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

Carissa Feliciano

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 Extraction and Data 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 median, minimum, maximum, and interquartile range of survival time were calculated for each racial/ethnic group. The frequency distribution for sex, histology, and cancer stage at diagnosis were summarized for each racial/ethnic group. In Figure 1, boxplots were used to display the differences in survival time among the different racial/ethnic groups. In Figure 2, a barchart was used to display the differences in median survival time for each stage at diagnosis, stratified by race/ethnicity. In Figure 3, a boxplot was used to display the differences in median survival time for each histologic subtype, stratified by race/ethnicity.

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. The patients in this cohort identified as one of the following races/ethnicities: Non-Hispanic White (NHW, n = 1,719), Non-Hispanic Black (NHB, n = 438), Hispanic (n = 1,060), Non-Hispanic Asian or Pacific Islander (NHAPI, n = 1,145), or Non-Hispanic American Indian / Alaska Native (NHAIAN, n = 22). 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 histologic diagnosis was adenocarcinoma.

Table 1: Patient Characteristics by Race/Ethnicity

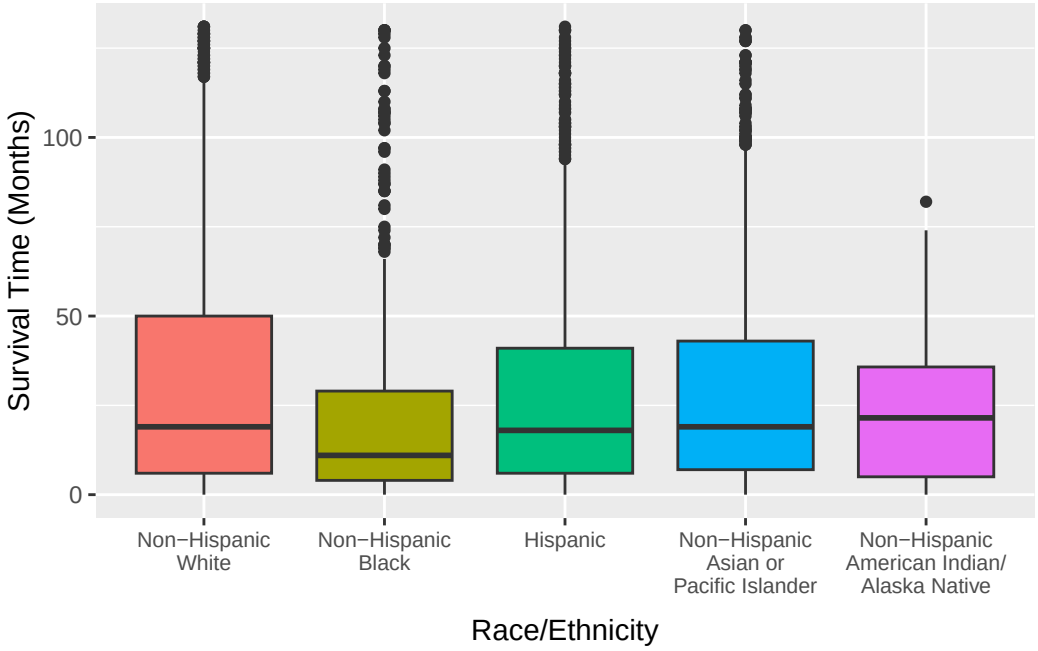
Characteristic	**Non-Hispanic White** N = 1,719[1]	**Non-Hispanic Black** N = 438[1]	**Hispanic (All Races)** N = 1,060[1]	**Non-Hispanic Asian or Pacific Islander** N = 1,145[1]	**Non-Hispanic American Indian/Alaska Native** N = 22[1]
Age (Years)	46 (42, 49)	47 (42, 49)	45 (38, 48)	46 (41, 49)	47 (42, 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 (Months) by Race/Ethnicity

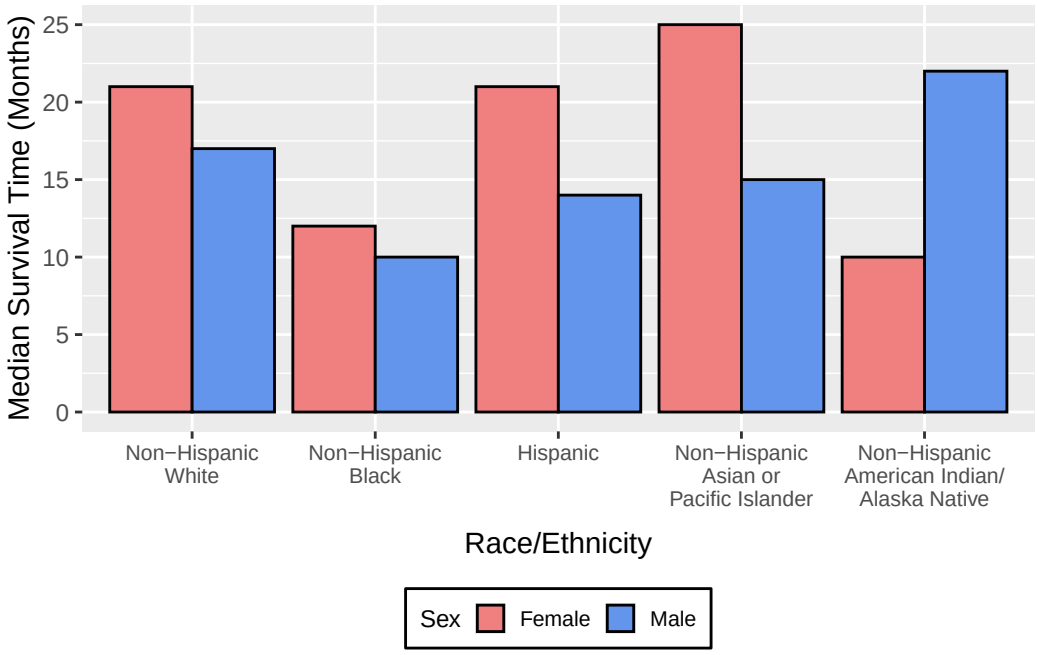
Race/Ethnicity	**Number of Observations**	**Median**	**Min**	**Max**	**1st Quartile**	**3rd Quartile**
Non-Hispanic White	1719	19.0	0	131	6	50.00
Non-Hispanic Black	438	11.0	0	130	4	29.00
Hispanic (All Races)	1060	18.0	0	131	6	41.00
Non-Hispanic Asian or Pacific Islander	1145	19.0	0	130	7	43.00
Non-Hispanic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

Figure 1: Survival Time by Race/Ethnicity



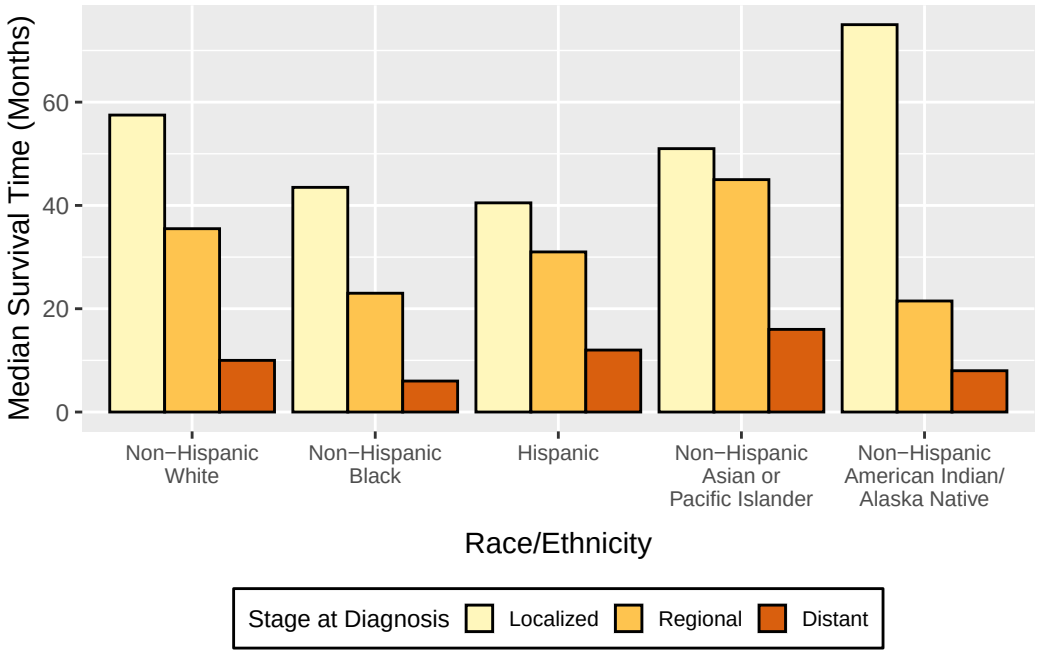
The median survival time was lowest for the NHB group (11.0 months) and highest for the NHAIAN group (21.5 months) (Table 1). The median survival time was similar for the NHW (19 months), Hispanic (18 months), and NHAPI (19 months) groups. The small sample size of the NHAIAN group (n = 22) limits the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

Figure 2: Median Survival Time by Race/Ethnicity and Sex



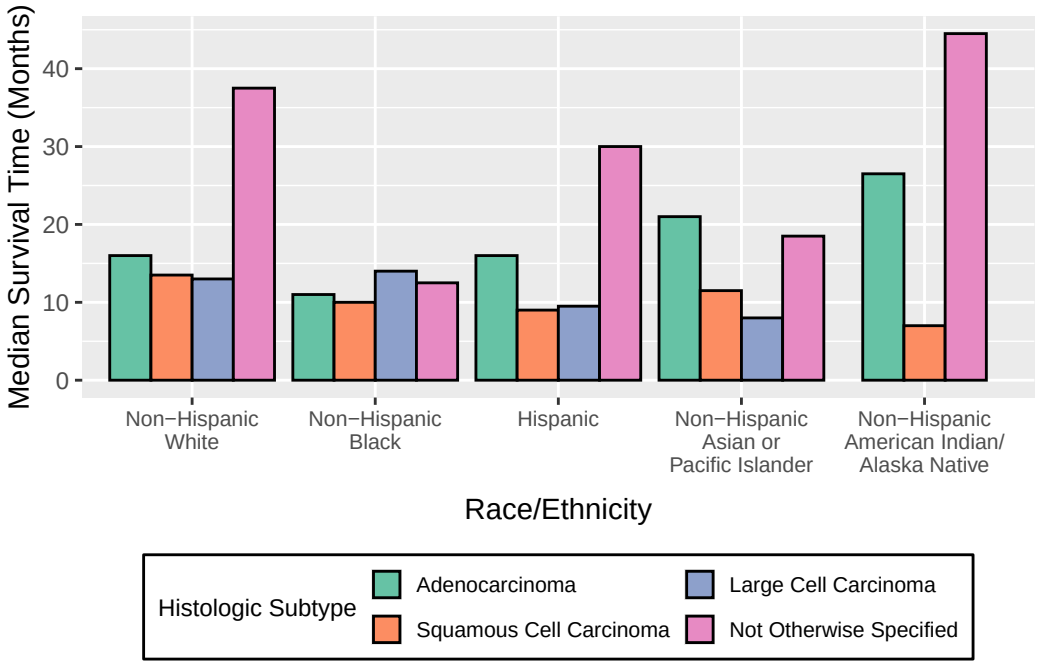
The NHB group had a similar percentage of female subjects (52%) as the NHW (50%) and NHAPI (54%) groups. When stratifying by race/ethnicity and sex, there still remained differences in median survival between the different racial ethnic groups for males and females. Females in the NHB group had a significantly lower median survival than females in the NHW, Hispanic, and NHAPI groups. The differences in the distribution of males/females likely do not explain the differences in survival across the racial/ethnic groups.

Figure 3: Median Survival Time by Race/Ethnicity and Stage at Diagnosis



The NHB group had a lower percentage of patients diagnosed with localized cancer (15%) compared to the NHW (20%) and Hispanic (19%) groups, which may have contributed to the lower median survival time. However, NHAPI had a lower percentage of patients diagnosed with localized cancer (11%) compared to the NHB group (15%) despite having a higher median survival time. When stratifying by race/ethnicity and stage at diagnosis, there still remained differences in medial survival across racial/ethnic groups among those with the same stage at diagnosis (Fig. 3). Differences in the distribution of cancer stage may contribute to differences in the median survival time across the racial/ethnic groups but do not fully explain the differences.

Figure 4: Median Survival Time by Race/Ethnicity and Histologic Diagnosis



The NHB group had a similar distribution of adenocarcinoma and NOS as the NHW and Hispanic groups. When stratifying by race/ethnicity and histologic subtype, there still remained differences in median survival across the different racial/ethnic groups for the same histologic subtype (Fig. 4). In the NHW, Hispanic, and NHAIAN groups, patients with a histologic diagnosis classified as "Not Otherwise Specified" had a significantly higher median survival compared to the other histologic subtypes. However, this was not true for the NHB group. It appears that differences in distribution of histologic subtypes did not contribute much to the differences in median survival.

Conclusion

In conclusion, the median survival appeared to be significantly lower for the NHB group compared to the NHW, Hispanic, NHAPI, and NHAIAN groups. These differences may be partially explained by differences in stage at diagnosis. Upon preliminary analysis, it does not appear that differences in the frequency distributions of sex and histology contributed much to the differences in median survival across the different racial/ethnic groups.

References