Survival Analysis of Males in the U.S. with Prostate Cancer Based on Race and Ethnicity

Ranni Tewfik

2024-02-23

# 1. Introduction

Because of the discrepancies in the literature regarding the relationship between race/ethnicity and mortality among patients with prostate cancer, one of the motivations for this project is to better understand that relationship using a large national registry of prostate cancer patients. This project is different from what has already been done because other studies have used smaller sample sizes or have not included several racial and ethnic groups in the study population. Therefore, the research question for this project is: Among males diagnosed with prostate cancer, are race and ethnicity associated with time to all-cause mortality?

# 2. Methods

## 2.1 Data Source

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) provides information on cancer incidence and survival in the United States. SEER uses population-based cancer registries that cover about half of the national population. The SEER registries collect data on tumor site and morphology, stage at diagnosis, treatment, vital status, and patient demographics, including race and ethnicity. This project used the SEER\*Stat software (version 8.4.2) to access the SEER 18 (2000-2018) Registry for data on 1,048,575 prostate cancer patients.

## 2.2 Project Design

This project investigated male patients with prostate cancer in the U.S. from 2000 to 2018. All ages and races/ethnicities were included in the study, however, only patients with localized, regional, or distant cancer stage at diagnosis were included. Patients with unknown/missing survival time were excluded.

## 2.3 Covariates

The main outcome of interest was all-cause mortality. The main exposure of interest was race/ethnicity, which included the following categories: “Non-Hispanic White”, “Non-Hispanic Black”, “Non-Hispanic Asian or Pacific Islander”, “Hispanic (all races)”, and “Other”. The other covariates, age at diagnosis and cancer stage at diagnosis, were categorical variables. The age groups were coded as “0-49 years”, “50-59 years”, “60-69 years”, “70-79 years”, and “≥ 80 years”. The categories for stage at diagnosis were “localized” (i.e., cancer is limited to where it originated), “regional” (i.e., cancer has spread to nearby lymph nodes, tissues, or organs), and “distant” (i.e., cancer has spread to distant body parts).

## 2.4 Statistical Analysis

All analyses were conducted using R software (version 4.3.1). Survival analysis was performed for Kaplan-Meier survival estimates and curves. Cox regression was used to compute adjusted hazard ratios and 95% confidence intervals. The Cox proportional hazards assumption was tested using log-log survival curves and extended Cox models for each covariate.

# 3. Results

## 3.1 Exploratory/Descriptive analysis

This project included 776,196 males in the U.S. with prostate cancer from the SEER database. Of these, 540,057 (69.6%) were Non-Hispanic White; 526,468 (67.8%) were age 60-79 years at diagnosis; and 629,222 (81.1%) had localized stage of prostate cancer at diagnosis (Table 1).

Warning: package 'here' was built under R version 4.3.2

here() starts at C:/Users/ranni/Desktop/MADA/Tewfik-MADA-project

Table 1. Study sample characteristics of males in the U.S. with prostate cancer by race and ethnicity.

| **Characteristic** | **Asian**, N = 37,168 | **Black**, N = 112,597 | **Hispanic**, N = 70,220 | **Other**, N = 16,154 | **White**, N = 540,057 |
| --- | --- | --- | --- | --- | --- |
| Age at diagnosis (years) | NA | NA | NA | NA | NA |
| 0-49 years | 566 (1.5%) | 5,830 (5.2%) | 2,272 (3.2%) | 506 (3.1%) | 11,989 (2.2%) |
| 50-59 years | 5,664 (15%) | 30,990 (28%) | 15,275 (22%) | 3,657 (23%) | 103,790 (19%) |
| 60-69 years | 14,724 (40%) | 46,626 (41%) | 28,269 (40%) | 6,557 (41%) | 218,550 (40%) |
| 70-79 years | 11,998 (32%) | 23,489 (21%) | 18,864 (27%) | 4,310 (27%) | 153,081 (28%) |
| 80+ years | 4,216 (11%) | 5,662 (5.0%) | 5,540 (7.9%) | 1,124 (7.0%) | 52,647 (9.7%) |
| Cancer stage at diagnosis | NA | NA | NA | NA | NA |
| Distant | 2,733 (7.4%) | 7,669 (6.8%) | 5,208 (7.4%) | 490 (3.0%) | 30,085 (5.6%) |
| Localized | 28,969 (78%) | 92,605 (82%) | 55,313 (79%) | 14,687 (91%) | 437,648 (81%) |
| Regional | 5,466 (15%) | 12,323 (11%) | 9,699 (14%) | 977 (6.0%) | 72,324 (13%) |
| Survival time (months) | 74 (51) | 74 (49) | 72 (51) | 64 (51) | 78 (50) |
| Vital status | NA | NA | NA | NA | NA |
| 0 | 29,182 (79%) | 83,648 (74%) | 55,450 (79%) | 14,840 (92%) | 405,853 (75%) |
| 1 | 7,986 (21%) | 28,949 (26%) | 14,770 (21%) | 1,314 (8.1%) | 134,204 (25%) |

The histogram for survival time in months shows a right-skewed (positively skewed) distribution in which the mean is greater than the median (Figure 1). In assessing time to death (all-cause mortality), the overall 10-year survival was 68.5%. Non-Hispanic Black males had lower survival than other races and ethnicities across the full follow-up time (1-year = 95.7%; 5-year = 82.2%; 10-year = 66.1%) (Figure 2).

Figure 1. Histogram of survival time distribution in months.

include\_graphics("../../results/figures/survtime\_distribution.png")

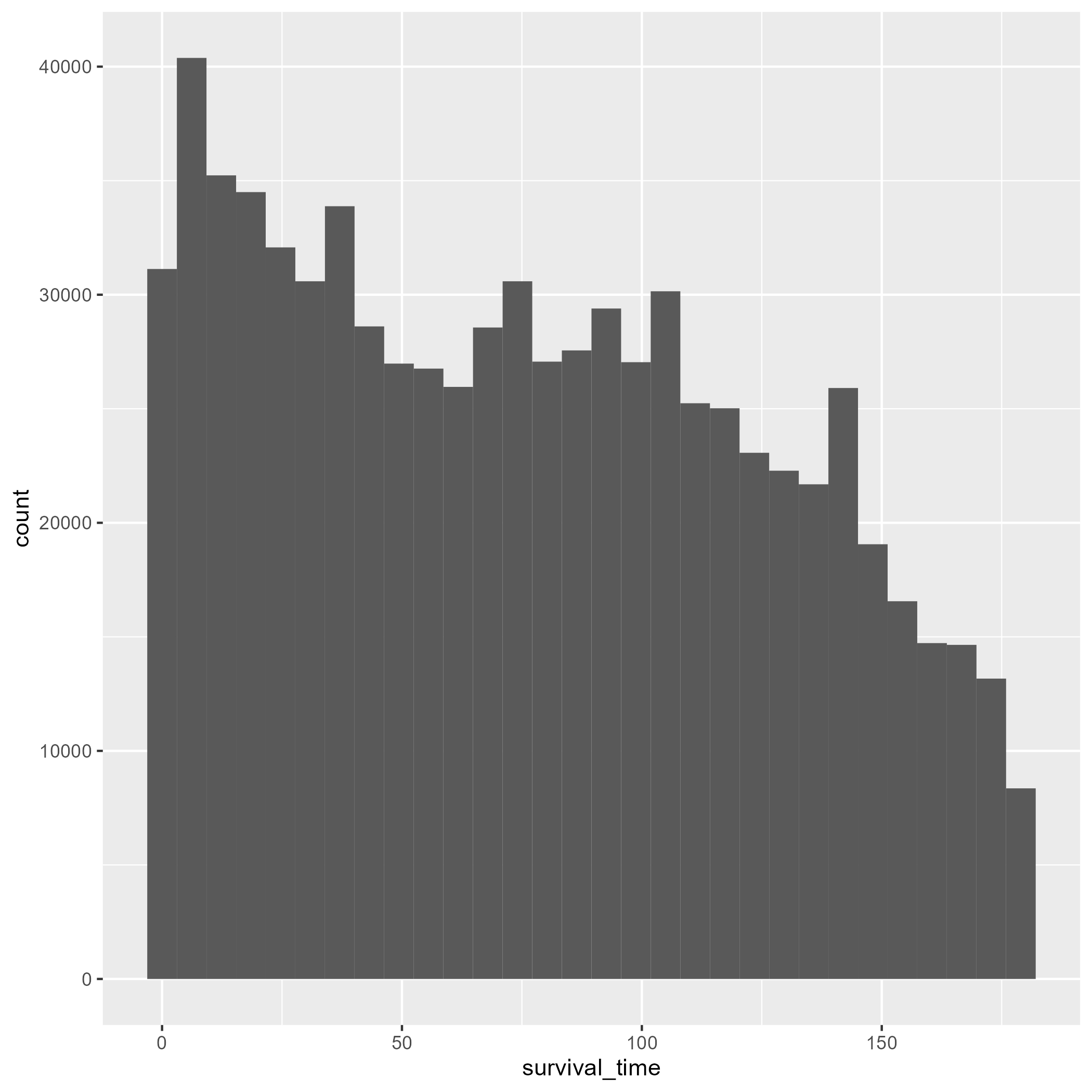
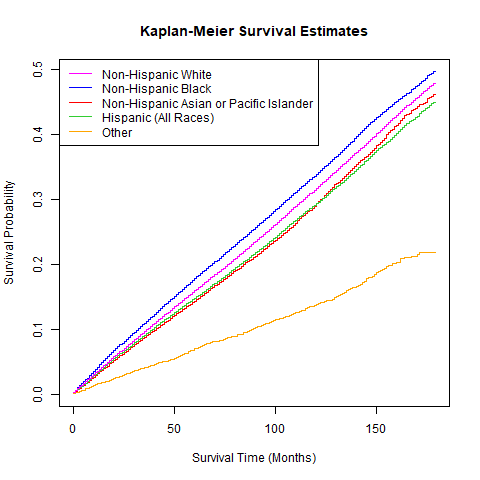


Figure 2. Kaplan-Meier survival curves for males in the U.S. with prostate cancer, stratified by race and ethnicity.

include\_graphics("../../results/figures/survivalcurve.png")



## 3.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

fig-result shows a scatterplot figure produced by one of the R scripts.

## 3.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example tbl-resulttable2 shows a summary of a linear model fit.

|  |
| --- |
| Table 1: Linear model fit table. |

# 4. Discussion

## 4.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 4.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 4.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper [@leek2015] discusses types of analyses.

These papers [@mckay2020; @mckay2020a] are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 5. References