# BIOST 515/518 Homework 4

Latera Tesfaye Olana

24 February, 2023

## Responses:

### The data:

Prior to commencing the literature review and providing justifications for which variable ought to be incorporated into the modeling, it is first necessary to provide a brief description of the data. Three observations were found to be missing and were therefore excluded from the analysis. The initial step in undertaking this study is to identify the predictor of interest and the primary outcome. As specified in the research question, the objective is to estimate odds ratios for mortality between groups that differ in their creatinine levels while controlling for various covariates. Consequently, we may reasonably deduce that the predictor of interest is creatinine, while the outcome is mortality, as represented by the occurrence of a death event. Creatinine is a waste product of muscles that is excreted by the kidneys and it was measured at the time of Magnetic Resonance Imaging (MRI) of the study participants (mg/dl). In the given dataset, creatinine level is represented using a variable *crt*. Death is represented by a variable *death* and it is an indicator that the participant was observed to die during the period of the study. The given covarites are *sex*, *race*, *age*, *smoking history*, *pack years*, and *DSST*. The following list provides description of these variables and additional variables created (encoded) for the modelling purposes:

*sex*: The sex of the participant. Only Male and Female are represented. In thus study Male would be represented by a value of 1, where as female by a value of zero.

*race*: Describes participant's race and the values are: White, Black, Asian, or Subject did not identify as White, Black or Asian. Three encoded variable are created to represent this variable. First, *black* which would be 1 if the participants identify or assigned as a black, 0 other wise. Second, *subject_not_identified*: which would be 1 if the participants did not identify themselves as one of the three given races or not assigned to one. Lastly, *white* which would be 1 if the participants identify or assigned as white, 0 otherwise. If all these new encoded variables have a value of zero, then this implies the participants identify or assigned as Asian.

*smoking_histroy*: This represents smoking history of the study participants. Three dichotomous variables were created for representing smoking history. *former_smoker* if the participants are identified to be former smokers and then this variable would be 1, for former smokers and 0 otherwise. *non_smoker* if the participants are identified to be non-smokers. Accordingly, this variable would be 1, for participants who never smoke and 0 otherwise. If an indicator for *former_smoker* and *non_smoker* is zero, then the participants are identified as current smokers.

*pack years*: Participant smoking history in pack years ( pack year = smoking 1 pack of cigarettes per day for 1 year). Pack years is a measure of smoking intensity, and may provide additional information on the relationship between smoking and mortality.

*dsst*: A measure of cognitive function (ability to think) for the participant at the time of MRI.

*Please note that I do not possess a medical background. My background is in quantitative analysis. Identifying associations or causal effects between the given variables requires a profound understanding of the field, which*

*cannot be achieved in a mere three days. While consulting various research works is a useful approach to this issue, I am unable to ascertain the credibility of each research mentioned here.*

Table one provides a summary of all the given variables.

Table 1: Summary of our data

| | Alive | Dead | Total |
|---|---|---|---|
| | (N=595) | (N=125) | (N=720) |
| age | | | |
| Mean (SD) | 74.1 (5.08) | 76.4 (6.07) | 74.5 (5.34) |
| Median [Min, Max] | 73.0 [65.0, 99.0] | 75.0 [67.0, 91.0] | 73.0 [65.0, 99.0] |
| packyrs | | | |
| Mean (SD) | 17.8 (23.9) | 28.0 (35.9) | 19.6 (26.6) |
| Median [Min, Max] | 4.50 [0, 111] | 18.5 [0, 240] | 7.00 [0, 240] |
| crt | | | |
| Mean (SD) | 1.03 (0.244) | 1.20 (0.472) | 1.06 (0.302) |
| Median [Min, Max] | 1.00 [0.500, 1.90] | 1.10 [0.500, 4.00] | 1.00 [0.500, 4.00] |
| sex | | | |
| Female | 318 (53.4%) | 47 (37.6%) | 365 (50.7%) |
| Male | 277 (46.6%) | 78 (62.4%) | 355 (49.3%) |
| race | | | |
| Asian | 36 (6.1%) | 9 (7.2%) | 45 (6.3%) |
| Black | 82 (13.8%) | 20 (16.0%) | 102 (14.2%) |
| Subject did not identify as White, Black or Asian | 9 (1.5%) | 3 (2.4%) | 12 (1.7%) |
| White | 468 (78.7%) | 93 (74.4%) | 561 (77.9%) |
| smoking_history | | | |
| Current smoker | 79 (13.3%) | 20 (16.0%) | 99 (13.8%) |
| Former smoker | 251 (42.2%) | 60 (48.0%) | 311 (43.2%) |
| Non smoker | 265 (44.5%) | 45 (36.0%) | 310 (43.1%) |
| dsst | | | |
| Mean (SD) | 42.5 (12.3) | 34.3 (12.5) | 41.0 (12.7) |
| Median [Min, Max] | 42.0 [0, 82.0] | 34.0 [8.00, 76.0] | 40.0 [0, 82.0] |

## Selecting variables

If found, I will endeavor to utilize evidence derived from systematic review articles, as they are widely regarded as possessing high levels of methodological rigor and quality.

**Is sex a variable of interest?**

Sex could be a variable of interest when studying the relationship between creatinine level and death, as it has been shown that there are differences in creatinine levels between males and females. It was found that women have lower creatinine levels than men, even after adjusting for different factors impacting factors [1]. This difference is thought to be due to hormonal and metabolic differences between the sexes. Furthermore, there are also differences in the prevalence of chronic kidney disease (CKD) and its associated mortality between males and females and CKD progressed more rapidly and was associated with a higher risk of mortality in males[2]. Therefore, when studying the relationship between creatinine level and death through estimating odds ratios of mortality for groups different in their creatinine levels, it would be important to consider sex as a potential confounding variable. This would involve stratifying the analysis by sex or adjusting for sex in our regression model.

Based on the evidence provided above, it appears that sex may be a potential confounding factor. As has been previously noted in this course, it is not feasible to identify confounders solely through data analysis or

statistical methods. While I have formed a clear stance on the relationship between creatinine and mortality in relation to sex, I am curious whether the data suggests any symptoms of confounding of sex for the relationship between creatinine level and death. Looking at figure 1, there seems to a difference in mean of creatinine level in mg/dl comparing values for study participants who died and still alive at the end of the study period. The next question is the difference is by chance (at least as far we are in the space of this data)? In that case, we can do t-test (parametric test). As shown in table 1, this presumed difference is significant (P-value = 0.00085). The last important question is does this mean difference still exists when the data is dis-aggregated by sex? For female this difference completely disappears as it is shown in figure 2 (the p-value of the parametric hypothesis testing in table 3 shows this as well). For male there seems to be a still a difference in mean value and the parametric t-test also supports this (the overall chi-squared test has a p-value of 0.001499). There are several factors that could influence the outcome of the aforementioned test, including the accuracy of the data and the tools utilized. However, it is crucial to supplement our scientific understanding with additional data. In certain circumstances, the data may contradict our initial conclusions, which highlights the need for re-evaluating our scientific discoveries.

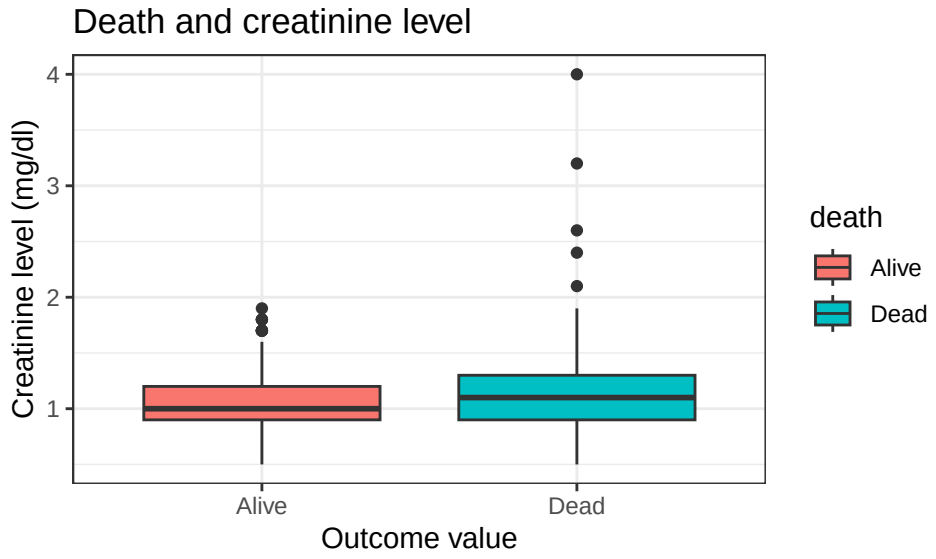



Figure 1: The relationship between creatinine and death

Table 2: Difference in actual mean of creatinine across both outcomes

| estimate | estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
|---|---|---|---|---|---|---|---|---|---|
| -0.16293 | 1.03227 | 1.1952 | -3.75705 | 0.00025 | 138.2609 | -0.24868 | -0.07718 | Welch Two Sample t-test | two.sided |

Table 3: Difference in actual mean in creatinine and death for females

| estimate | estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
|---|---|---|---|---|---|---|---|---|---|
| -0.12108 | 0.91509 | 1.03617 | -1.75636 | 0.08529 | 48.85668 | -0.25962 | 0.01747 | Welch Two Sample t-test | two.sided |

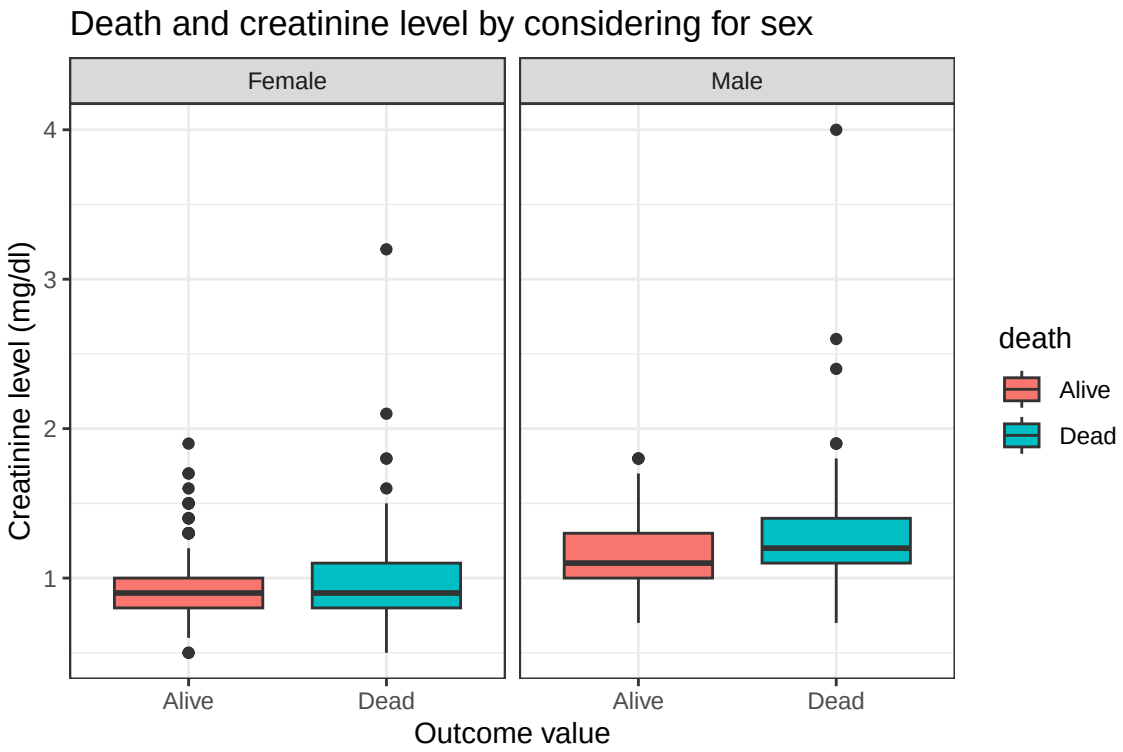



Figure 2: The relationship between creatinine and death adjusted for sex

Table 4: Difference in actual mean in creatinine and death for male

| estimate | estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
|---|---|---|---|---|---|---|---|---|---|
| -0.12424 | 1.16679 | 1.29103 | -2.3596 | 0.02055 | 86.25866 | -0.2289 | -0.01957 | Welch Two Sample t-test | two.sided |

**Is race a variable of interest?**

The 2009 Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) and the 1999 Modification of Diet in Renal Disease Study (MDRD) equations are the most commonly used methods for estimating Glomerular Filtration Rate (GFR), which considers creatinine levels for detecting kidney disease. Black individuals are assigned race multiplicative coefficients of 1.16 and 1.21 in these equations. These formulas were developed by combining direct measurements of kidney function based on iothalamate infusion and urinary clearance with statistical modeling to create equations that include coefficients for race, gender, and age. The race coefficients were included to account for differences in blood creatinine levels between Black and white individuals with similar measured GFR. The MDRD study hypothesized that differences in blood creatinine levels were due to greater muscle mass in Black individuals [3], but this claim was not scientifically substantiated, and subsequent studies have not definitively proven this hypothesis. The inclusion of race in these formulas represents an example of ecological fallacy [4], as race is an estimated variable that cannot be accurately measured or captured. Furthermore, the racial group of Black individuals is genetically and socially diverse, including multiracial individuals, and the inclusion of race in these formulas assumes a homogeneity that is not present [5]. It is important to acknowledge and address the social injustice and institutionalized racism that negatively impact the well-being of disadvantaged racial groups and work towards creating a system that is not biased towards any particular racial group. In general, I will not be using race in my model.

What does the data have to say? The box-plots in figure 3 shows the association between creatinine level

across outcome groups of the study participants. In white race group, there seems to be a large difference in mean of creatinine level across the two outcome groups (Alive and dead). However, as we have specified race is not a scientific variable, and this association only implies to other factors surrounding this issue. As an example, it would be simple to observe a connection between individuals who were not classified or designated as belonging to any particular racial group, by examining creatinine levels across participants with various results. However, simply stating and characterizing this difference on its own would not provide any meaningful information.

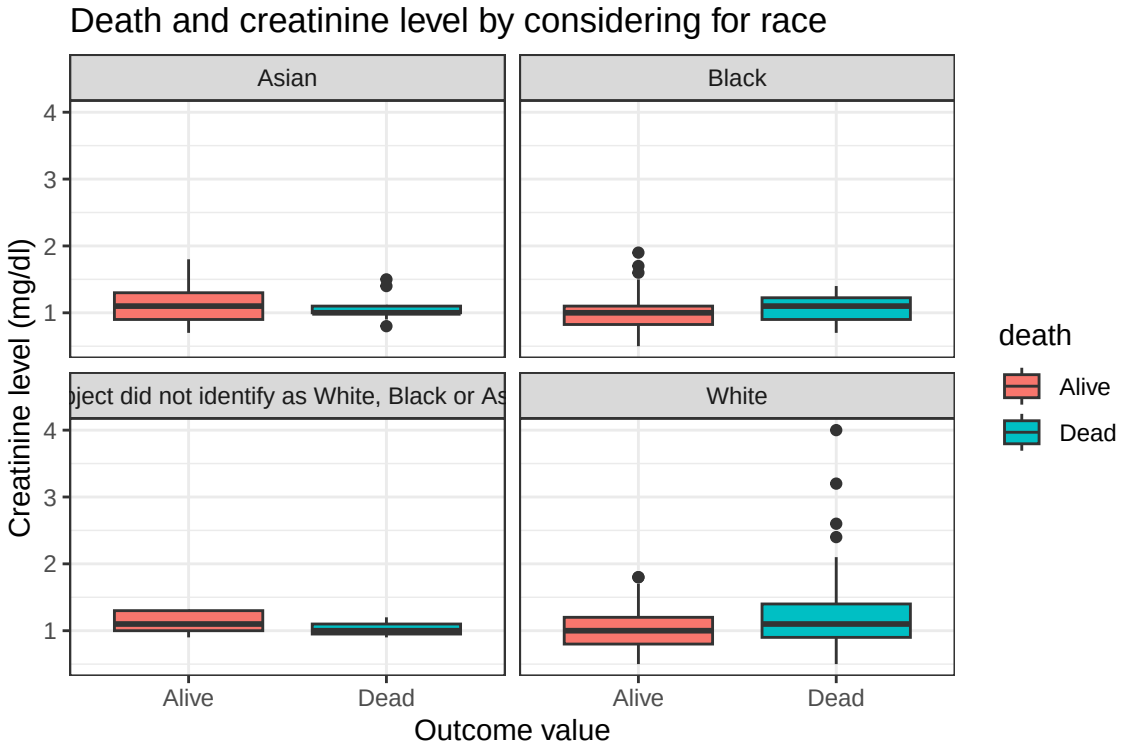



Figure 3: Death and creatinine level by considering for race

**Is age is a variable of interest?**

The concentration of serum creatinine tends to rise gradually with age, specifically in females from the age of 40 years and males from the age of 60 years. The impact of age on serum creatinine levels is particularly notable in the 60 to 80-year-old age group [6]. Age is a relevant variable to consider when studying the relationship between creatinine level and death because both creatinine levels and the risk of death tend to increase with age. Age is also associated with a higher prevalence of chronic kidney disease and other comorbidities that can affect kidney function and mortality risk. Therefore, it is important to adjust for age when studying the relationship between creatinine level and mortality to avoid confounding [7]. Considering age in our modelling process is benificail. What this data says about our stand? As shown in table 1, the mean difference in age for the study groups differing in their outcome (parametric hypothesis test, p-value $< 0.001$), and as also shown in table 6 and figure 4. Figure 6, shows density plot of age of the participants across the two outcome.

Table 5: Difference in mean age for groups varying by their outcome

| estimate | estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
|---|---|---|---|---|---|---|---|---|---|
| -2.35301 | 74.07899 | 76.432 | -4.04617 | 8e-05 | 162.5016 | -3.50136 | -1.20466 | Welch Two Sample t-test | two.sided |

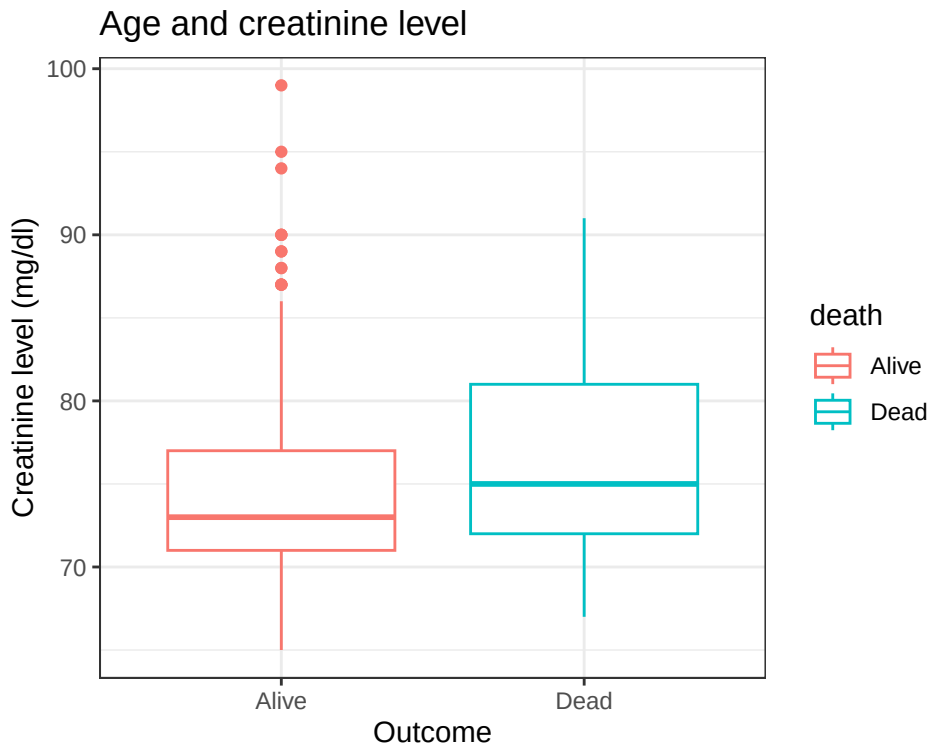
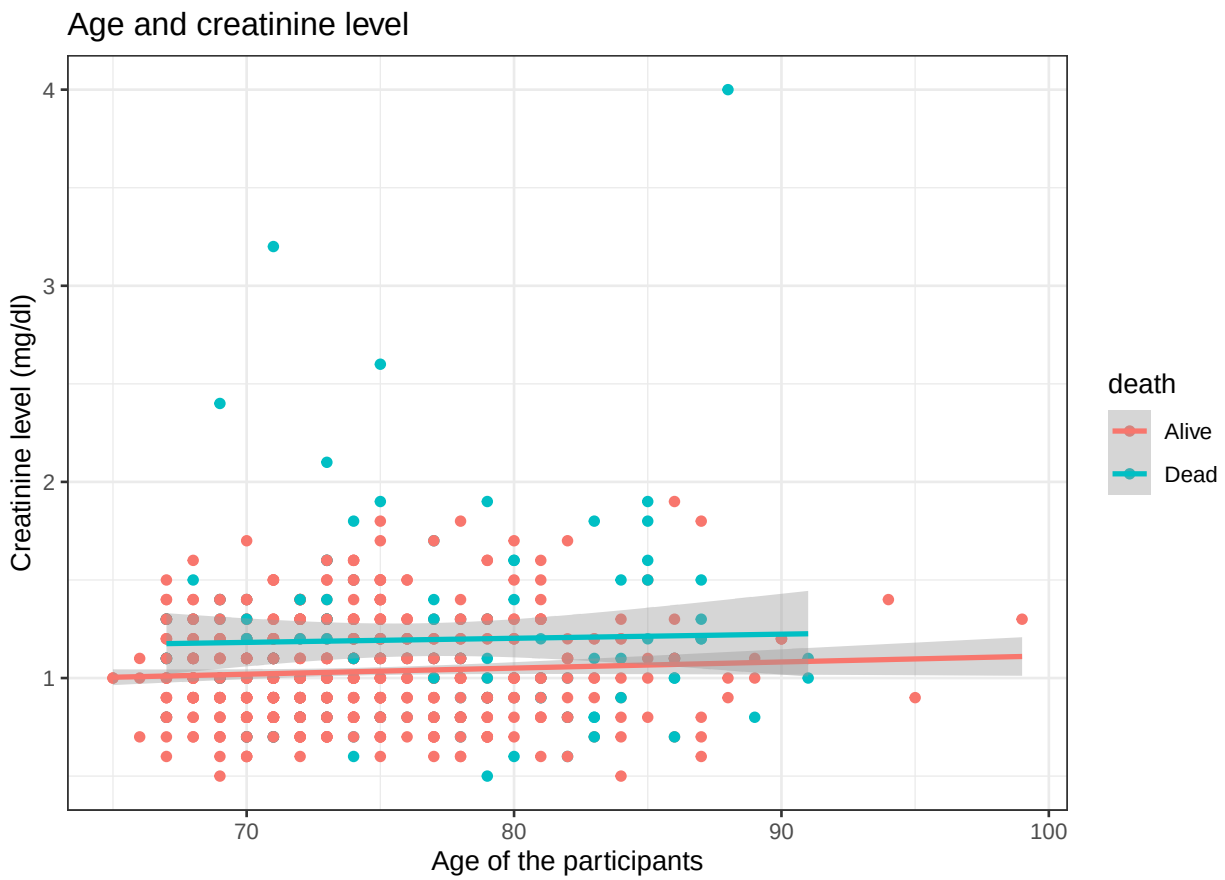



Figure 4: Age and outcome groups




As shown in figure 5, there seems to an increasing trend in creatinine level with increasing age (but very small increments, fitting linear regression with outcome creatinine level and predictor age give us a p-value of 0.02503). However, these relationships are highly impacted by the nature of the data.

**Is smoking a variable of interest?**

From the above given description of variables for this model I will consider a *smoking_histroy*. This variable consists both *packyrs* and *yrsquit* variables described above. The variable for representing packs of cigarettes smoked perday for a year is an important variable indicating the intensity of smoking. However, creatinine level might rapidly change over time and impacted by other factors as well. Accordingly, the impact of heavy smoking years ago might not have impact on creatinine level measured at certain point in time. The creatinine level of current smokers was found to be higher than that of former smokers and individuals who have never smoked. Interestingly, research shows that in the general population, smokers do not demonstrate a lower creatinine level compared to individuals who have never smoked [8]. In addition, the habit of smoking also impacts mortality [9]. Therefore, it is important to adjust for smoking history as characterized in this report (current, former and never smoking statuses).

**Is DSST a variable of interest**

The Development of the Digit Symbol Substitution Test (DSST) evaluates a variety of cognitive processes. To excel in the DSST, one must possess unimpaired motor speed, attention, and visuoperceptual skills such as scanning, writing, or drawing (i.e., fundamental manual dexterity). Associative learning may also impact performance [9]. Even tough, DSST might be impacted by different factors, which also impact creatinine levels and mortality, for the objective of studying the association between creatinine level and mortality for the given study group this has a little of scientific interest. This is an example where making decisions based solely on data may lead us astray. If we investigate the association between variables such as death, creatinine levels, and DSST measures, we may find indications of a relationship. However, if there is no scientific interest in including a particular variable, our statistical tests become meaningless. Hence, statistics should be a

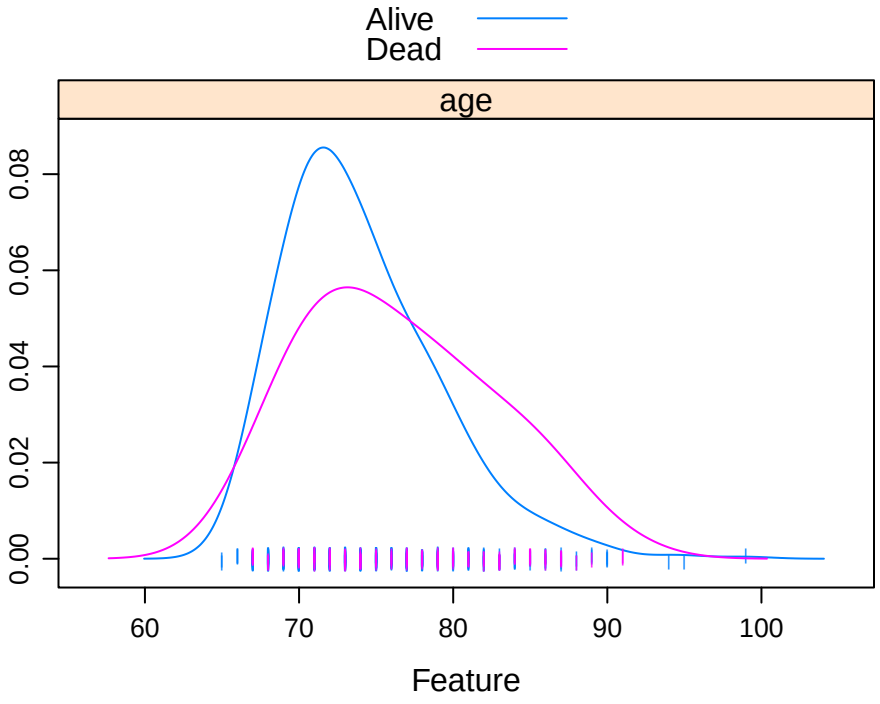



Figure 5: Density plot of age and creatinine level

means of reinforcing our scientific beliefs.

## The model

Using the selected variables, the model can be written as follow:

$$logit(P(death_i = 1 \mid crt_i, age_i, formerSmoker_i, neverSmoker_i, sex_i)) = \beta_0 + \beta_1 \cdot crt_i + \beta_2 \cdot age_i \beta_3 \cdot formerSmoker_i +$$

$$\beta_4 \cdot formerSmoker_i + \beta_5 \cdot *sex_i$$

At baseline we have smoking status, *current* and sex, *female*.

$logit(P(death_i = 1)$ represents the log odds. $age_i$ age of the ith study participant; $formerSmoker_i$ is 1 if the ith participant is a former smoker, 0 otherwise; $neverSmoker_i$ is 1 if the ith participant never smoked, 0 otherwise; $sex_i$ if the ith participant is male, 0 other wise.

Table 6: Transformed logistic fit coefficients

| | e(Est) | e(95%L) | e(95%H) | F stat | df | Pr(>F) |
|---|---|---|---|---|---|---|
| Intercept | 0.00017 | 0.00001 | 0.00340 | 32.65073 | 1 | 0.00000 |
| crt | 3.25436 | 1.64283 | 6.44674 | 11.48751 | 1 | 0.00074 |
| age | 1.08134 | 1.03987 | 1.12447 | 15.41467 | 1 | 0.00010 |
| smoking_history | NA | NA | NA | 1.30612 | 2 | 0.27155 |
| Former smoker | 0.79945 | 0.42489 | 1.50418 | 0.48344 | 1 | 0.48711 |
| Non smoker | 0.60434 | 0.31370 | 1.16426 | 2.27420 | 1 | 0.13201 |

| | e(Est) | e(95%L) | e(95%H) | F stat | df | Pr(>F) |
|---|---|---|---|---|---|---|
| sexMale | 1.37963 | 0.87222 | 2.18224 | 1.89897 | 1 | 0.16865 |

Table 7: Hypothesis test for the logistic regression

| Resid. Df | Resid. Dev | Df | Deviance | Rao | Pr(>Chi) |
|---|---|---|---|---|---|
| 682 | 586.2158 | NA | NA | NA | NA |
| 681 | 575.0961 | 1 | 11.1197 | 11.8508 | 0.00058 |

## Scientific interpretation

We estimate that the odds of a death event within 11 years of the study period is 3.25 times greater for study populations that differ in 1 mg/dl creatinine level but have the same age, sex and smoking history (95% CI for odds ratio: 1.643 6.447 ). Using a likelihood ratio test we conclude that there is a significant difference in the odds of a death event when comparing two populations who differ in one mg/dl creatinine level ($p < 0.00058$).

# Reference


1. Bjornsson TD. Use of serum creatinine concentrations to determine renal function. Clin Pharmacokinet. 1979 May-Jun;4(3):200-22. doi: 10.2165/00003088-197904030-00003. PMID: 383355.
2. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl (2011). 2022 Apr;12(1):7-11. doi: 10.1016/j.kisu.2021.11.003. Epub 2022 Mar 18. PMID: 35529086; PMCID: PMC9073222.
3. Levey, A. S. et al. A new equation to estimate glomerular filtration rate. Ann. Intern. Med. 150, 604–612 (2009).
4. Schwartz, S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. Am. J. Public Health 84, 819–824 (1994).
5. Eneanya, N.D., Boulware, L.E., Tsai, J. et al. Health inequities and the inappropriate use of race in nephrology. Nat Rev Nephrol 18, 84–94 (2022). https://doi.org/10.1038/s41581-021-00501-8
6. Jim Yu-Hsiang Tiao, James B Semmens, John R.L Masarei, Michael M.D Lawrence-Brown, The effect of age on serum creatinine levels in an aging population: relevance to vascular surgery, Cardiovascular Surgery, Volume 10, Issue 5, 2002, Pages 445-451, ISSN 0967-2109, https://doi.org/10.1016/S0967-2109(02)00056-X.
7. Kovesdy, C. P., et al. (2008). Elevated serum creatinine predicts mortality in elderly patients with chronic heart failure: The role of medication use. European Heart Journal, 29(6), 1749-1755. doi: 10.1093/eurheartj/ehn190
8. Jean-Michel Halimi, Bruno Giraudeau, Sylviane Vol, Emile Cacès, Hubert Nivet, Yvon Lebranchu, Jean Tichet, Effects of current smoking and smoking discontinuation on renal function and proteinuria in the general population, Kidney International, Volume 58, Issue 3, 2000, Pages 1285-1292, ISSN 0085-2538, https://doi.org/10.1046/j.1523-1755.2000.00284.x.
9. Lariscy JT, Hummer RA, Rogers RG. Cigarette Smoking and All-Cause and Cause-Specific Adult Mortality in the United States. Demography. 2018 Oct;55(5):1855-1885. doi: 10.1007/s13524-018-0707-2. PMID: 30232778; PMCID: PMC6219821.

## Code Appendix

```
### Setting up the packages
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
# check if packages are installed; if not, install them
packages <- c("tidyverse", "readr", "ggExtra", "plotly",
              "ggplot2","ggstatsplot","ggside","rigr","nlme","lmtest",
              "sandwich")
not_installed <- setdiff(packages, rownames(installed.packages()))
if (length(not_installed)) install.packages(not_installed)

# load packages
library("MASS")
library(sandwich)
library(readr)
library(lmtest)
library(nlme)
library(ggstatsplot)
library(ggside)
library(caret)
library(rigr)
library(ggExtra)
library(broom)
library(plotly)
library(ggplot2)
library(table1)
library(tidyverse) # don't load tidyverse package due to conflict with dplyr
### ------------------------------------------------------------
#Loading working directory of the raw data

#Please load your data/directory by changing it with your work directory
#Throughout this code module you will see a tone of places, where
#data is read and written, so please make sure to change them to your
#working directory folder format

working_directory_data <- setwd("C:/Users/latera/Desktop/Bio_ass")

mri <- read_csv("Data/mri.csv")

mri%>%
  dplyr::select(death, age, packyrs,crt, sex, race,yrsquit,dsst) %>%
  mutate(death = factor(death, levels = c(0, 1),
                            labels = c("Alive", "Dead"))) -> mri

mri$smoking_history <- with(mri, ifelse((packyrs != 0) & (yrsquit == 0), 'Current smoker',
                          ifelse((packyrs == 0) & (yrsquit == 0), "Non smoker", "Former smoker")))

mri$smoking_history <- as.factor(mri$smoking_history)
mri$race <- as.factor(mri$race)
mri$sex <- as.factor(mri$sex)
mri <- na.omit(mri)
```

```
knitr::kable(table1(~ age + packyrs + crt + sex+ race + smoking_history + dsst | death, data=mri,
       overall="Total"),

      digits = 5,caption = "Summary of our data")
ggplot(mri, aes(x = death, y = crt, fill=death)) +
  geom_boxplot() +
  xlab("Outcome value") + ylab("Creatinine level (mg/dl)")+
  ggtitle("Death and creatinine level")+
  theme(legend.position = "none")+
  theme_bw()
ggplot(mri, aes(x = death, y = crt, fill=death)) +
  geom_boxplot() +
  xlab("Outcome value") + ylab("Creatinine level (mg/dl)")+
  ggtitle("Death and creatinine level by considering for sex")+
  facet_wrap(~sex)+
  theme(legend.position = "none")+
  theme_bw()
knitr::kable(tidy(t.test(crt~death, data = mri)),
      digits = 5,caption = "Difference in actual mean of creatinine across
      both outcomes")

mri %>%
  filter(sex == "Female") -> female_mri

knitr::kable(tidy(t.test(crt~death, data = female_mri)),
      digits = 5,caption = "Difference in actual mean in creatinine and death for females")

mri %>%
  filter(sex == "Male") -> male_mri

knitr::kable(tidy(t.test(crt~death, data = male_mri)),
      digits = 5,caption = "Difference in actual mean in creatinine and death
      for male")

ggplot(mri, aes(x = death, y = crt, fill=death)) +
  geom_boxplot() +
  xlab("Outcome value") + ylab("Creatinine level (mg/dl)")+
  ggtitle("Death and creatinine level by considering for race")+
  facet_wrap(~race)+
  theme(legend.position = "none")+
  theme_bw()
knitr::kable(tidy(t.test(age~death, data = mri)),
      digits = 5,caption = "Difference in mean age for groups varying by their outcome")
ggplot(mri, aes(x = death, y = age, col=death)) +
  geom_boxplot() +
  xlab("Outcome") + ylab("Creatinine level (mg/dl)")+
  ggtitle("Age and creatinine level")+
  theme(legend.position = "none")+
  theme_bw()
ggplot(mri, aes(x = age, y = crt, col=death)) +
  geom_point() +
  xlab("Age of the participants") + ylab("Creatinine level (mg/dl)")+
```

```
  ggtitle("Age and creatinine level")+
  theme(legend.position = "none")+
  geom_smooth(method = "lm")+
  theme_bw()
lm(crt~age, data = mri)

featurePlot(x = mri[, c("age")],
y = mri$death,
plot = "density",
scales = list(x = list(relation = "free"),
y = list(relation = "free")),
adjust = 1.5,pch = "|",
layout = c(1, 1),
auto.key = list(columns = 1))

mri <- read_csv("Data/mri.csv")

mri$smoking_history <- with(mri, ifelse((packyrs != 0) & (yrsquit == 0), 'Current smoker',
                         ifelse((packyrs == 0) & (yrsquit == 0), "Non smoker", "Former smoker")))

mri$smoking_history <- as.factor(mri$smoking_history)
mri$race <- as.factor(mri$race)
mri$sex <- as.factor(mri$sex)
mri <- na.omit(mri)

model_reg <- regress("odds", death ~ crt + age + smoking_history + sex, data = mri)


knitr::kable(model_reg$transformed,

      digits = 5,caption = "Transformed logistic fit coefficients")
hy_test <- anova(glm(death ~  age + smoking_history + sex, data = mri, family = "binomial"),
 glm(death ~ crt + age + smoking_history + sex, data = mri, family = "binomial"),
 test = "Rao")

knitr::kable(hy_test,

      digits = 5,caption = "Hypothesis test for the logistic regression")
```