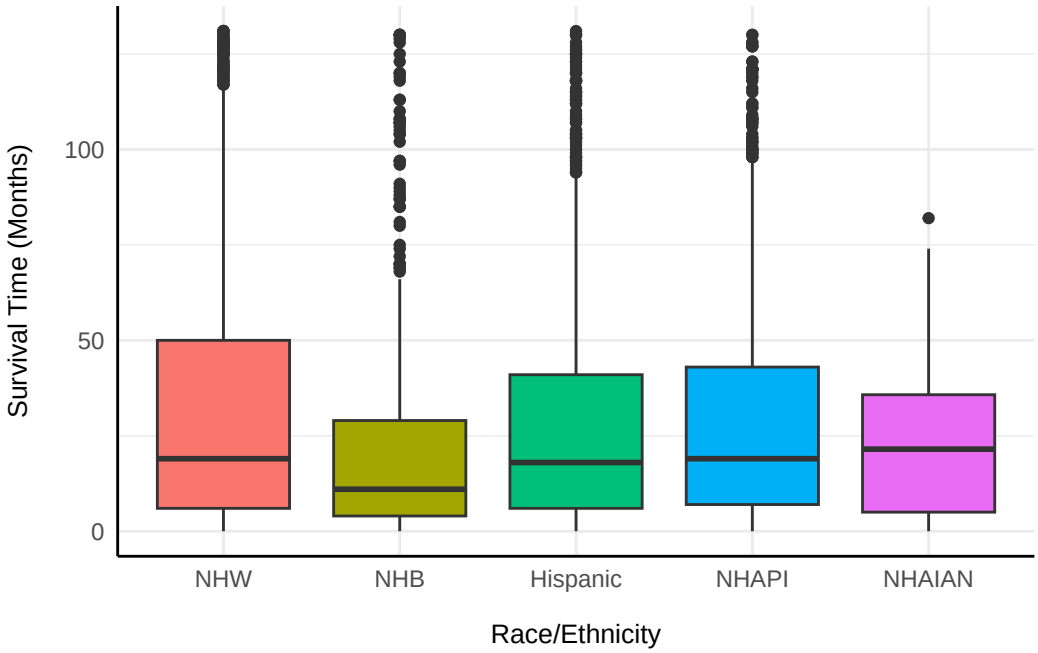

Figure 1: Distribution of survival times by race/ethnicity. The middle line within each box represents the median survival time. The upper bound of the box represents the 75th percentile (Q3), and the lower bound of the box represents the 25th percentile (Q1). The upper whisker represents the maximum (highest value within 1.5*IQR of Q3), and the lower whisker represents the minimum (lowest value within 1.5*IQR of Q1). Points outside the whiskers represent outliers. NHW, non-Hispanic White; NHB, non-Hispanic Black; Hispanic, Hispanic all races; NHAPI, non-Hispanic Asian or Pacific Islander; NHAIAN, non-Hispanic American Indian/Alaska Native.

Figure 1 shows the distribution of survival times by race/ethnicity. The median survival time was lowest for the NHB group (11.0 months) and highest for the NHAIAN group (21.5 months) (Table 2). The median survival time was similar for the NHW (19 months), Hispanic (18 months), and NHAPI (19 months) groups. The small sample size of the NHAIAN group (n = 22) limits the ability to draw definitive conclusions about this group.

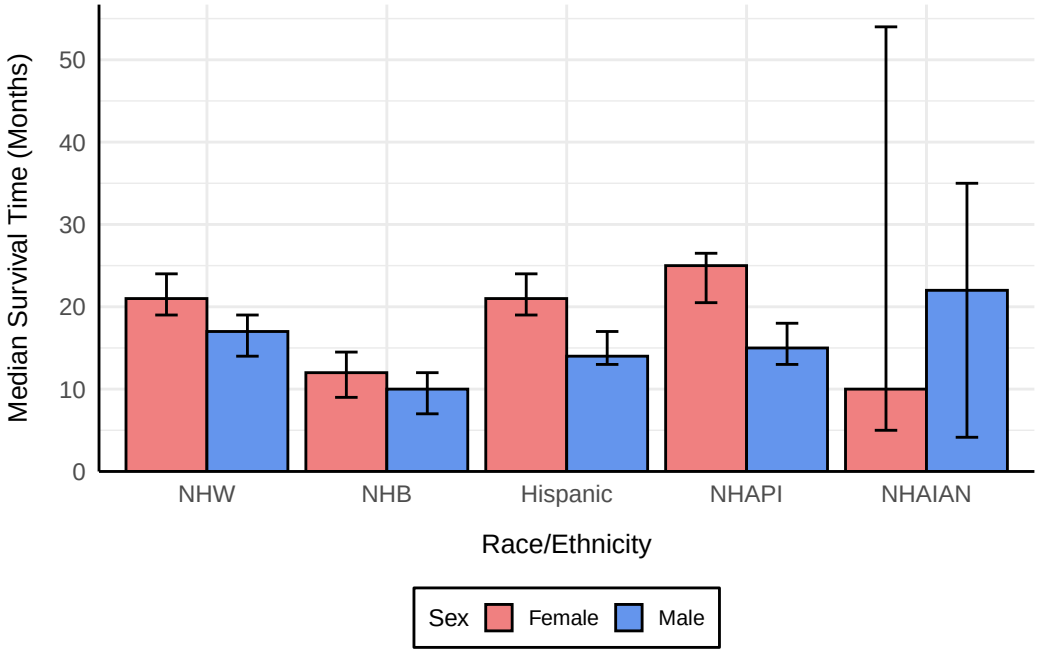

Figure 2: Median survival time by race/ethnicity and sex. Error bars represent 95% confidence intervals (CIs). NHW, non-Hispanic White; NHB, non-Hispanic Black; Hispanic, Hispanic all races; NHAPI, non-Hispanic Asian or Pacific Islander; NHAIAN, non-Hispanic American Indian/Alaska Native.

Among NHB patients, 52% were female, compared to 50% of NHW, 57% of Hispanic, 54% of NHAPI, and 59% of NHAIAN patients (Table 1). Figure 2 shows the median survival time stratified by sex and race/ethnicity. The median survival time of NHB females (12 months) was significantly lower than that of NHW (21 months), Hispanic (21 months), and NHAPI (25 months) females. Similarly, the median survival time of NHB males (10 months) was significantly lower than that of NHW (17 months), Hispanic (14 months), and NHAPI (15 months) males. Racial/ethnic disparities in survival time persisted after stratifying by sex. This suggests that differences in the distribution of sex across the racial/ethnic groups do not explain the differences in survival time.

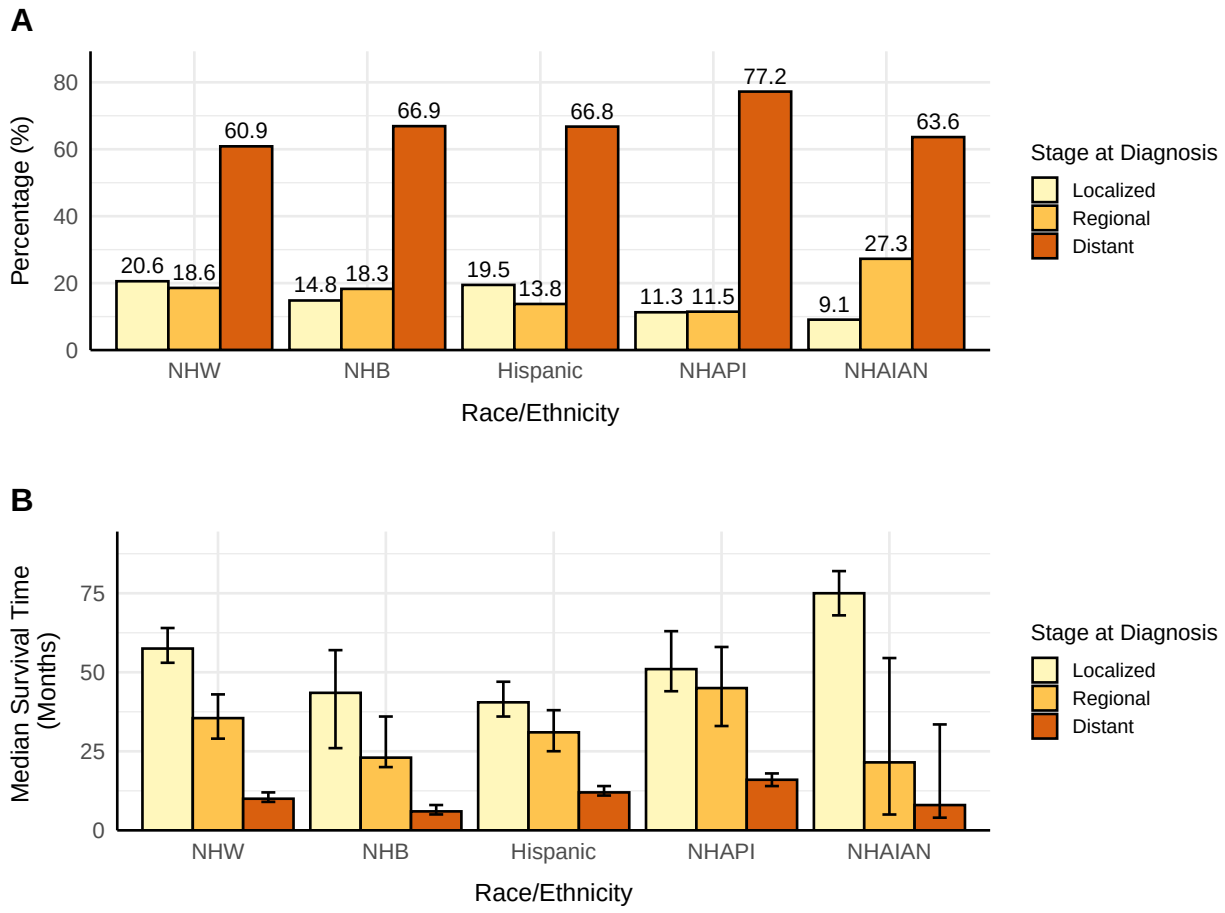

Figure 3: Racial/ethnic differences in cancer stage at diagnosis and median survival time. A) Distribution of cancer stage at diagnosis within each racial/ethnic group. B) Median survival time by race/ethnicity and stage at diagnosis. Error bars represent 95% CIs. NHW, non-Hispanic White; NHB, non-Hispanic Black; Hispanic, Hispanic all races; NHAPI, non-Hispanic Asian or Pacific Islander; NHAIAN, non-Hispanic American Indian/Alaska Native.

Figure 3a shows the distribution of stage at diagnosis for each racial/ethnic group. The NHB group had a lower percentage of patients diagnosed with localized cancer (14.6%) compared to the NHW (20.6%) and Hispanic (19.5%) groups. This may partially explain the lower median survival time of NHB patients, as patients with localized cancer have better survival outcomes overall than patients diagnosed with regional or distant cancer. Of note, the NHAPI group had a lower percentage of patients diagnosed with localized cancer (11%) compared to the NHB group (15%) despite having a higher median survival time.

To explore whether stage at diagnosis contributed to the racial/ethnic disparities in survival time, we assessed the median survival time by race/ethnicity and stage at diagnosis (Fig. 3b). Among patients with localized cancer at diagnosis, NHB patients had a lower median survival time (43.5 months) than NHW (57.5 months), NHAPI (51 months), and NHAIAN (75 months) patients. However, these differences were not statistically significant as the 95% CIs overlapped. The median survival time for Hispanic patients (40.5 months) was similar to that of NHB patients (43.5 months).

Among patients with regional cancer at diagnosis, NHB patients had a lower median survival time (23 months) than NHW (35.5 months), Hispanic (31 months), and NHAPI (45 months) patients. These differences were not statistically significant. Among patients with distant cancer at diagnosis, the median survival time of NHB patients (6 months) was significantly lower than that of NHW (10 months), Hispanic (12 months), and NHAPI (16 months) patients.

The lower median survival time of NHB patients compared other racial/ethnic groups may be partially explained by a higher frequency of late stage diagnoses. However, stage at diagnosis does not fully explain the racial/ethnic disparities in survival, as these disparities persist when stratifying by stage at diagnosis.

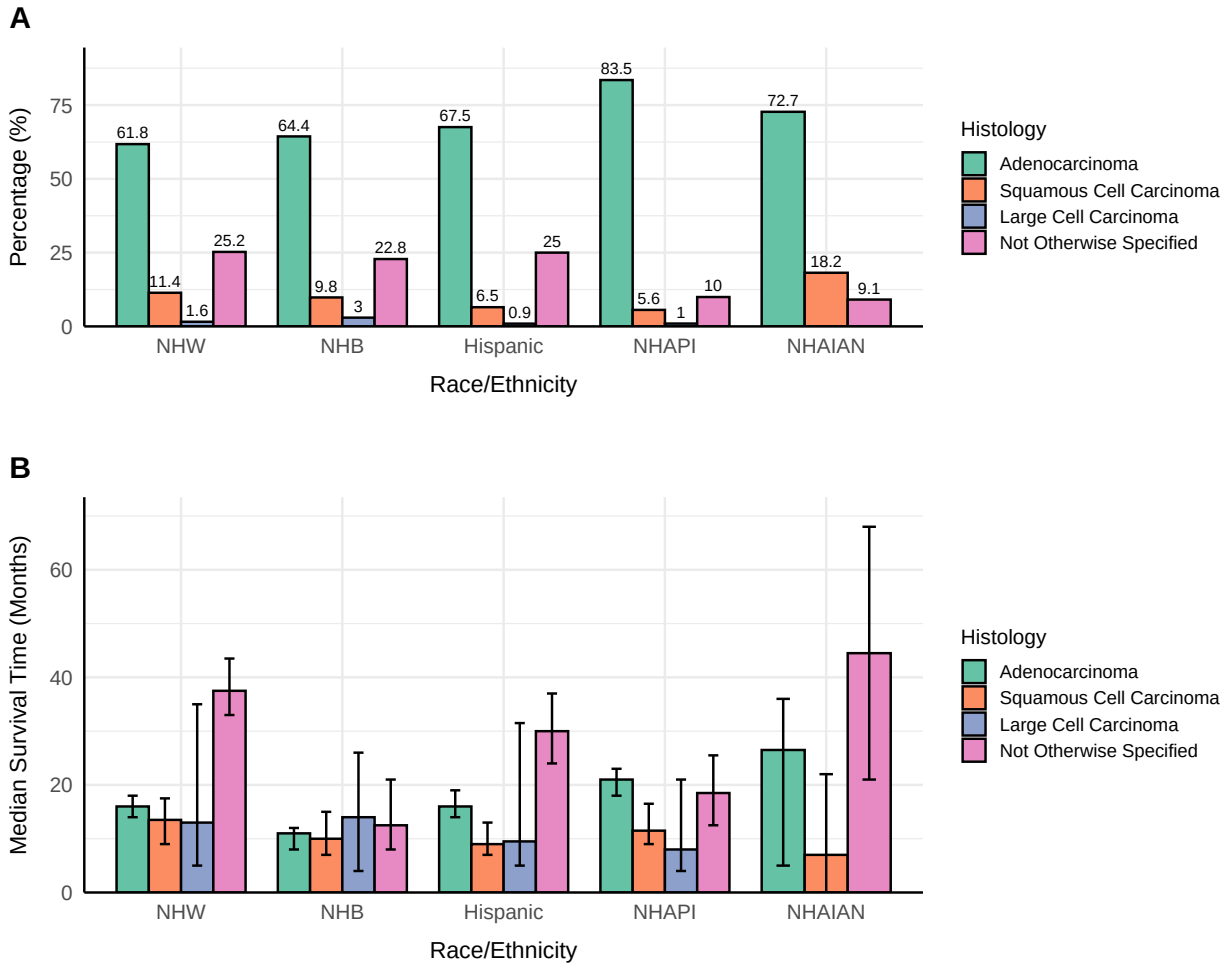

Figure 4: Racial/ethnic differences in histological diagnoses and median survival time. A) Distribution of histological diagnoses within each racial/ethnic group. B) Median survival time by race/ethnicity and histological diagnosis. Error bars represent 95% CIs. NHW, non-Hispanic White; NHB, non-Hispanic Black; Hispanic, Hispanic all races; NHAPI, non-Hispanic Asian or Pacific Islander; NHAIAN, non-Hispanic American Indian/Alaska Native.

Figure 4a shows the distribution of histological diagnoses for each racial/ethnic group. For the NHW, NHB, Hispanic, and NHAPI groups, adenocarcinoma is the most common histological di-

agnosis, followed by NOS. The NHB group had a similar percentage of patients diagnosed with adenocarcinoma (64.4%) as the NHW (61.8%) and Hispanic (67.5%) groups. The NHB group also had a similar percentage of patients diagnosed with NOS (22.8%) as the NHW (25.2%) and Hispanic (25%) groups.

To explore whether differences in the frequency distributions of histological diagnoses contributed to the racial/ethnic disparities in survival time, we assessed the median survival time by race/ethnicity and histological diagnosis (Fig. 4b). Among patients diagnosed with adenocarcinoma, NHB patients had a statistically significantly lower median survival (11 months) than NHW (16 months), Hispanic (16 months), and NHAPI (21 months) patients. Among patients diagnosed with squamous cell carcinoma, NHB patients had a lower median survival (10 months) than NHW (13.5 months) and NHAPI (11.5 months). However, this difference was not statistically significant. Among patients with a histological diagnosis of NOS, NHB patients (12.5 months) had a statistically significantly lower median survival than NHW (37.5 months) and Hispanic (30 months). NHB patients had a lower median survival (12.5 months) than NHAPI patients (18.5 months) as well, but this difference appeared to be not statistically significant.

Less than 3% of patients from each group were diagnosed with large cell carcinoma. The small sample size of this subgroup limits the ability to draw definitive conclusions, as reflected by the wide 95% CIs.

Conclusion

This project found that among patients aged 18 to 50 years with non-small cell lung cancer in California, NHB patients had a lower survival time compared to NHW, Hispanic, and NHAPI patients. The median survival time for NHB patients was 11.0 months, compared to 19 months for NHW, 18 months for Hispanic, 19 months for NHAPI, and 21.5 months for NHAIAN patients. Additionally, the NHB group had a lower percentage of patients with localized cancer (14.8%) compared to 20.6% for the NHW group and 19.5% for the Hispanic group. While stage at diagnosis likely contributed to the racial/ethnic disparity in survival, it could not fully explain the disparity. It appears that sex and histological diagnosis did not contribute much to the differences in median survival across the racial/ethnic groups as these disparities generally persisted when stratifying for sex and histological diagnosis.

References

1. American Cancer Society. Lung Cancer Statistics. American Cancer Society. Updated January 29, 2024. Accessed October 27, 2024. https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html

2. Howlader N, Noone AM, Krapcho M, et al. SEER Cancer Statistics Review 1975-2018. Published April 15, 2021. Accessed October 27, 2024. https://seer.cancer.gov/archive/csr/1975_2018/results_merged/sect_01_overview.pdf

3. Thomas A, Chen Y, Yu T, Jakopovic M, Giaccone G. Trends and Characteristics of Young Non-Small Cell Lung Cancer Patients in the United States. Front Oncol. 2015;5:113.

4. Suidan AM, Roisman L, Belilovski Rozenblum A, et al. Lung Cancer in Young Patients: Higher Rate of Driver Mutations and Brain Involvement, but Better Survival. J Glob Oncol. 2019;5:1-8.

5. Ellis L, Canchola AJ, Spiegel D, Ladabaum U, Haile R, Gomez SL. Racial and Ethnic Disparities in Cancer Survival: The Contribution of Tumor, Sociodemographic, Institutional, and Neighborhood Characteristics. J Clin Oncol. 2018;36(1):25-33.

6. Surveillance, Epidemiology, and End Results (SEER) Program, National Cancer Institute. SEER*Stat Database: Incidence - SEER Research Data, 17 Registries (excl AK), Nov 2023 Sub (2000-2021). National Cancer Institute, DCCPS, Surveillance Research Program. Released April 2024. Accessed October 17, 2024.

7. Hansen RN, Zhang Y, Seal B, et al. Long-term survival trends in patients with unresectable stage III non-small cell lung cancer receiving chemotherapy and radiation therapy: a SEER cancer registry analysis. BMC Cancer. 2020;20(1):276.

8. Ganti AK, Klein AB, Cotarla I, Seal B, Chou E. Update of Incidence, Prevalence, Survival, and Initial Treatment in Patients With Non-Small Cell Lung Cancer in the US. JAMA Oncol. 2021;7(12):1824-32.