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

Ranni Tewfik

2024-03-15

# 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

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; 629,222 (81.1%) had localized stage of prostate cancer at diagnosis; 401,176 (51.7%) had at least a college degree; and 490,476 (63.2%) had a family history of cancer (Table 1). The mean survival time for each ethnic group of prostate cancer patients ranged from 64 months (“Other”) to 78 months (“White”) (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%) |
| Educational attainment | NA | NA | NA | NA | NA |
| college or higher | 19,241 (52%) | 58,406 (52%) | 36,207 (52%) | 8,338 (52%) | 278,984 (52%) |
| high school or less | 13,459 (36%) | 40,463 (36%) | 25,585 (36%) | 5,834 (36%) | 194,320 (36%) |
| some college | 4,468 (12%) | 13,728 (12%) | 8,428 (12%) | 1,982 (12%) | 66,753 (12%) |
| Family history of cancer | 23,368 (63%) | 71,355 (63%) | 44,362 (63%) | 10,166 (63%) | 341,225 (63%) |
| 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%) |

In assessing time to death (all-cause mortality), the overall 10-year survival was 68.5% (Figure 1). Non-Hispanic Black males had statistically significantly 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). Based on the log-log survival curves and extended Cox models, the proportional hazards assumption holds for all covariates (“Supplementary Material”).

Figure 1. Kaplan-Meier survival curves for males in the U.S. with prostate cancer.

knitr::include\_graphics(here("results/figures/survivalcurve\_total.png"))

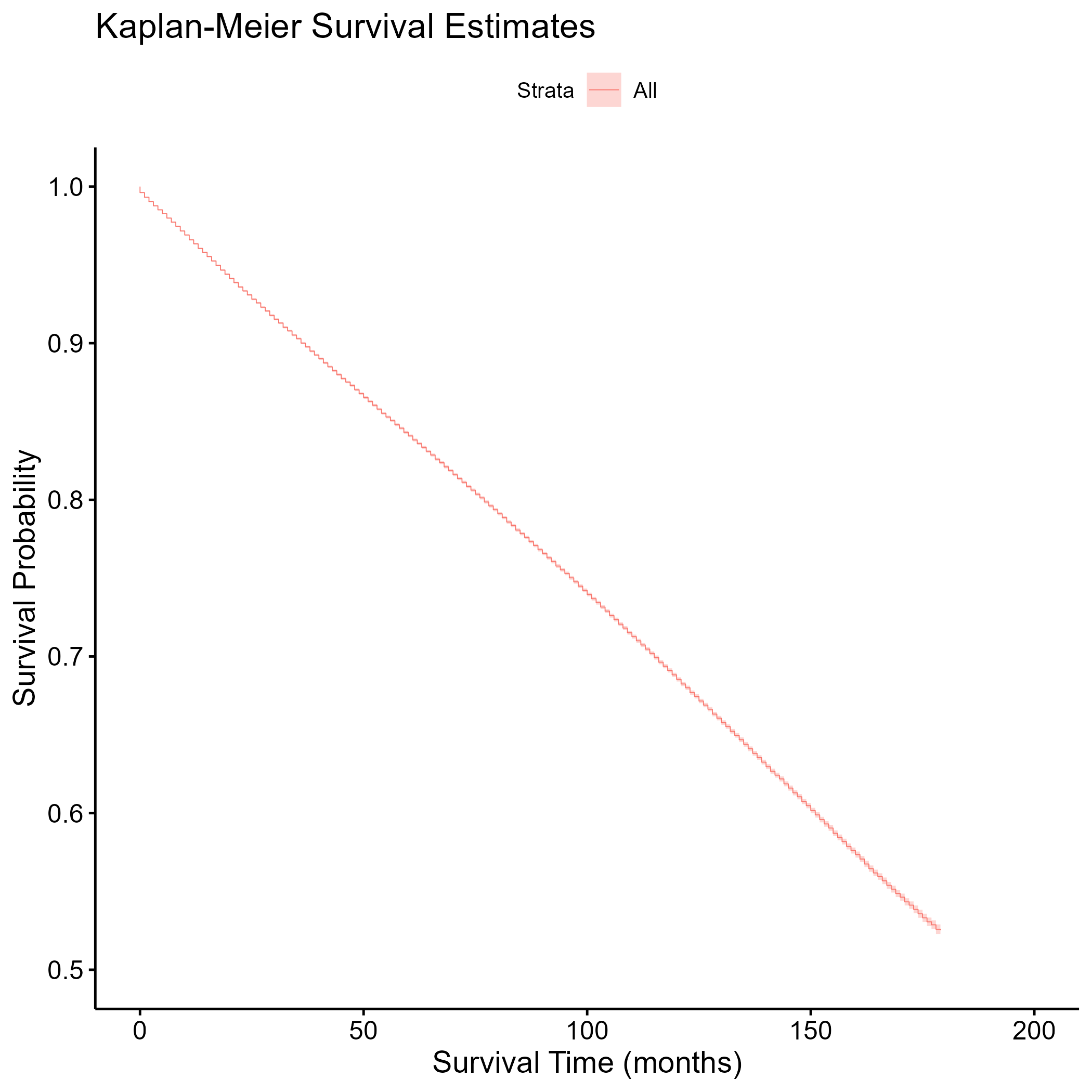
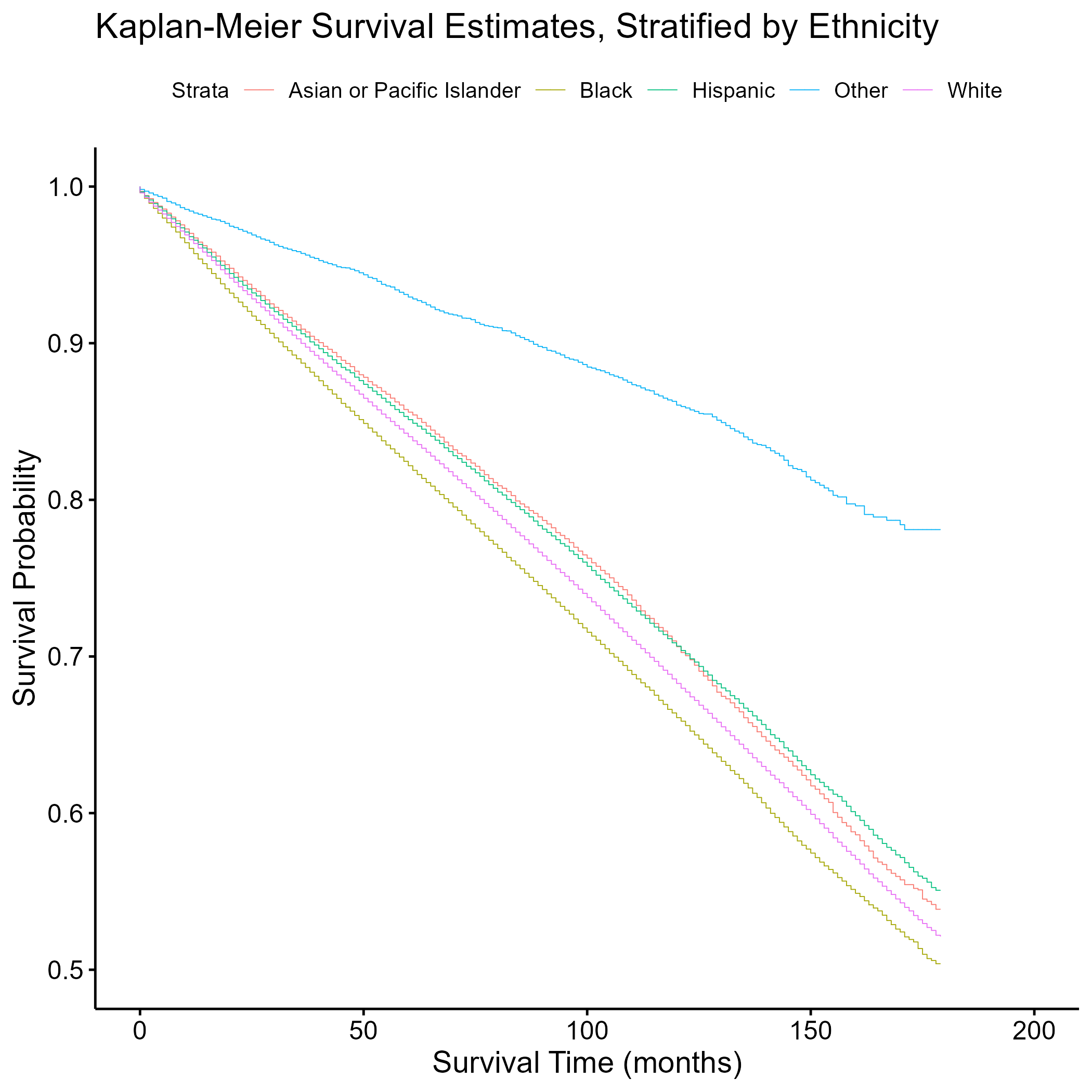


Figure 2. Kaplan-Meier survival curves for males in the U.S. with prostate cancer, stratified by race and ethnicity.

knitr::include\_graphics(here("results/figures/survivalcurve\_ethnicity.png"))



After adjustment for age and cancer stage at diagnosis, the rate of all-cause mortality was 1.32 times higher in Non-Hispanic Black males than Non-Hispanic White males (aHR = 1.32; 95% CI = 1.30 – 1.34) (Table 2). Compared to Non-Hispanic White males, the adjusted hazard ratios were less than 1.00 and statistically significant for males who were Non-Hispanic Asian/Pacific Islander, Hispanic, and other race/ethnicity (Table 2). Controlling for the other covariates, males older than 80 years at diagnosis (vs. males 0-49 years) had the highest adjusted hazard ratio for all-cause mortality among all age groups (aHR = 14.38; 95% CI = 13.70 – 15.09), and males with distant stage of prostate cancer (vs. localized stage) had a higher adjusted hazard ratio than males with regional stage (aHR = 7.27; 95% CI = 7.17 – 7.36) (Table 2).

Table 2. Adjusted hazard ratios and 95% confidence intervals for time to death (all-cause mortality) among males in the U.S. with prostate cancer (SEER database).

|  | Characteristic | Adjusted.hazard.ratio | X95..CI..lower. | X95..CI..upper. |
| --- | --- | --- | --- | --- |
| relevel(factor(ethnicity), ref = “White”)Asian | Asian | 0.7710781 | 0.7538515 | 0.7886984 |
| relevel(factor(ethnicity), ref = “White”)Black | Black | 1.3181728 | 1.3013971 | 1.3351648 |
| relevel(factor(ethnicity), ref = “White”)Hispanic | Hispanic | 0.9160922 | 0.9006471 | 0.9318021 |
| relevel(factor(ethnicity), ref = “White”)Other | Other | 0.4601053 | 0.4357671 | 0.4858028 |
| relevel(factor(agegroup), ref = “0-49 years”)50-59 years | 50-59 years | 1.4482922 | 1.3785081 | 1.5216089 |
| relevel(factor(agegroup), ref = “0-49 years”)60-69 years | 60-69 years | 2.5463864 | 2.4274974 | 2.6710980 |
| relevel(factor(agegroup), ref = “0-49 years”)70-79 years | 70-79 year | 5.7221279 | 5.4558116 | 6.0014440 |
| relevel(factor(agegroup), ref = “0-49 years”)80+ years | 80+ years | 14.3773894 | 13.7013153 | 15.0868235 |
| relevel(factor(stage), ref = “Localized”)Distant | Distant | 7.2662180 | 7.1714705 | 7.3622173 |
| relevel(factor(stage), ref = “Localized”)Regional | Regional | 1.0860275 | 1.0687881 | 1.1035450 |

|  |
| --- |
| 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