Group 23 (Group LGL)

Jinglun Li

Fuyu Guo

Xinhao Li

HW4: 1st Project Check-In

1. What is the general domain/subject area of this project?

This project will focus on the area of cardiovascular disease (CVD). In this study, we aim to study the association between serum biomarkers and mortality rates among patients with diagnosed heart failure.

2. What data will you use, and what is the source?

This project will be based on medical records of 299 patients with heart failure, with left ventricular systolic dysfunction. The dataset was collected and analyzed by Ahmad et al [1]. Then in January 2020, the dataset was elaborated and donated to the University of California Irvine (UCI) Machine Learning Repository by Davide Chicco under the same Attribution 4.0 International (CC BY 4.0) copyright [2]. We downloaded this dataset from the UCI Machine Learning Repository and will use it under the same Attribution 4.0 International (CC BY 4.0) copyright. The dataset can be accessed through https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.

3. What primary questions will you seek to answer?

The association between serum creatine, a biomarker indicating renal dysfunction[3], and mortality rates for patients with heart failure.

4. What secondary questions will you seek to answer

- The association between serum creatine and death risks 30 days after beginning follow-up among patients with heart failure.
- Also, we will explore whether and how demographical features and health conditions modify the association between serum creatine and mortality rates/Death 30-day in the patients with heart failure.

5. What outcome(s)/endpoint(s) will you use?

- The primary outcome in this study is the survival time of the patients.
- The secondary outcome is the death (0 for alive patient, and 1 for dead patient) 30 days after beginning the follow-up of the patients. It is a binary outcome.

6. Statistical Analysis Plan

a) Data cleaning: although the dataset has been elaborated by Davide Chicco, we will check any potential missing data in the outcome, exposure, and covariates. We will report the number of missing data and if the number is less than 10%, we will consider including a "missing" indicator for categorical variables and imputing continuous variables. We will report the final number of patients included in our study.

b) Checking the exposure:

The primary exposure, serum creatine, which is continuous, is usually categorized into two different levels ($\leq$ 1.5 mg/dL for the normal level vs, and > 1.5 mg/dL for the abnormal level). In this project, we will first treat the serum creatine as a continuous variable and

calculate its sample mean, standard deviation, median, and range. Then we will treat it as a binary variable and calculate the proportion of normal and abnormal levels in patients.

c) Checking the outcomes:

We will calculate the average person-time until death in the normal serum creatinine group and the abnormal serum creatinine group. We will also calculate the Death 30-day in these two groups. Survival plots will be made to visualize the mortality rates in serum creatinine groups. Chis-squared tests will be applied to test the difference of Death 30-day in these groups.

d) Checking other covariates.

In this project, age (continuous), sex (male vs. female), anemia (yes vs. no), diabetes (yes vs. no), ejection fraction (≤ 30, 31-44, and ≥ 45), smoking (yes vs. no), platelets (continuous, kilo platelets/mL), and serum sodium (continuous, mEq/L) will be considered as covariates. High blood pressure (yes vs. no) is included in the datasets but the diagnosis criteria is unclear, so we decide not to include this in our analysis. Creatinine phosphokinase value which the author called CPK is also in the dataset. However, after reading other literature, the CPK usually refers to creatine phosphokinase. Due to this inconsistency, we decide not to include this variable in our analysis, either.

The proportions for categorical covariates and mean (standard deviations) for continuous covariates in normal serum creatinine group and abnormal serum creatinine group will be calculated and compared using chi-squared tests and t-tests respectively.

e) Modeling analysis

First, we will do a crude model to assess the association between serum creatine (normal vs. abnormal) and mortality rates using a Cox proportional-hazards model. The outcome is

the survival time with the event (0 for censored and 1 for death) and the only predictor is serum creatine (normal vs. abnormal). Second, based on the crude model we will adjust for age, sex, anemia, diabetes, ejection fraction. Last, we will further adjust for other serum biomarkers including serum sodium and platelets value to get a fully adjusted model.

For Death 30-day, logistic regressions models will be applied following the same adjustment procedure mentioned above from the crude model to the fully adjusted model.

To check whether the findings from models using categorical serum creatinine are robust, we replace the exposure with continuous serum creatinine value for all the models mentioned above.

f) Subgroup analysis for potential effect modification.

To check whether effect modification exists, we will include an interaction term between serum creatine (normal vs. abnormal) and age (<65 vs. ≥65), sex, diabetes, anemia, ejection fraction, smoking, platelets, and serum in turns into the model. The p-value of the interaction term will be used to determine if there are any effects modifications. The same process will be applied for continuous serum creatine values.

g) Checking nonlinearity

To check whether the relationship between serum creatine and mortality risks is linear, we will replace the linear term in the fully adjusted model with a natural spline of serum creatine. The knots and degrees of freedom of the spline will be determined during the following analysis.

7. What are the biggest challenges you foresee in answering your proposed questions and completing this project.

We foresee the biggest problem may be the violation of the proportional-hazards assumption. Also, we are concerned that the limited sample size may prohibit us from including such a number of predictors in the model.

8. Will you seek domain expertise? Why or why not? If so, from whom?

Yes. Because this project is very closely linked to clinical medicine especially CVD and renal diseases, we need to discuss with experts in this area. However, given the limited time of this project and the limited resources, we would like to read related articles and learn background knowledge first. Then, we will put unresolved questions into the discussion board on Canvas to see if any classmates can help us. We believe our class incorporates students from different disciplines related to human health. There must be somebody who can help us.

9. What software package(s) will you use to complete this project?

- R 4.1.1 (R core team) will be used to manipulate data and to do model analysis
- Tidyverse package will be used to make out work easier.
- Survival package will be used to conduct Cox proportional-hazard models.
- ggfortify and gglot2 will be used to visualize our results.

10. Exploratory data analysis (The codes are showed in the end of this file)

- In this project, we will include variables age, sex, anemia, diabetes., ejection fraction, smoking, platelets, serum creatinine, serum sodium, time to the events, and death event.
- We checked missing values in this dataset and found no missing values. Finally, 299 participants' data with 11 variables were included in the analysis.

- The average level of serum creatinine among patients is 1.394 (sd = 1.034) mg/dL. The value ranges from 0.500 mg/dL to 9.400 mg/dL, with the median at 1.394 mg/dL. Based on the histogram and boxplot below, we found the distribution is right-skewed, with outliers with high values.

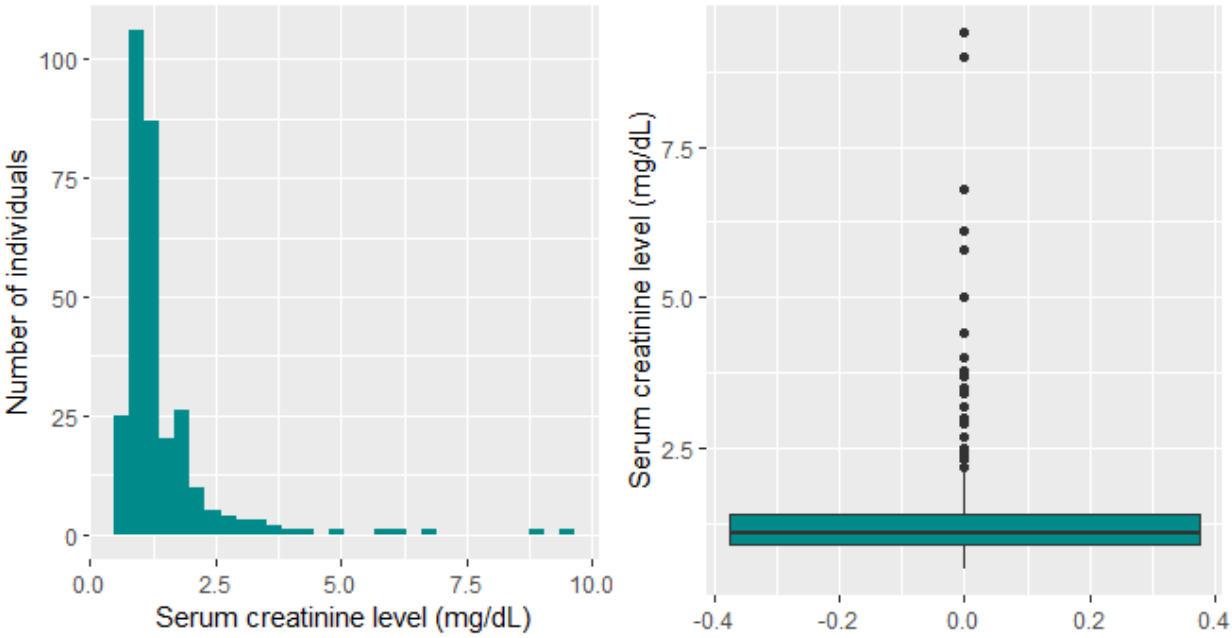


Figure 1. Distribution of serum creatinine in the 299 patients.

- We also calculate the proportion of normal level (≤ 1.5 mg/dL) abnormal level (> 1.5 mg/dL) in the serum creatinine. A total of 232 (77.6%) patients were considered to have normal level serum creatinine value and 67 (22.4%) with abnormal level.
- We calculate the average time to the events in this project. The time to the event in this dataset was measured in the unit of days. A total of 38,948 person-days were contributed by the patients in this study. The average time to the events is 130.3 (sd = 77.61) days. The time ranges from 4 days to 285 days, with a median of 115 days.

  Among patients who died at the end of follow-up, the average time to death is 70.89 (sd = 62.38) days. The time ranges from 4 days to 241 days, with a median of 44.50 days. As

shown in figure 3, in general the time to events among dead participants is greater than the patients who were finally censored.

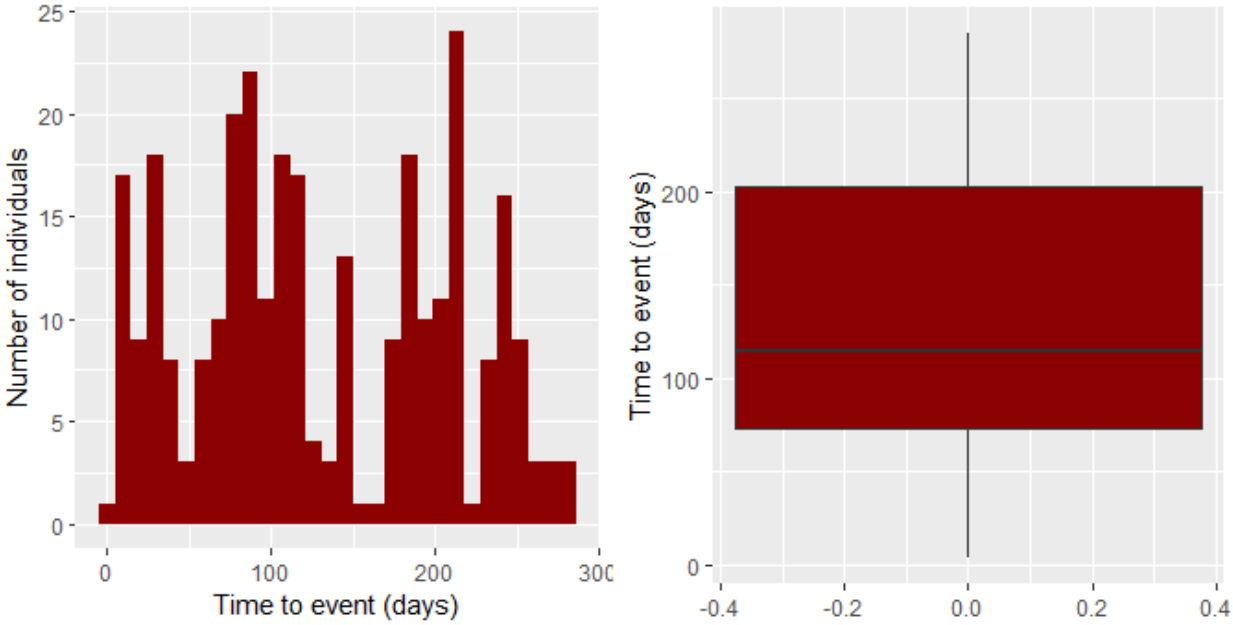


Figure 2. Distribution of the time to the event among the 299 patients.

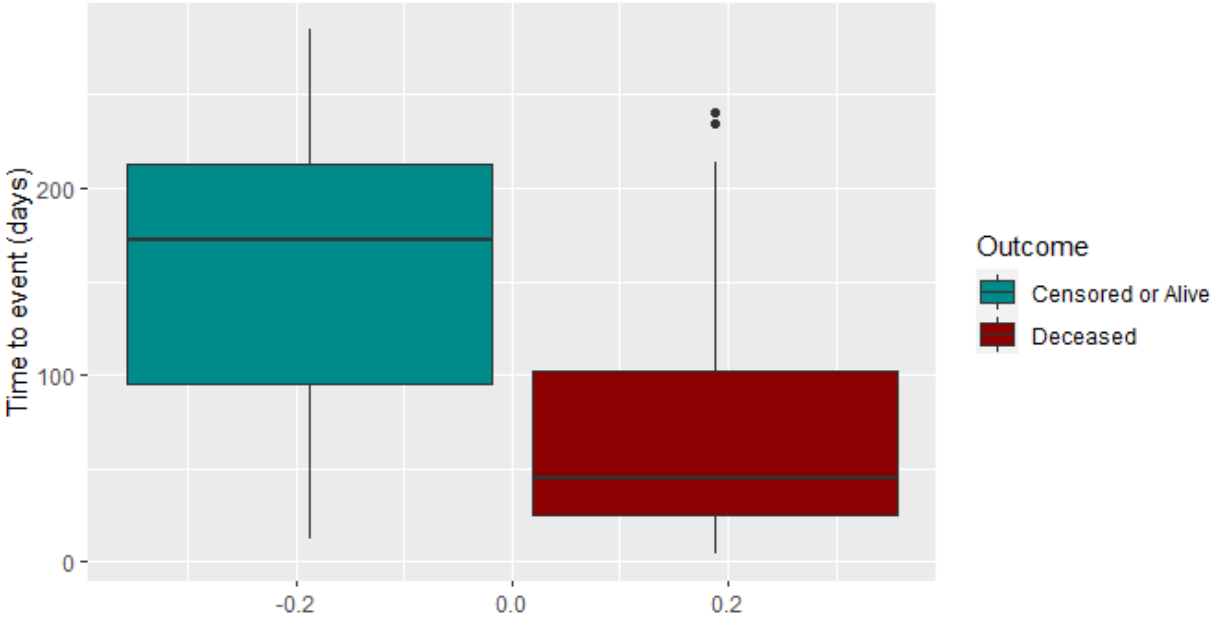


Figure 3. Distribution of the time to the event by death among deceased and alive/censored patients.

- We also calculated the proportion of death 30 days after beginning the follow-up (Death 30-day). A total of 35 (11.7%) patients died within 30 days, and 5 (1.7%) patients were censored, and 259 (86.6%) people survived greater than 30 days.
- We check the distribution of covariates among the patients by serum creatinine level. Details are shown in table 1.

Table 1. Characteristics of the 299 patients by serum creatinine level[a]

	Normal serum creatinine	**Abnormal serum creatinine**	**p-value**
Number	$n = 232$	$n = 67$	
Average serum creatinine[b] (SD)	1.03 (0.21)	2.67 (1.60)	< 0.001
Average time to events (SD)	136.34 (75.81)	109.22 (80.66)	0.0157
Deaths	53 (22.84%)	43 (64.18%)	< 0.001
Death 30-days			0.0085
Censored	4	1	
Death	20	15	
Alive	208	51	
Average age (SD)	59.64 (11.49)	64.97 (12.41)	0.0022
Sex			0.7651
Male	149 (64.22%)	45 (67.16%)	
Female	83 (35.78%)	22 (32.84%)	
Smokers, Yes	78 (33.62%)	18 (26.87%)	0.3710
Anemia, Yes	102 (43.97%)	27 (40.30%)	0.6937
Diabetes, Yes	98 (42.24%)	27 (40.30%)	0.8860
Ejection fraction			0.0025
≤ 30	61 (26.29%)	32 (47.76%)	
31-44	102 (43.97%)	24 (35.82%)	
≥ 45	69 (29.74%)	11 (16.42%)	
Average platelets[c] (SD)	265270.8 (96762.6)	256734.7 (101796.5)	0.5424
Average serum sodium[d] (SD)	137.32 (3.61)	134.21 (5.89)	<0.001

a. Number and proportion are reported for categorical variables. Average and standard deviation (SD) are reported for continuous variables. Chi-squared tests and t-tests were applied to categorical and continuous variables respectively.

b. The unit of serum creatinine is mg/dL.

c. The unit of platelets is kilo platelets/mL.

d. The unit of serum sodium is mEq/L.

- We also drew a survival plot to explore the crude association between mortality rates and serum creatinine level. As shown in figure 4, the survival rates for patients with abnormal level serum creatinine are greater than the rates for patients with normal level serum creatinine.
-

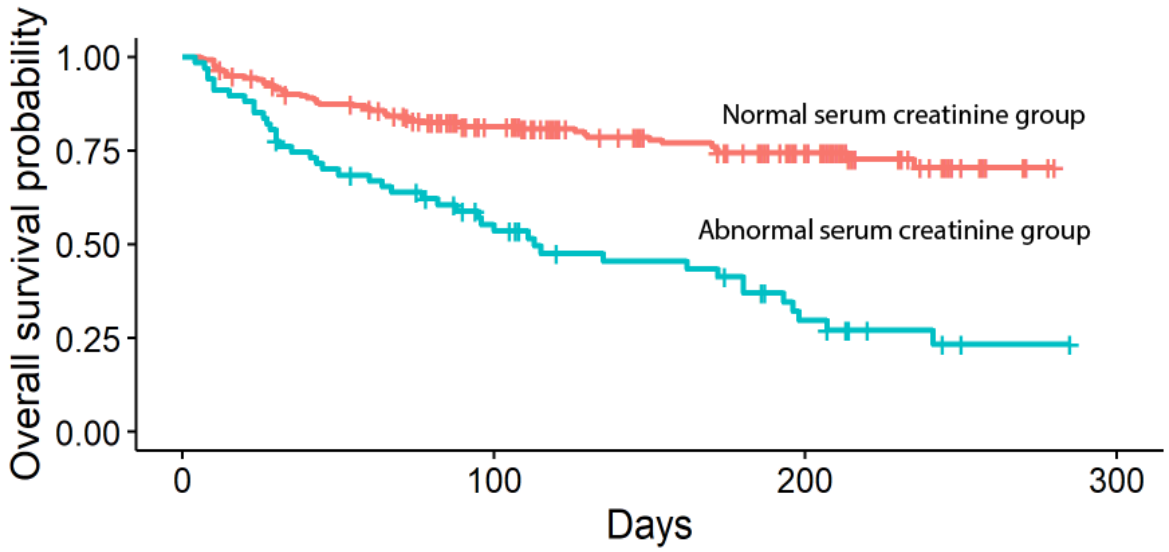


Figure 4. Survival rates of the 299 patients by serum creatinine level.

11. Project Attestation

No member of this group is using these data or same/similar questions in any other course or course project, at HSPH. By listing our name as a group member on our project, and submitting this assignment, we are attesting to this statement above.

Reference

1. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure

patients: A case study. *PLoS One*. 2017;12(7):e0181001. doi:10.1371/JOURNAL.PONE.0181001

2. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Informatics Decis Mak 2020 201*. 2020;20(1):1-16. doi:10.1186/S12911-020-1023-5

3. Metra, M., et al., The role of the kidney in heart failure. European Heart Journal, 2012. **33**(17): p. 2135-2142.

Codes and outputs:

20211010_first_check_in

Jinglun Li; Fuyu Guo; Xinhao Li

10/10/2021

```
library(tidyverse)
library(survival)
library(ggpubr)
library(survminer)
```

```
############################
# load data
dta <- read.csv("heart_f.csv")

# select variables we will use
dta <- dplyr::select(dta,
                     "age", "sex", "anaemia",
                     "diabetes", "ejection_fraction", "smoking",
                     "platelets", "serum_creatinine", "serum_sodium",
                     "time", "DEATH_EVENT")

# rename the variables to make our work easier
names(dta) <- c("age", "sex", "anemia",
                     "dbt", "ef", "smoking",
                     "plat", "ser_crt", "ser_na",
                     "time", "death")

# check sample size
dim(dta)
```

```
## [1] 299  11
```

```
# check if there is any missing value in variables
complete.cases(dta) %>% all()
```

```
## [1] TRUE
```

```
# no missing value



#########################################################
#########################################################
# check exposure distribution
summary(dta$ser_crt)
```

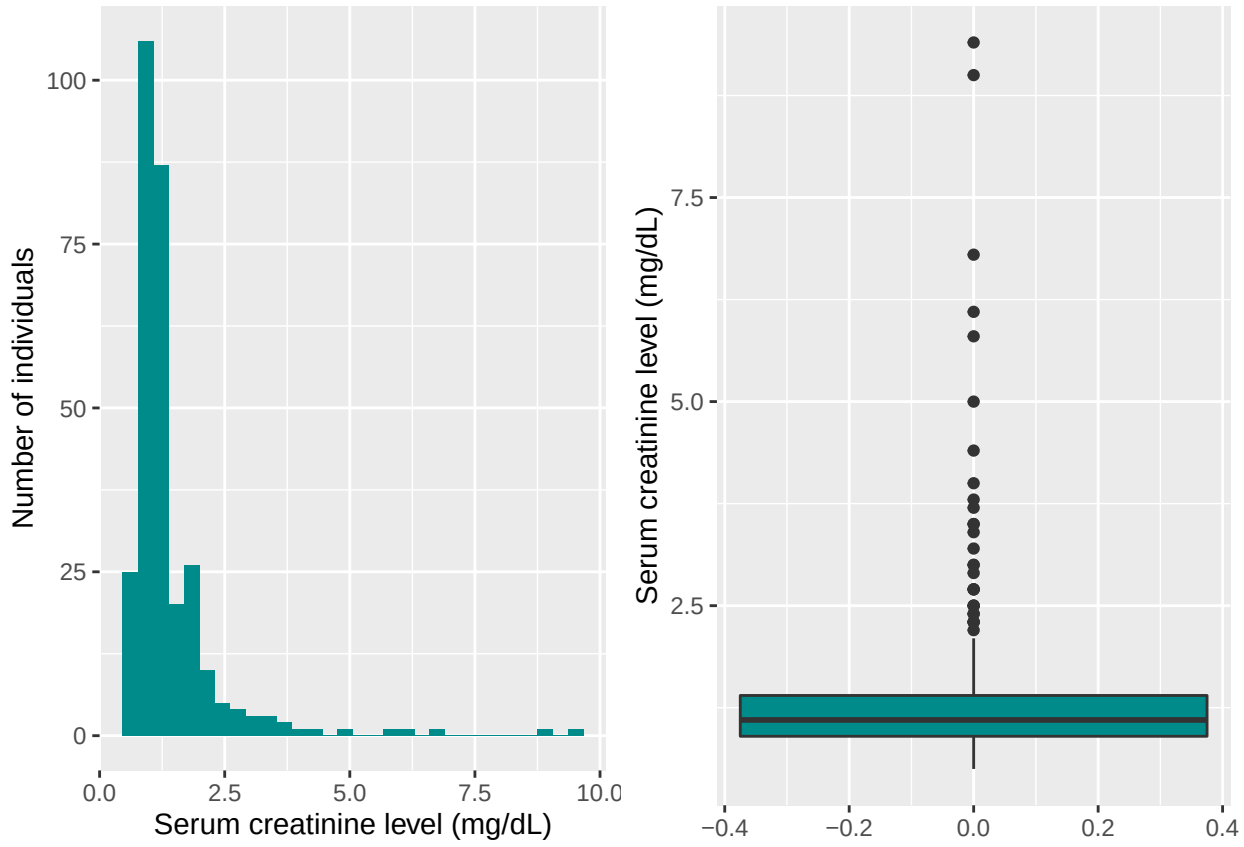
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   0.900   1.100   1.394   1.400   9.400
```

```
sd(dta$ser_crt)
```

```
## [1] 1.03451
```

```
p1 <- ggplot(dta) +
        geom_histogram(aes(x = ser_crt), fill = "DarkCyan") +
    ylab("Number of individuals") +
    xlab("Serum creatinine level (mg/dL)")
p2 <- ggplot(dta) +
        geom_boxplot(aes(y = ser_crt), fill = "DarkCyan") +
    ylab("Serum creatinine level (mg/dL)")
ggarrange(p1, p2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggsave("20211001_exposure_distribution.png", width = 150, height = 80,
       units = "mm")


# calculate the normal level proportion
dta$ser_crt_group <- if_else(dta$ser_crt <= 1.5, "normal", "abnormal")
dta$ser_crt_group %>% table
```

```
## .
## abnormal   normal
##       67      232
```

```
dta$ser_crt_group %>% table %>% prop.table()
```

```
## .
##  abnormal    normal
## 0.2240803 0.7759197
```

```
############################################################
############################################################
# check total person-time
sum(dta$time)
```

```
## [1] 38948
```

```
summary(dta$time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    73.0   115.0   130.3   203.0   285.0
```

```
sd(dta$time)
```

```
## [1] 77.61421
```

```
# check time to the event among patients who died finally
summary(dta$time[dta$death == 1])
```

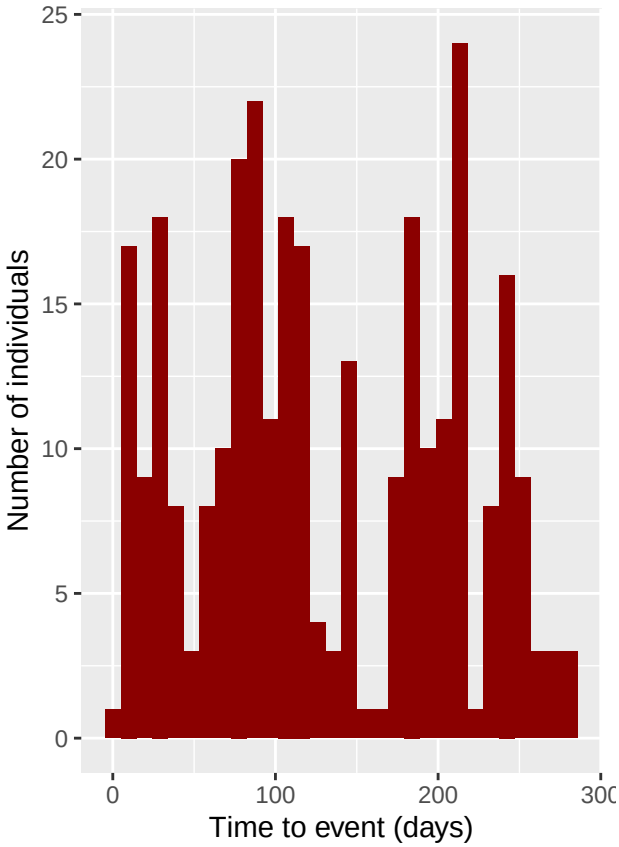
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   25.50   44.50   70.89  102.25  241.00
```

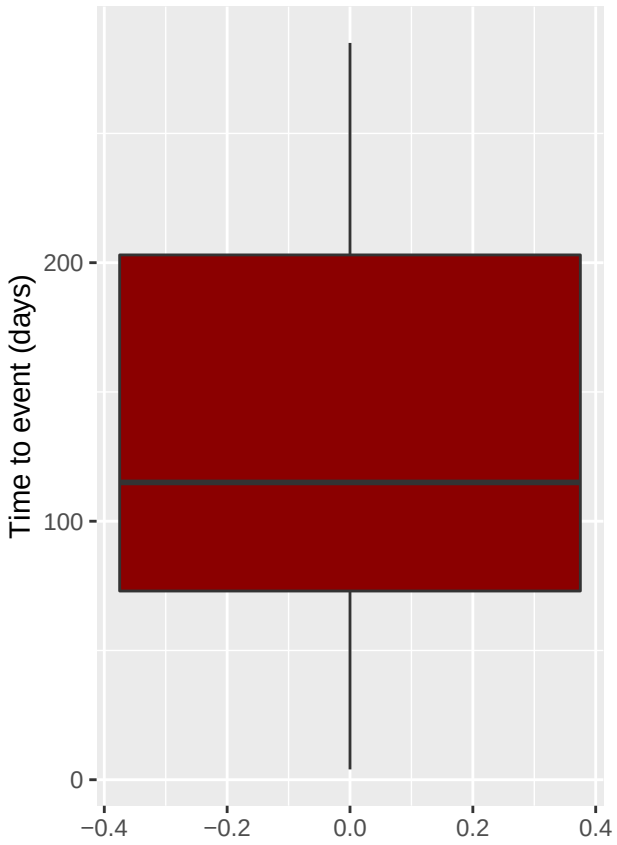
```
sd(dta$time[dta$death == 1])
```

```
## [1] 62.37828
```

```
# make plots
p1 <- ggplot(dta) +
        geom_histogram(aes(x = time), fill = "DarkRed") +
    xlab("Time to event (days)") +
    ylab("Number of individuals")
p2 <- ggplot(dta) +
        geom_boxplot(aes(y = time), fill = "DarkRed") +
    ylab("Time to event (days)")
ggarrange(p1, p2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

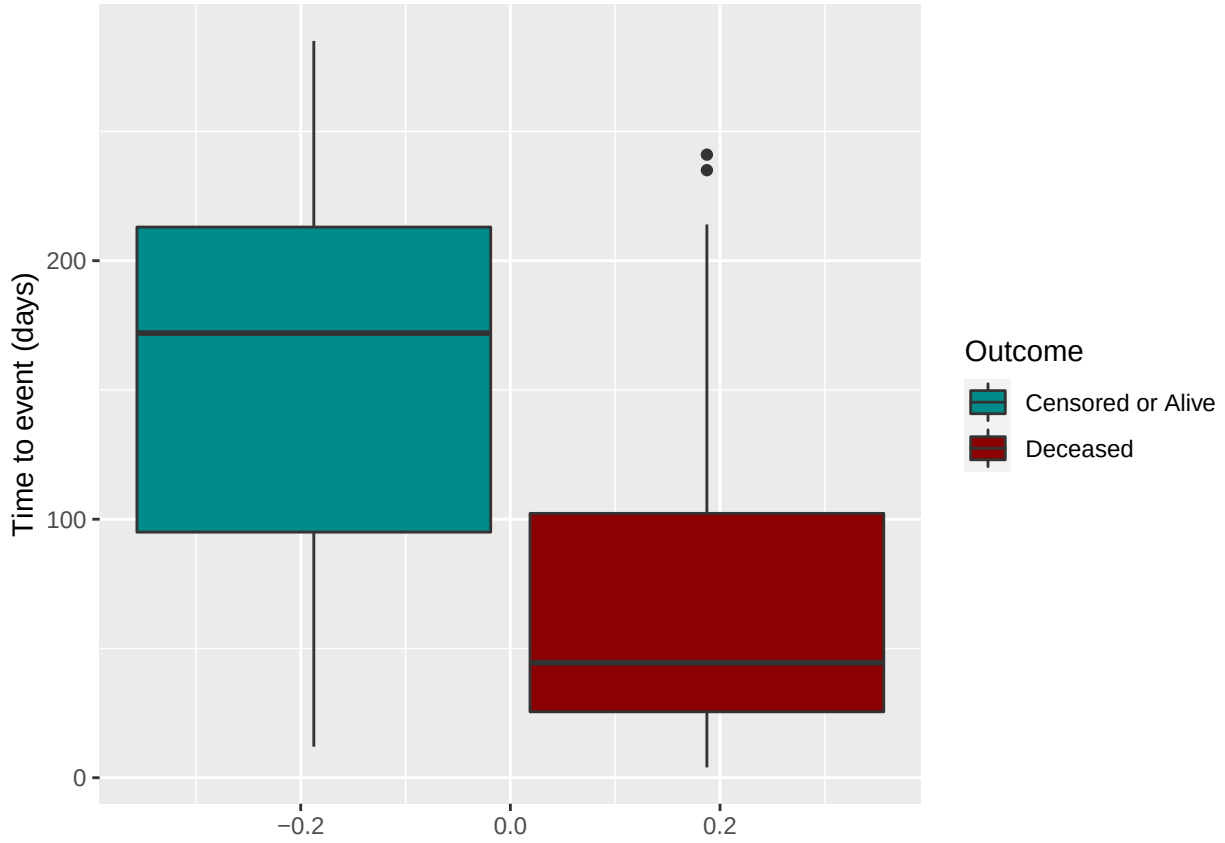





```
ggsave("20211001_death_distribution.png", width = 150, height = 80,
       units = "mm")

# make a plot by death = 0

dta1<-dta %>% mutate(Outcome = case_when(death=="0"~"Censored or Alive",death=="1"~"Deceased"))
ggplot(dta1) +
        geom_boxplot(aes(y = time, group = death, fill = Outcome)) +
        scale_fill_manual(values = c("DarkCyan", "DarkRed")) +
    ylab("Time to event (days)")
```




```
ggsave("20211001_death_distribution_by_death.png", width = 150, height = 80,
       units = "mm")



# calculate the proportion of death in 30 days
dta$death_30 <- NA
dta$death_30[dta$death == 1& dta$time <= 30] <- "Yes"
dta$death_30[(dta$death == 1& dta$time > 30) |
               (dta$death == 0& dta$time > 30)] <- "No"
dta$death_30[dta$death == 0& dta$time <= 30] <- "Censored"
table(dta$death_30)
```

```
##
## Censored       No      Yes
##        5      259       35
```

```
table(dta$death_30) %>% prop.table()
```

```
##
##   Censored         No        Yes
## 0.01672241 0.86622074 0.11705686
```

```
##############################################
##############################################
# Check Characteristics of the 299 patient by serum creatinine level
#  Table 1 #
# check continuous variables
re1 <- dta %>% group_by(ser_crt_group) %>%
        summarise(avg_crt = mean(ser_crt),
                  sd_crt = sd(ser_crt),
                  avg_time = mean(time),
                  sd_time = sd(time),
                  avg_age = mean(age),
                  sd_age = sd(age),
                  avg_plat = mean(plat),
                  sd_plat = sd(plat),
                  avg_na = mean(ser_na),
                  sd_na = sd(ser_na)) %>%
        t()


re1 <- re1[, c(2,1)]
re1
```

```
##               [,1]        [,2]
## ser_crt_group "normal"    "abnormal"
## avg_crt       "1.025991"  "2.667761"
## sd_crt        "0.2071469" "1.5996475"
## avg_time      "136.3362"  "109.2239"
## sd_time       "75.80768"  "80.66162"
## avg_age       "59.63937"  "64.97015"
## sd_age        "11.49359"  "12.41330"
## avg_plat      "265270.8"  "256734.7"
## sd_plat       " 96762.64" "101796.45"
## avg_na        "137.3233"  "134.2090"
## sd_na         "3.609004"  "5.889223"
```

```
# t-test for continuous variables
lapply(c("ser_crt", "time", "age", "plat", "ser_na"),
       function (x){
               dta$x <- dta[,x]
               t.test(x~ser_crt_group, data = dta)$p.value %>% return()
       })
```

```
## [[1]]
## [1] 5.15281e-12
##
## [[2]]
## [1] 0.01574423
##
## [[3]]
## [1] 0.002168052
##
## [[4]]
## [1] 0.5423805
```

```
##
## [[5]]
## [1] 9.389263e-05
```

```
#################################################################
# categorical variables

# rename the categories
dta$ser_crt_group <- factor(dta$ser_crt_group , levels = c("normal", "abnormal"))
dta$sex <- if_else(dta$sex == 1, "male", "female")
dta$ef[dta$ef <= 30] <- "<=30"
dta$ef[dta$ef > 30 & dta$ef < 45] <- "41-44"
dta$ef[dta$ef >= 45] <- ">=45"


# check the proportion
lapply(c("death", "death_30", "sex", "smoking", "anemia", "dbt",
         "ef"),
       function(x){
        s1 <- table(dta[,x], dta$ser_crt_group)
        s2 <- table(dta[,x], dta$ser_crt_group) %>% as.matrix()
        s2 <- cbind(s2[,1]/sum(dta$ser_crt_group == "normal"),  s2[,1]/sum(dta$ser_crt_group == "abnorma
        s3 <- table(dta[,x], dta$ser_crt_group) %>% chisq.test()
        list(s1, s2, s3) %>% return()
         })
```

```
## Warning in chisq.test(.): Chi-squared approximation may be incorrect
```

```
## [[1]]
## [[1]][[1]]
##
##     normal abnormal
##   0    179       24
##   1     53       43
##
## [[1]][[2]]
##         [,1]      [,2]
## 0 0.7715517 2.6716418
## 1 0.2284483 0.7910448
##
## [[1]][[3]]
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 38.872, df = 1, p-value = 4.525e-10
##
##
##
## [[2]]
## [[2]][[1]]
##
##            normal abnormal
```

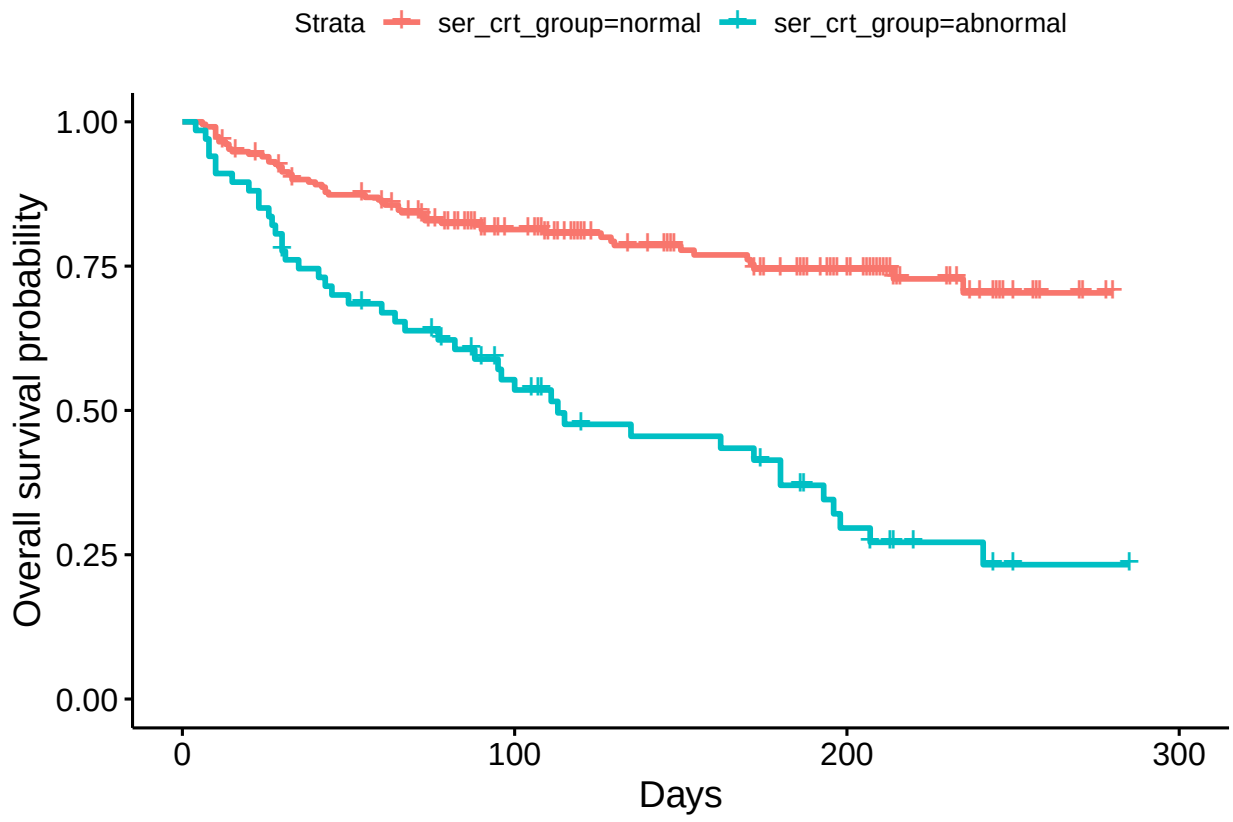
```
##   Censored      4        1
##   No          208       51
##   Yes          20       15
##
## [[2]][[2]]
##                [,1]       [,2]
## Censored 0.01724138 0.05970149
## No       0.89655172 3.10447761
## Yes      0.08620690 0.29850746
##
## [[2]][[3]]
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 9.534, df = 2, p-value = 0.008506
##
##
##
## [[3]]
## [[3]][[1]]
##
##          normal abnormal
##   female     83       22
##   male      149       45
##
## [[3]][[2]]
##              [,1]     [,2]
## female 0.3577586 1.238806
## male   0.6422414 2.223881
##
## [[3]][[3]]
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 0.089291, df = 1, p-value = 0.7651
##
##
##
## [[4]]
## [[4]][[1]]
##
##     normal abnormal
##   0    154       49
##   1     78       18
##
## [[4]][[2]]
##        [,1]     [,2]
## 0 0.6637931 2.298507
## 1 0.3362069 1.164179
##
## [[4]][[3]]
##
```

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 0.8004, df = 1, p-value = 0.371
##
##
##
## [[5]]
## [[5]][[1]]
##
##     normal abnormal
##   0    130       40
##   1    102       27
##
## [[5]][[2]]
##        [,1]     [,2]
## 0 0.5603448 1.940299
## 1 0.4396552 1.522388
##
## [[5]][[3]]
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 0.1551, df = 1, p-value = 0.6937
##
##
##
## [[6]]
## [[6]][[1]]
##
##     normal abnormal
##   0    134       40
##   1     98       27
##
## [[6]][[2]]
##        [,1]     [,2]
## 0 0.5775862 2.000000
## 1 0.4224138 1.462687
##
## [[6]][[3]]
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 0.020568, df = 1, p-value = 0.886
##
##
##
## [[7]]
## [[7]][[1]]
##
##         normal abnormal
##   <=30      61       32
```

```
##   >=45      69       11
##   41-44    102       24
##
## [[7]][[2]]
##            [,1]      [,2]
## <=30  0.2629310 0.9104478
## >=45  0.2974138 1.0298507
## 41-44 0.4396552 1.5223881
##
## [[7]][[3]]
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 11.971, df = 2, p-value = 0.002515
```

```
#####################################################################
#############################################################
# survival plots for crude association

ggsurvplot(
    fit = survfit(Surv(time, death) ~ ser_crt_group, data = dta,),
    xlab = "Days",
    ylab = "Overall survival probability")
```



```r
ggsave("20211001_survival_plots_crude.png", width = 150, height = 80,
       units = "mm")
```