Final Project Report

Erica Shin

1. Introduction

Background:

Leukemia, a type of blood cancer characterized by the rapid growth of abnormal white blood cells in the bone marrow, is the most common type of cancer diagnosed in children - accounting for approximately 28% of all pediatric cancers in the United States. The incidence of leukemia in the pediatric population has raised public health concerns due to its impact on the long-term survival and quality of life of young patients.

The National Cancer Institute (NCI) collects cancer incidence data from population-based cancer registries covering approximately 47.9 percent of the US population in the Surveillance, Epidemiology, and End Results Program (SEER). The SEER registries collect data on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, first course of treatment, and more. This report focuses specifically on leukemia cases within the SEER dataset to investigate trends and factors influencing survival among pediatric patients diagnosed with this condition.

Question of Interest:

This report aims to address the following research question: **What demographic factors (e.g., sex, race, median income, age) are associated with the likelihood of longer survival times in deceased pediatric patients under age 10 diagnosed with malignant leukemia in the US?**

While advancements in medical research and treatment methods have led to improved survival rates, disparities persist based on demographic factors such as sex, race, socioeconomic status, and age. Understanding these disparities is crucial for developing targeted interventions that can enhance treatment efficacy and support systems for vulnerable populations.

---

2. Methods

The data for this analysis was sourced from the SEER registries, which provides comprehensive cancer statistics and serves as a crucial resource for understanding trends in cancer incidence, treatment, and survival across the United States. To access and analyze the SEER data, the statistical software tool **SEER*Stat** was used. This tool allows for the selection and analysis of variables within the SEER registries and is available for download on the NCI's website.

The dataset utilized in this report comes from the SEER registry titled "**Incidence - SEER Research Limited-Field Data, 22 Registries, Nov 2023 Sub (2000-2021) - Linked to County Attributes - Time Dependent (1990-2022) Income/Rurality, 1969-2022 Counties**." This particular registry was chosen due to its extensive geographic coverage, representing approximately 47.9% of the U.S. population, as based on the 2020 U.S. Census. The dataset for the report was curated by selecting specific variables from the registry and offers a comprehensive view of the factors influencing survival among pediatric leukemia patients, with a focus on key demographic variables such as sex, race, and socioeconomic status.

To explore the question of interest, the dataset for this report focuses on the following 13 variables:

- **Patient ID**: Integer identifier for each patient.
- **Sex**: Character variable indicating the patient's sex (e.g., "Female", "Male").
- **Race (Recode: White, Black, Other)**: Character variable representing the patient's race.
- **Race Ethnicity**: Character variable detailing the specific ethnicity of the patient.
- **Age (Recode with Single Ages and 90+)**: Character variable indicating the patient's age.
- **Age (Recode with <1 Year Olds)**: Character variable representing age ranges.
- **Year of Diagnosis**: Integer indicating the year the patient was diagnosed.
- **Site (Recode ICD-O-3/WHO 2008)**: Character variable specifying the type of leukemia diagnosed.
- **Behavior Code (ICD-O-3)**: Character variable indicating the malignancy status.
- **Year of Death (Recode)**: Character variable denoting whether the patient is alive or the year of death.
- **Type of Reporting Source**: Character variable describing the source of the cancer data (e.g., hospital or clinic).
- **Median Household Income (Inflation Adjusted to 2022)**: Character variable representing the income bracket of the patient's household.

- **Rural-Urban Continuum Code**: Character variable indicating the population type of the area where the patient resides.

The names of all 13 variables were taken directly from SEER, with recodings also performed by SEER.

Data Exploration Tools:

The following tools were used to explore the data.

- **dim()**: Used to look at the dimensions of the initial dataset (26,916 observations by 13 variables).
- **str()**: Provided information about the types of variables contained in the initial dataset (2 integer and 11 character variables).
- **colnames()**: Provided information about the variable or column names.
- **head()** and **tail**(): Utilized to look at the first and last six observations.
- **table()**: Used to count occurrences of specific variables.
- **summary()**: Provided statistical summaries like mean, median, min, max, and quartiles for specific variables.

Cleaning and Wrangling:

To clean the data, the column names were renamed to simplify the variable names for easier use in coding. Missing observations in key variables (sex, race, age, age range, and median income) were then addressed. The only variable with missing values was median income (med_income), and the corresponding observations were removed from the dataset.

For the analysis, the focus was placed on the survival time of patients who had already passed away. Therefore, observations categorized as "Alive at last contact" were excluded, retaining only those with a recorded year of death. It is important to note that this filtering process may introduce some bias, as factors associated with better survival are less likely to be represented in the dataset. After filtering, the dataset contained 3,457 observations and 14 variables.

Subsequently, the variables for year of death (year_death) and age (age) were originally character variables, so they were converted to numeric types for easier calculations and visualizations. A new variable, "surv," was then created to calculate the survival time for each patient by subtracting the year of diagnosis from the year of death.

---

3. Results

a. Summary Statistics

Table 1. Summary Statistics of Survival Time by Sex

Sex	Count of Patients	Average Survival (Years)	Median Survival (Years)	Standard Deviation
Female	1564	2.167519	1	2.946973
Male	1893	2.494453	1	3.141080

Table 2. Summary Statistics of Survival Time by Race

Race	Count of Patients	Average Survival (Years)	Median Survival (Years)	Standard Deviation
Black	427	2.025761	1	2.749612
Other (American Indian/AK Native, Asian/Pacific Islander)	307	2.446254	1	3.343056
Unknown	20	2.450000	1	3.531438
White	2703	2.385128	1	3.066261

Table 3. Summary Statistics of Survival Time by Race and Sex

Race	Sex	Count of Patients	Average Survival (Years)	Median Survival (Years)	Standard Deviation
Black	Female	200	1.7800000	1	2.614532
Black	Male	227	2.2422907	1	2.851544
Other (American Indian/AK Native, Asian/Pacific Islander)	Female	147	2.3401361	1	3.274008
Other (American Indian/AK Native, Asian/Pacific Islander)	Male	160	2.5437500	1	3.412605
Unknown	Female	7	0.8571429	0	1.463850
Unknown	Male	13	3.3076923	3	4.049375
White	Female	1210	2.2181818	1	2.959109
White	Male	1493	2.5204287	1	3.144917

Table 4. Summary Statistics of Survival Time by Age and Age Range of Diagnosis

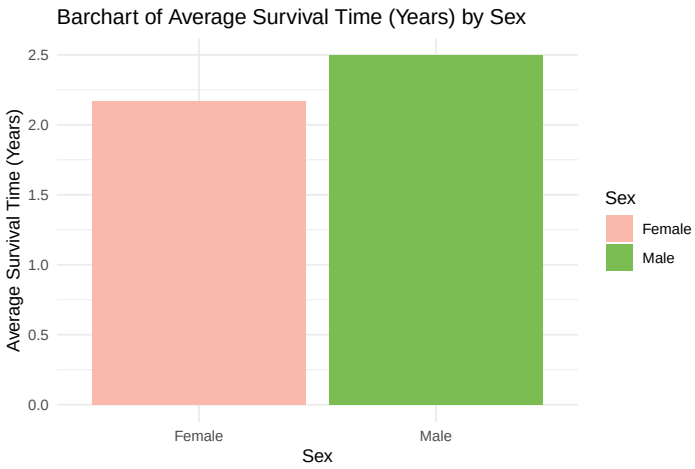
Age of Diagnosis	Age Range of Diagnosis	Count of Patients	Average Survival (Years)	Median Survival (Years)	Standard Deviation
0	00 years	731	1.142271	1	1.723807
1	01-04 years	479	2.016701	1	2.622667
2	01-04 years	419	2.622912	2	3.249741
3	01-04 years	381	2.787402	2	3.209271
4	01-04 years	270	3.362963	2	3.984370
5	05-09 years	238	2.760504	2	3.240923
6	05-09 years	212	3.047170	2	3.538233
7	05-09 years	242	3.078512	2	3.461410
8	05-09 years	230	2.500000	1	2.866384
9	05-09 years	255	2.427451	1	3.198430

Table 5. Summary Statistics of Survival Time by Median Income

Median Income Range	Count of Patients	Average Survival (Years)	Median Survival (Years)	Standard Deviation
$120,000+	115	2.426087	1	3.214716
$110,000 - $119,999	138	2.644928	1	2.929340
$100,000 - $109,999	227	2.392070	1	3.239760
$95,000 - $99,999	146	2.150685	1	2.634892
$90,000 - $94,999	159	2.496855	1	2.999735
$85,000 - $89,999	146	2.486301	1	3.132117
$80,000 - $84,999	332	2.481928	1	3.301969
$75,000 - $79,999	271	2.261993	1	3.242312
$65,000 - $69,999	345	2.342029	1	2.861960
$60,000 - $64,999	338	2.059172	1	2.761980
$55,000 - $59,999	235	2.293617	1	3.009061
$50,000 - $54,999	145	2.013793	1	2.728416
$45,000 - $49,999	128	2.234375	1	3.225109
$40,000 - $44,999	104	1.692308	1	1.946253
< $40,000	25	1.920000	1	2.413849
NA	603	2.601990	1	3.338708

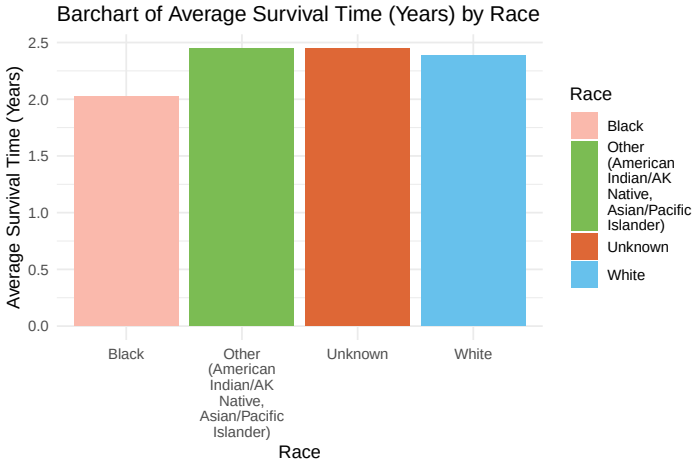
b. Visualizations

Figure 1. Barchart of Average Survival Time (Years) by Sex



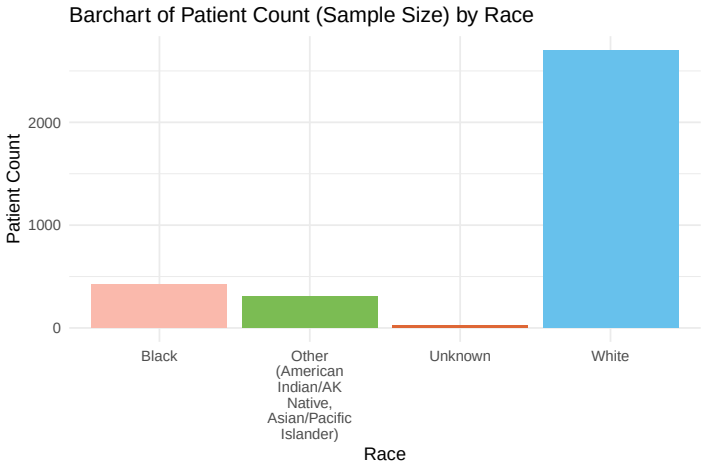
Males had a higher average survival time (2.494 years) compared to females (2.168 years).

Figure 2. Barchart of Average Survival Time (Years) by Race



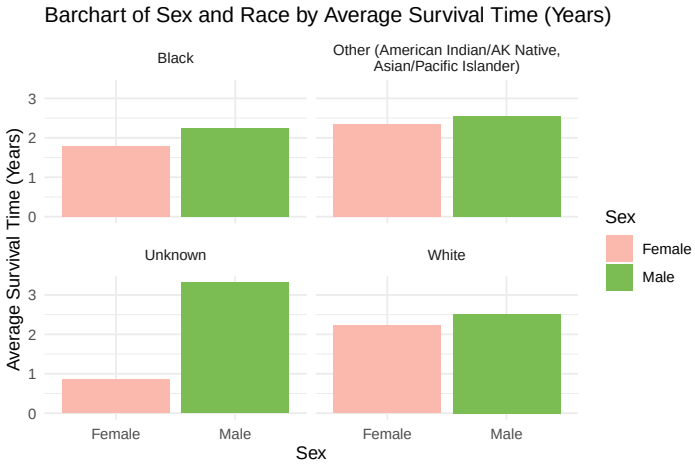
Among racial categories, the "Unknown" group had the highest average survival time at 2.450 years, closely followed by "Other (American Indian/AK Native, Asian/Pacific Islander)" at 2.446 years.

Figure 3. Barchart of Patient Count (Sample Size) by Race



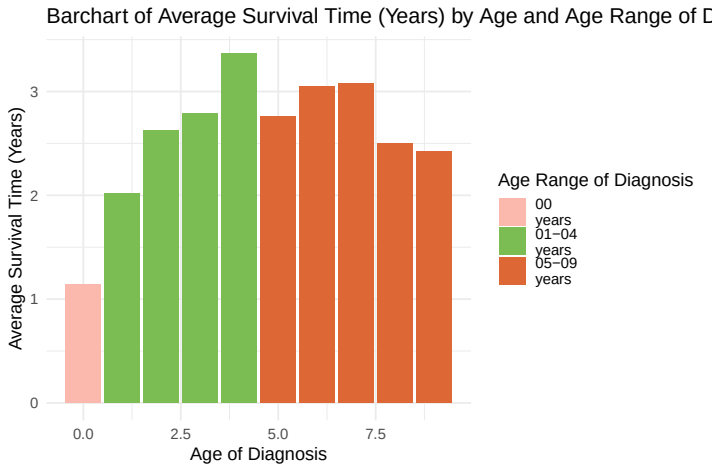
The majority of patients in the dataset were classified as "White" (2703 out of 3457 patients or 78%) while the "Unknown" category contained only 20 patients (5.8%), compared to 307 for "Other" (8.9%) and 427 for "Black" (12.4%) - can also refer to **Table 2.** Summary Statistics of Survival Time by Race.

Figure 4. Barchart of Average Survival Time (Years) by Race and Sex



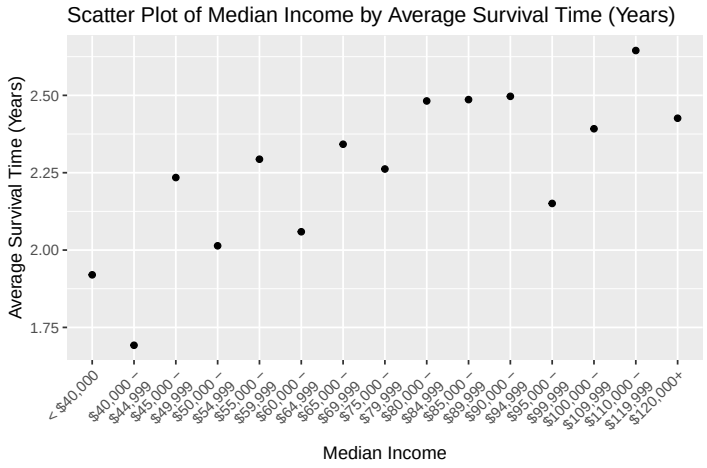
Among the four race categories, the Unknown" group had the largest discrepancy in average survival time between males and females.

Figure 5. Barchart of Average Survival Time (Years) by Age and Age Range of Diagnosis



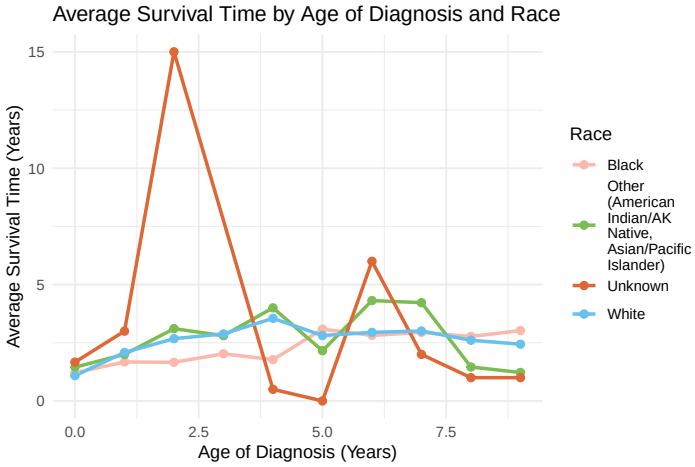
The majority of observations were in the 1-4 and 5-9 year ranges. Newborns (0 years) showed significantly lower average survival times, likely due to their increased vulnerability to disease as well the age range spanning less years than the other categories.

Figure 6. Scatterplot of Average Survival Time (Years) by Median Income



There appears to be a positive association between median income and average survival time, indicating that higher income is correlated with better survival outcomes.

Figure 7. Lineplot Average Survival Time (Years) by Age of Diagnosis and Race



The discrepancy in sample sizes suggests that average survival times for the smaller race categories may be more susceptible to outliers, especially in the "Unknown" group, which exhibited an unusually high average survival time of around 15 years for patients approximately 2.5 years old - can also refer to **Figure 7.** Lineplot Average Survival Time (Years) by Age of Diagnosis and Race.

---

4. Conclusion and Summary

In conclusion, the preliminary analysis indicates that several demographic factors—such as gender, income level, and age—are associated with longer survival times in deceased pediatric patients under age 10 diagnosed with malignant leukemia in the US. Males, older children, and those from higher-income backgrounds generally exhibited longer survival times.

It is important to note that the distribution of patients across racial categories was uneven, with the majority classified as "White" (78% of the dataset). The smaller sample sizes in categories such as "Unknown" (5.8%) and "Other" (8.9%) suggest that these groups' survival data may be more sensitive to outliers, particularly in the "Unknown" category, which reported unusually high survival times, possibly due to a small number of cases with long survival.

Further analysis is required to explore these relationships in more depth and account for potential confounding variables.

The analysis reveals potential disparities that warrant further investigation and suggests that targeted healthcare strategies may be necessary to address these disparities. These findings can inform future research and guide healthcare professionals in providing better support to vulnerable pediatric populations affected by leukemia.