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

Ranni Tewfik

2024-02-01

#Load packages and data  
library(here)

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

here() starts at C:/Users/ranni/OneDrive - University of Georgia/Desktop/MADA/Tewfik-MADA-project

library(knitr)  
library(tidyverse)

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

Warning: package 'ggplot2' was built under R version 4.3.2

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

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

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

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.2

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(haven)

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

library(Hmisc)

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

Attaching package: 'Hmisc'  
  
The following objects are masked from 'package:dplyr':  
  
 src, summarize  
  
The following objects are masked from 'package:base':  
  
 format.pval, units

library(psych)

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

Attaching package: 'psych'  
  
The following object is masked from 'package:Hmisc':  
  
 describe  
  
The following objects are masked from 'package:ggplot2':  
  
 %+%, alpha

library(survival)

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

library(survminer)

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

Loading required package: ggpubr

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

Attaching package: 'survminer'  
  
The following object is masked from 'package:survival':  
  
 myeloma

seer1 <- read.csv("SEER.csv")  
  
attach(seer1)

# 1. Summary/Abstract

Research Question

Among males diagnosed with prostate cancer, are race and ethnicity associated with time to all-cause mortality?

Background

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.

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 will use 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 (i.e, observations).

Study Design

The data registry is based on a prospective cohort study that investigated male patients with prostate cancer in the U.S. from 2000 to 2018. All ages and races/ethnicities will be included in this project, however, only patients with localized, regional, or distant cancer stage at diagnosis will be included. Patients with unknown/missing survival time will be excluded.

Covariates

The main outcome of interest is all-cause mortality. The main exposure of interest is race/ethnicity, which includes 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, are categorical variables.

Statistical Analysis

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

#Get an overview and summary of raw data  
str(seer1)

'data.frame': 1048575 obs. of 5 variables:  
 $ race : chr "Non-Hispanic White" "Non-Hispanic White" "Hispanic (All Races)" "Non-Hispanic White" ...  
 $ age : chr "70-74 years" "65-69 years" "55-59 years" "75-79 years" ...  
 $ stage : chr "Localized" "Localized" "Localized" "Blank(s)" ...  
 $ survival\_months: chr "94" "61" "139" "191" ...  
 $ vital\_status : chr "Alive" "Dead" "Alive" "Dead" ...

summary(seer1)

race age stage survival\_months   
 Length:1048575 Length:1048575 Length:1048575 Length:1048575   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
 vital\_status   
 Length:1048575   
 Class :character   
 Mode :character

#Descriptive analysis of raw data  
table(race, useNA = "always")

race  
 Hispanic (All Races)   
 93169   
Non-Hispanic American Indian/Alaska Native   
 3701   
 Non-Hispanic Asian or Pacific Islander   
 49183   
 Non-Hispanic Black   
 148432   
 Non-Hispanic Unknown Race   
 21181   
 Non-Hispanic White   
 732909   
 <NA>   
 0

table(age, useNA = "always")

age  
 00 years 01-04 years 05-09 years 10-14 years 15-19 years 20-24 years   
 5 22 7 7 16 13   
25-29 years 30-34 years 35-39 years 40-44 years 45-49 years 50-54 years   
 22 56 455 4968 22045 68999   
55-59 years 60-64 years 65-69 years 70-74 years 75-79 years 80-84 years   
 135492 186358 217630 176269 123864 67752   
 85+ years <NA>   
 44595 0

table(stage, useNA = "always")

stage  
 Blank(s) Distant In situ Localized   
 225648 46228 1 629393   
 Regional Unknown/unstaged <NA>   
 100816 46489 0

table(vital\_status, useNA = "always")

vital\_status  
 Alive Dead <NA>   
706604 341971 0

survival\_months <- as.numeric(survival\_months)

Warning: NAs introduced by coercion

summary(survival\_months)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 0.00 35.00 82.00 88.46 134.00 227.00 8217

# 2. Introduction

## 2.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above (here dataanalysis\_template\_references.bib) and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in [@mckay2020; @mckay2020a].

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). fig-schematic is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

## 3.2 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 3.3 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

tbl-summarytable shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below.

|  |
| --- |
| Table 1: Data summary table. |

## 4.2 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.

## 4.3 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 2: Linear model fit table. |

# 5. Discussion

## 5.1 Summary and Interpretation

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

## 5.2 Strengths and Limitations

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

## 5.3 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.

# 6. References