Exploratory Data Analysis

(EDA)

Núria Jolis Orriols

2 de juny, 2022

Contents

Step 1 - Obtaining the data

Dataset description and modification

```
## [1] 1176   51
```

```
## 'data.frame':    1176 obs. of  51 variables:
##  $ group                    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ID                       : int  125047 139812 109787 130587 138290 154653 194420 153461 113076 147
##  $ outcome                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age                      : int  72 75 83 43 75 76 72 83 61 67 ...
##  $ gender                   : int  1 2 2 2 2 1 1 2 2 1 ...
##  $ BMI                      : num  37.6 NA 26.6 83.3 31.8 ...
##  $ hypertensive             : int  0 0 0 0 1 1 1 1 1 1 ...
##  $ atrialfibrillation       : int  0 0 0 0 0 1 0 1 1 0 ...
##  $ CHD.with.no.MI           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ diabetes                 : int  1 0 0 0 0 0 0 1 1 1 ...
##  $ deficiencyanemias        : int  1 1 1 0 1 1 0 1 0 0 ...
##  $ depression               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hyperlipemia             : int  1 0 0 0 0 1 1 0 0 0 ...
##  $ renal.failure            : int  1 0 1 0 1 1 1 0 1 0 ...
##  $ COPD                     : int  0 1 0 0 1 1 1 0 0 0 ...
##  $ heart.rate               : num  68.8 101.4 72.3 94.5 67.9 ...
##  $ systolic.blood.pressure  : num  156 140 135 126 157 ...
##  $ diastolic.blood.pressure: num  68.3 65 61.4 73.2 58.1 ...
##  $ respiratory.rate         : num  16.6 20.9 23.6 21.9 21.4 ...
##  $ temperature              : num  36.7 36.7 36.5 36.3 36.8 ...
##  $ SP.O2                    : num  98.4 96.9 95.3 93.8 99.3 ...
##  $ urine.output             : num  2155 1425 2425 8760 4455 ...
##  $ hematocrit               : num  26.3 30.8 27.7 36.6 29.9 ...
##  $ RBC                      : num  2.96 3.14 2.62 4.28 3.29 ...
##  $ MCH                      : num  28.2 31.1 34.3 26.1 30.7 ...
##  $ MCHC                     : num  31.5 31.7 31.3 30.4 33.7 ...
##  $ MCV                      : num  89.9 98.2 109.8 85.6 91 ...
##  $ RDW                      : num  16.2 14.3 23.8 17 16.3 ...
##  $ leucocyte                : num  7.65 12.74 5.48 8.22 8.83 ...
##  $ platelets                : num  305 246 204 216 251 ...
##  $ neutrophils              : num  74.7 NA 68.1 81.8 NA ...
##  $ basophils                : num  0.4 NA 0.55 0.15 NA 0.3 0.2 NA 0.55 NA ...
##  $ lymphocyte               : num  13.3 NA 24.5 14.5 NA ...
##  $ PT                       : num  10.6 NA 11.3 27.1 NA ...
##  $ INR                      : num  1 NA 0.95 2.67 NA ...
##  $ NT.proBNP                : num  1956 2384 4081 668 30802 ...
##  $ creatine.kinase          : num  148 60.6 16 85 111.7 ...
##  $ creatinine               : num  1.958 1.122 1.871 0.586 1.95 ...
##  $ urea.nitrogen            : num  50 20.3 33.9 15.3 43 ...
##  $ glucose                  : num  115 148 149 128 146 ...
##  $ blood.potassium          : num  4.82 4.45 5.83 4.39 4.78 ...
##  $ blood.sodium             : num  139 139 141 138 137 ...
##  $ blood.calcium            : num  7.46 8.16 8.27 9.48 8.73 ...
##  $ chloride                 : num  109.2 98.4 105.9 92.1 104.5 ...
##  $ anion.gap                : num  13.2 11.4 10 12.4 15.2 ...
##  $ magnesium.ion            : num  2.62 1.89 2.16 1.94 1.65 ...
##  $ pH                       : num  7.23 7.22 7.27 7.37 7.25 ...
##  $ bicarbonate              : num  21.2 33.4 30.6 38.6 22 ...
##  $ lactic.acid              : num  0.5 0.5 0.5 0.6 0.6 ...
##  $ PCO2                     : num  40 78 71.5 75 50 ...
##  $ EF                       : int  55 55 35 55 55 35 55 75 50 55 ...
```

The dataset consists of 1176 observations and 51 variables, all numeric being 36 of type "double" and 15 of

type “integer”.

Dataset modifications:

1. The first two variables are to be discarded. The variable, “group”, is removed as it was created by Li 2021 to separate the data for training and testing their models, and the ID is the patient’s identification which will not be useful to predict the outcome.
2. A new variable is created to combine systolic and diastolic blood pressure. It is called mean arterial pressure (MAP) and it follows the next equation: MAP = [Systolic + 2*Diastolic]/3.
3. 11 of the variables are numerical binary, they will be converted into factors: “outcome”, “gender”, “hypertensive”, “atrialfibrillation”, “CHD.with.no.MI”, “diabetes”, “deficiencyanemias”, “depression”, “hyperlipemia”, “renal.failure” and “COPD”.

```
## 'data.frame':    1176 obs. of  48 variables:
##  $ outcome           : Factor w/ 2 levels "Survivor","Non-survivor": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age               : int  72 75 83 43 75 76 72 83 61 67 ...
##  $ gender            : Factor w/ 2 levels "M","F": 1 2 2 2 2 1 1 2 2 1 ...
##  $ BMI               : num  37.6 NA 26.6 83.3 31.8 ...
##  $ hypertensive      : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 2 2 2 ...
##  $ atrialfibrillation: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 2 1 ...
##  $ CHD.with.no.MI    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ diabetes          : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 2 2 2 ...
##  $ deficiencyanemias : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 1 2 1 1 ...
##  $ depression        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ hyperlipemia      : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 2 1 1 1 ...
##  $ renal.failure     : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
##  $ COPD              : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 2 2 1 1 1 ...
##  $ heart.rate        : num  68.8 101.4 72.3 94.5 67.9 ...
##  $ respiratory.rate  : num  16.6 20.9 23.6 21.9 21.4 ...
##  $ temperature       : num  36.7 36.7 36.5 36.3 36.8 ...
##  $ SP.O2             : num  98.4 96.9 95.3 93.8 99.3 ...
##  $ urine.output      : num  2155 1425 2425 8760 4455 ...
##  $ hematocrit        : num  26.3 30.8 27.7 36.6 29.9 ...
##  $ RBC               : num  2.96 3.14 2.62 4.28 3.29 ...
##  $ MCH               : num  28.2 31.1 34.3 26.1 30.7 ...
##  $ MCHC              : num  31.5 31.7 31.3 30.4 33.7 ...
##  $ MCV               : num  89.9 98.2 109.8 85.6 91 ...
##  $ RDW               : num  16.2 14.3 23.8 17 16.3 ...
##  $ leucocyte         : num  7.65 12.74 5.48 8.22 8.83 ...
##  $ platelets         : num  305 246 204 216 251 ...
##  $ neutrophils       : num  74.7 NA 68.1 81.8 NA ...
##  $ basophils         : num  0.4 NA 0.55 0.15 NA 0.3 0.2 NA 0.55 NA ...
##  $ lymphocyte        : num  13.3 NA 24.5 14.5 NA ...
##  $ PT                : num  10.6 NA 11.3 27.1 NA ...
##  $ INR               : num  1 NA 0.95 2.67 NA ...
##  $ NT.proBNP         : num  1956 2384 4081 668 30802 ...
##  $ creatine.kinase   : num  148 60.6 16 85 111.7 ...
##  $ creatinine        : num  1.958 1.122 1.871 0.586 1.95 ...
##  $ urea.nitrogen     : num  50 20.3 33.9 15.3 43 ...
##  $ glucose           : num  115 148 149 128 146 ...
##  $ blood.potassium   : num  4.82 4.45 5.83 4.39 4.78 ...
##  $ blood.sodium      : num  139 139 141 138 137 ...
##  $ blood.calcium     : num  7.46 8.16 8.27 9.48 8.73 ...
##  $ chloride          : num  109.2 98.4 105.9 92.1 104.5 ...
```

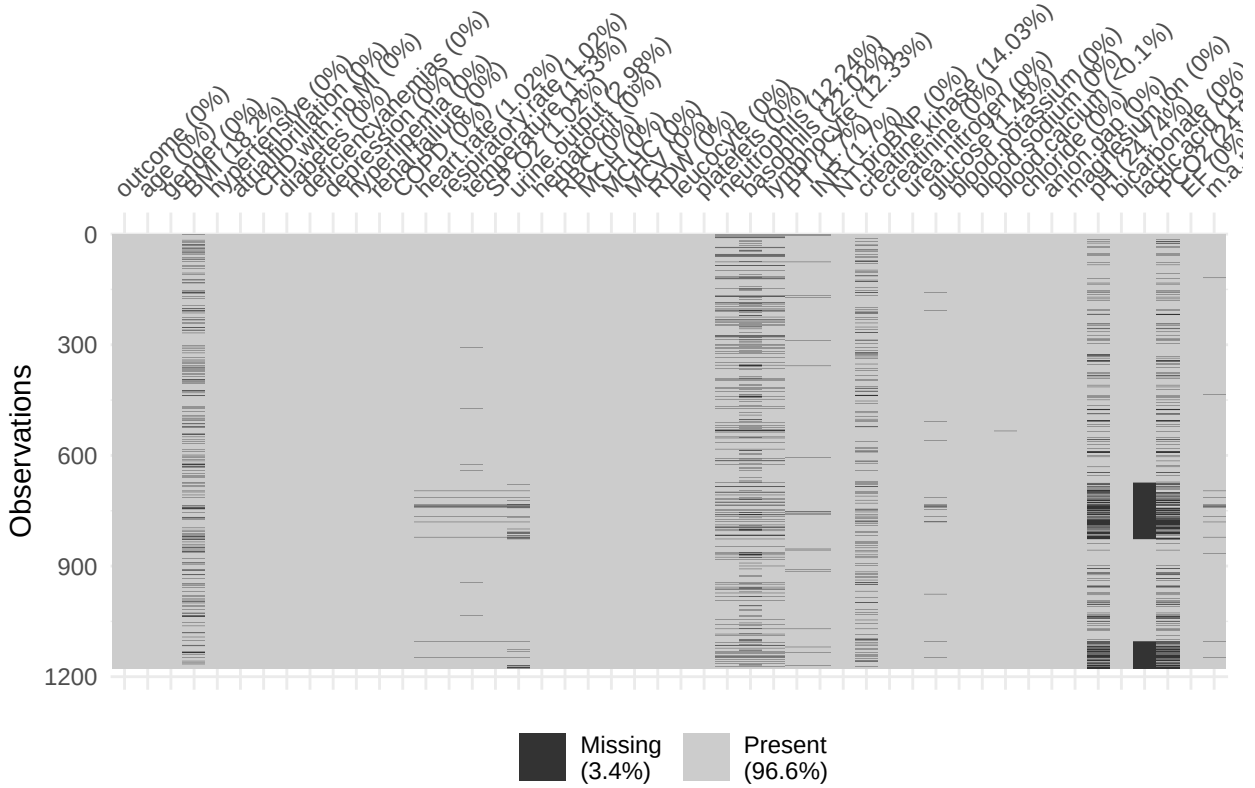
```
##  $ anion.gap         : num  13.2 11.4 10 12.4 15.2 ...
##  $ magnesium.ion     : num  2.62 1.89 2.16 1.94 1.65 ...
##  $ pH                : num  7.23 7.22 7.27 7.37 7.25 ...
##  $ bicarbonate       : num  21.2 33.4 30.6 38.6 22 ...
##  $ lactic.acid       : num  0.5 0.5 0.5 0.6 0.6 ...
##  $ PCO2              : num  40 78 71.5 75 50 ...
##  $ EF                : int  55 55 35 55 55 35 55 75 50 55 ...
##  $ m.a.p             : num  97.5 90 86 90.9 90.9 ...
```

Finally, we got a dataset of 1176 observations and 48 variables: 39 numeric and 11 factors. The outcome is the response variable whose behavior shall be modeled and the 47 variables left are considered to be candidate predictors.

Checking missing values

```
## [1] 1901
```

```
## [1] 3.4
```




```
##             BMI      heart.rate respiratory.rate     temperature           SP.O2
##            18.2             1.0              1.0             1.5             1.0
##  urine.output      neutrophils        basophils      lymphocyte              PT
##             3.0            12.2             22.0            12.3             1.7
##             INR  creatine.kinase          glucose   blood.calcium              pH
##             1.7            14.0              1.4             0.1            24.7
##     lactic.acid             PCO2            m.a.p
##            19.4            24.9              1.3
```

The 3.4% of the values are missing and are concentrated in 18 of the 48 variables. Of those 18 variables, 8 of them present more than 10% of missing values: “basophils”, “creatine.kinase”, “lactic.acid”, “BMI”, “neutrophils”, “lymphocyte”, “pH” and “PCO2”.

Step 2 - Variables description

Qualitative variables

Quantitative variables

Table 1: Descriptive table for quantitative variables

Variable	N	Mean	Median	Min	Max	NAs	%NAs
BMI	962	30.19	13.35	28.31	104.97	214	18.2
age	1176	74.05	19.00	77.00	99.00	0	0.0
heart.rate	1164	84.58	36.00	83.61	135.71	12	1.0
respiratory.rate	1164	20.80	11.14	20.37	40.90	12	1.0
temperature	1158	36.68	33.25	36.65	39.13	18	1.5
SP.O2	1164	96.27	75.92	96.45	100.00	12	1.0
urine.output	1141	1899.28	0.00	1675.00	8820.00	35	3.0
hematocrit	1176	31.91	20.31	30.80	55.42	0	0.0
RBC	1176	3.57	2.03	3.49	6.58	0	0.0
MCH	1176	29.54	18.12	29.75	40.31	0	0.0
MCHC	1176	32.86	27.82	32.99	37.01	0	0.0
MCV	1176	89.90	62.60	90.00	116.71	0	0.0
RDW	1176	15.95	12.09	15.51	29.05	0	0.0
leucocyte	1176	10.72	0.10	9.68	64.75	0	0.0
platelets	1176	241.52	9.57	222.67	1028.20	0	0.0
neutrophils	1032	80.12	5.00	82.47	98.00	144	12.2
basophils	917	6.23	0.10	0.30	675.00	259	22.0
lymphocyte	1031	12.23	0.97	10.47	83.50	145	12.3
PT	1156	17.49	10.10	14.64	71.27	20	1.7
INR	1156	4.07	0.87	1.30	975.00	20	1.7
NT.proBNP	1176	11011.04	50.00	5837.75	118928.00	0	0.0
creatine.kinase	1011	246.94	8.00	89.50	42987.50	165	14.0
creatinine	1176	16.00	0.27	1.33	975.00	0	0.0
urea.nitrogen	1176	36.29	5.36	30.61	161.75	0	0.0
glucose	1159	148.80	66.67	136.40	414.10	17	1.4
blood.potassium	1176	4.18	3.00	4.11	6.57	0	0.0
blood.sodium	1176	138.90	114.67	139.25	154.74	0	0.0
blood.calcium	1175	8.50	6.70	8.50	10.95	1	0.1
chloride	1176	102.29	80.27	102.52	122.53	0	0.0
anion.gap	1176	13.92	6.64	13.67	25.50	0	0.0
magnesium.ion	1176	2.12	1.40	2.09	4.07	0	0.0
pH	885	7.38	7.09	7.38	7.58	291	24.7
bicarbonate	1176	26.91	12.86	26.50	47.67	0	0.0
lactic.acid	948	8.36	0.50	1.62	975.00	228	19.4
PCO2	883	45.54	18.75	43.00	98.60	293	24.9
EF	1176	48.71	15.00	55.00	75.00	0	0.0
m.a.p	1161	79.02	51.16	77.30	129.01	15	1.3

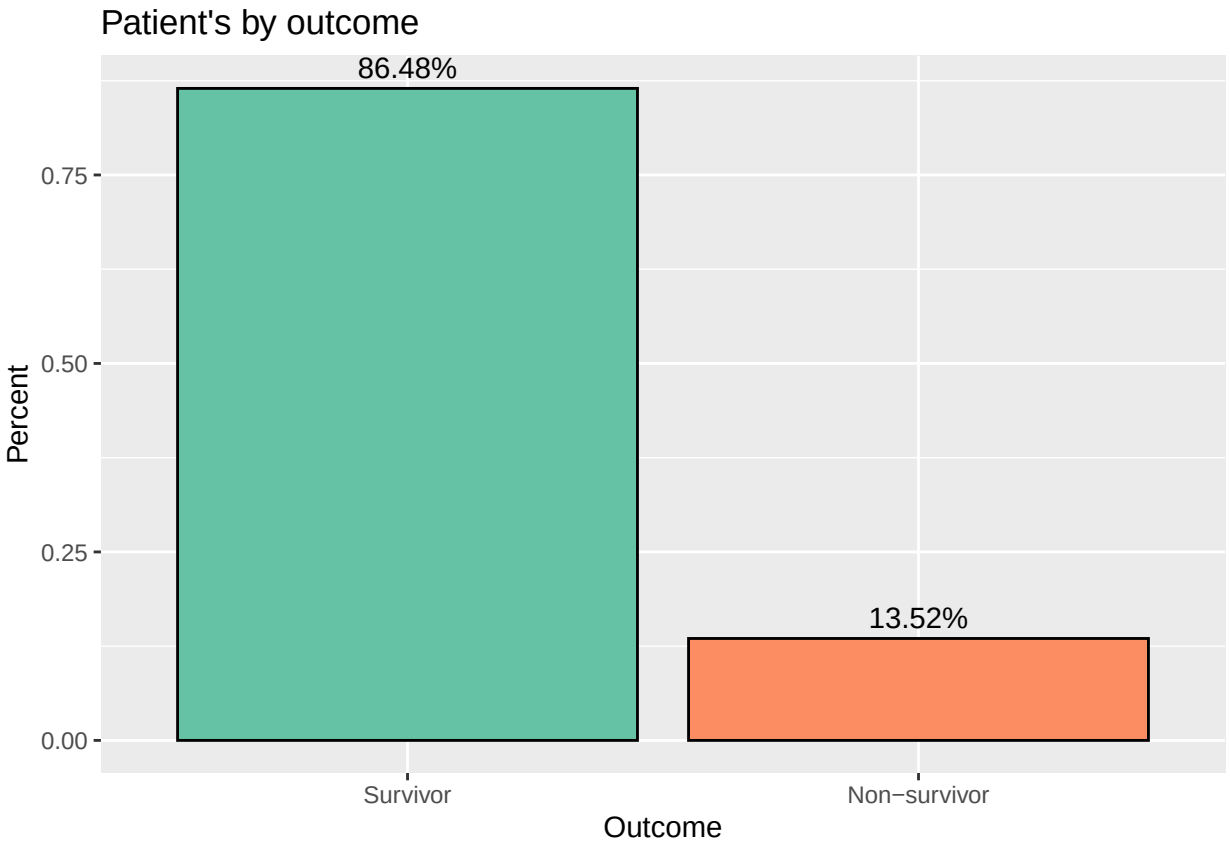
Step 3 - Data visualization

The 48 variables can be divided into 5 groups:

- Primary variable: outcome.
- Demographic features: age, gender and BMI.
- Vital signs: heart rate, m.a.p, respiratory.rate,temperature, SPO2 and urine output.
- Comorbidities: hypertension, atrial fibrillation, CHD.with.no.MI, diabetes, depression, deficiencyanemias, hiperlipiaemia, renal failure and COPD.
- Laboratory variables: the rest.

Exploring the primary variable

The primary response is the binary variable outcome (Survivor, Non-survivor) defined as the vital status at the time of hospital discharge.



As the graphic shows, at the end of the study, 86,48% (1017) of the patients survived whereas the 13,52% left (159) died. It can be concluded that the data is very imbalanced but this is normal in these types of studies.

Exploring demographic features

Among the demographic features there are two numeric variables (age and BMI) and one factor (gender).

```
##       age             BMI
```

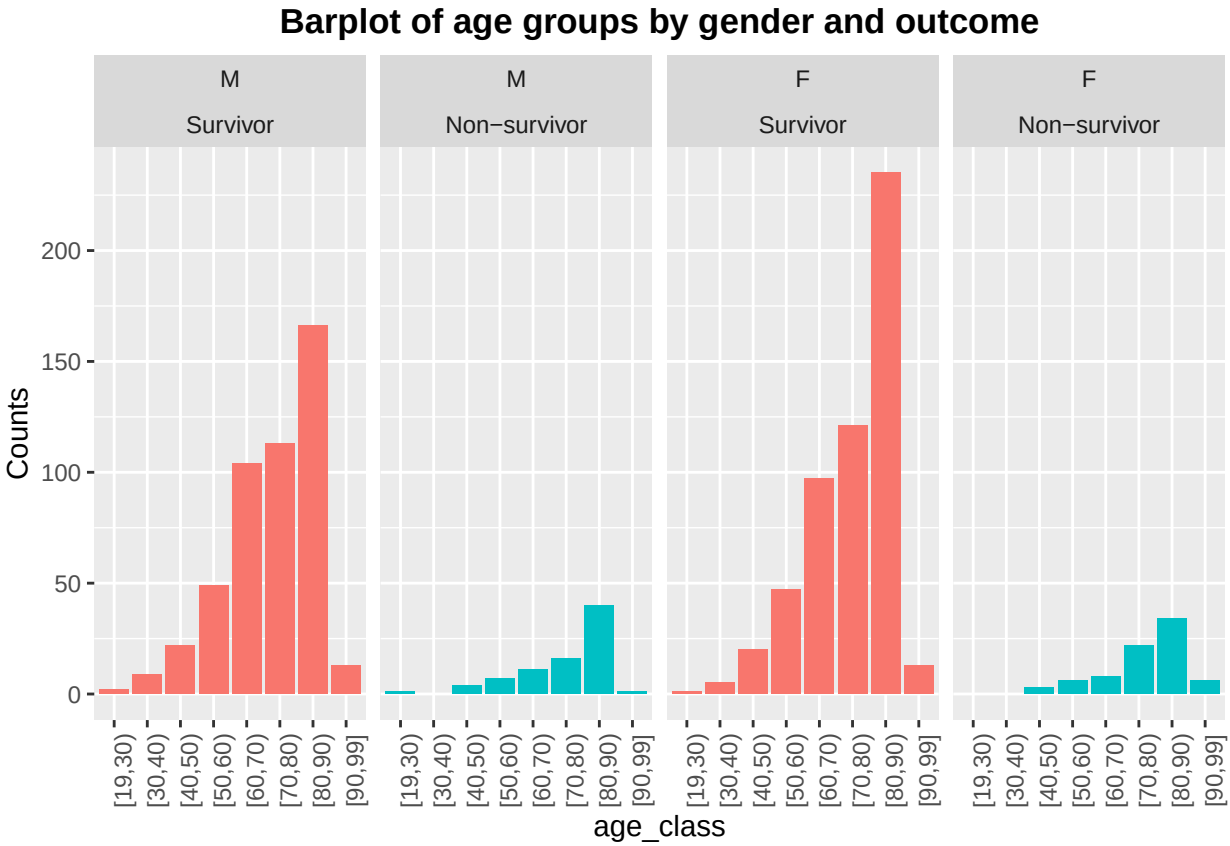
```
##  Min.   :19.00   Min.   : 13.35
##  1st Qu.:65.00   1st Qu.: 24.33
##  Median :77.00   Median : 28.31
##  Mean   :74.05   Mean   : 30.19
##  3rd Qu.:85.00   3rd Qu.: 33.63
##  Max.   :99.00   Max.   :104.97
##                  NA's   :214
```

```
##
##     M     F
## 47.45 52.55
```

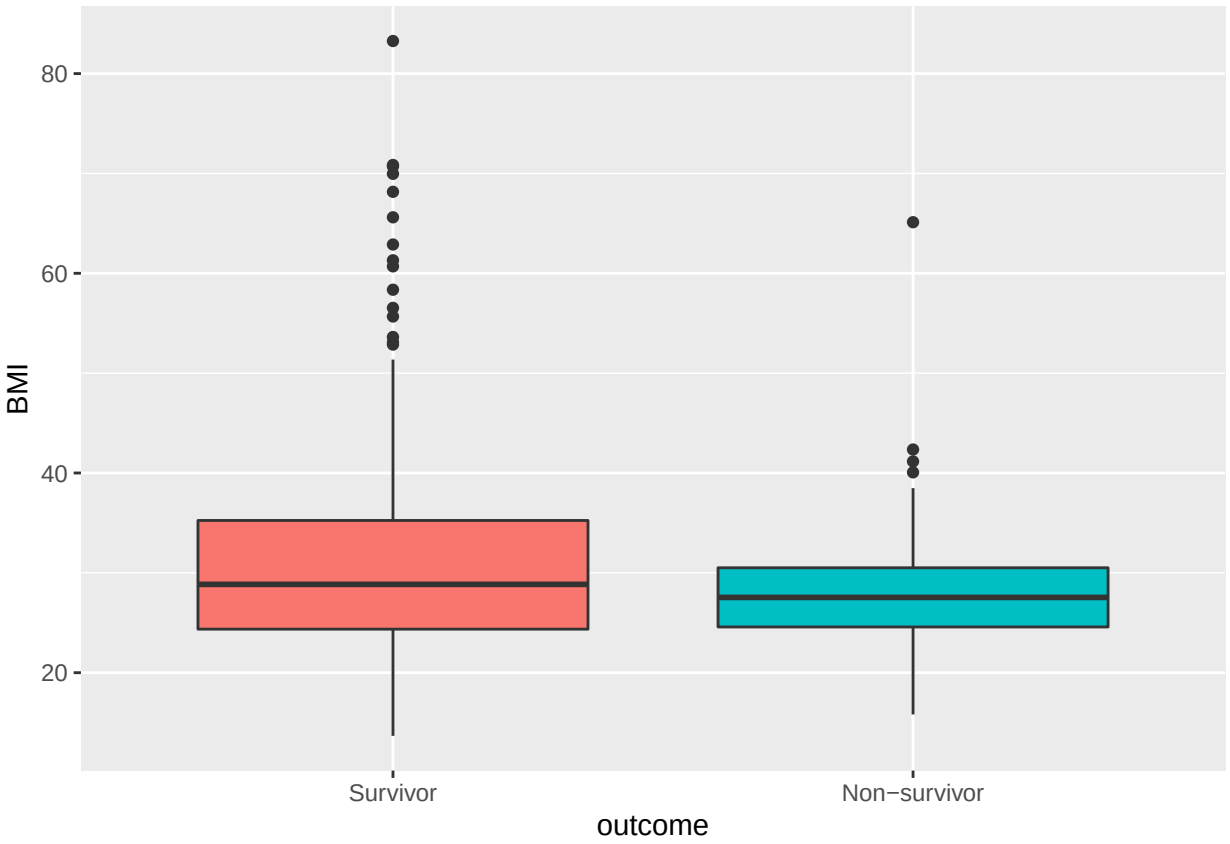
```
##
##   No  Yes
## 1087   89
```

The age of the patients in this study ranges from 19 to 99 and the BMI ranges from 13,35 to 104,95. There are 214 missing values of the BMI feature.

Regarding gender, 47,45% of the subjects are males and 52,55% are females. In this case, it can be said that the data is balanced.



The graphics show that the distribution of the age group bars between males and females is very similar; also among survivors and non-survivors. Among the survivors, the incidence is higher between 60 and 90 years old independently of the gender. Among the group of non-survivors, the incidence is higher between 80 and 90 years old. The patient whose outcome is unknown is a male between 80 and 90 years old.

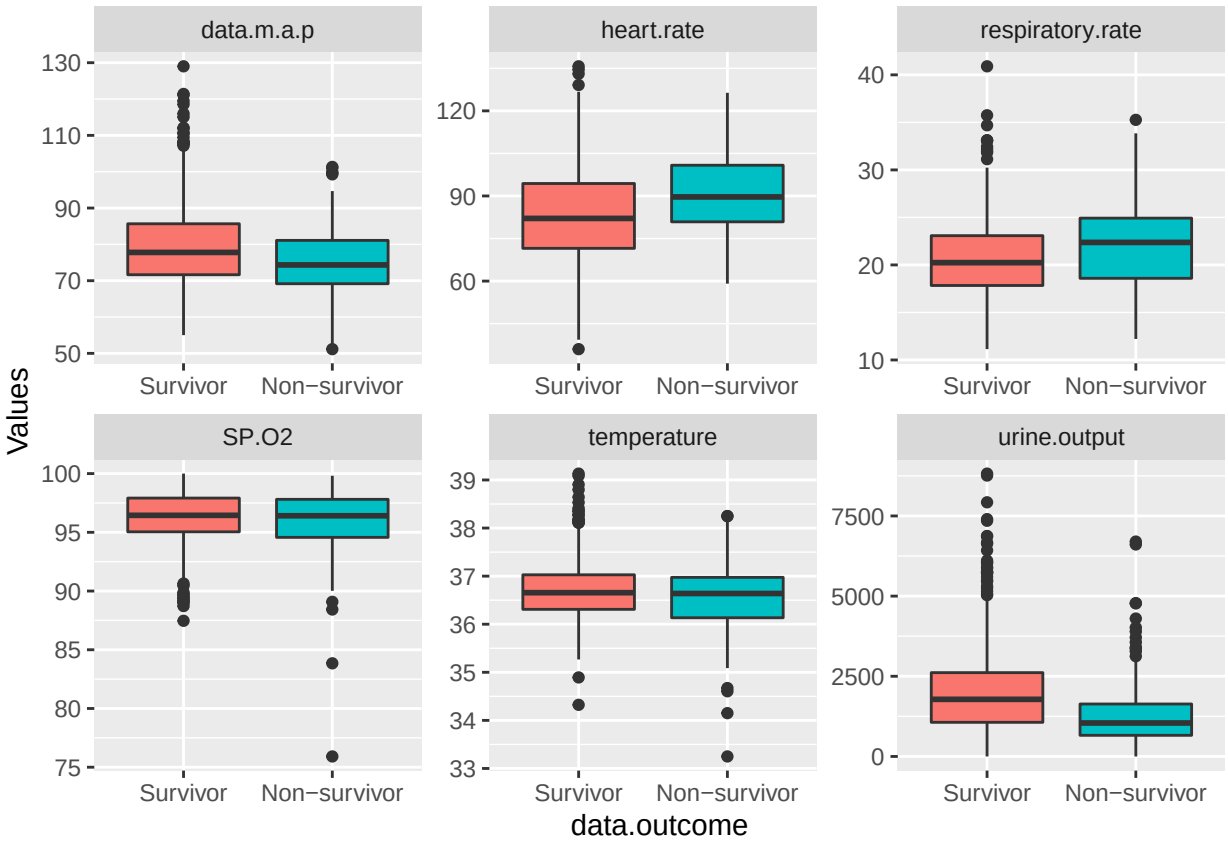


The BMI's median is 28.31. It is slightly higher for survivors than non-survivors. In the first plot, we observe that the median of the two groups seems similar but there are a few outliers that make the group survivor a little bit right-skewed.

Exploring vital signs

```
##    heart.rate     respiratory.rate  temperature        SP.O2         urine.output
##  Min.   : 36.00   Min.   :11.14    Min.   :33.25   Min.   : 75.92   Min.   :   0
##  1st Qu.: 72.37   1st Qu.:17.93    1st Qu.:36.29   1st Qu.: 95.00   1st Qu.: 980
##  Median : 83.61   Median :20.37    Median :36.65   Median : 96.45   Median :1675
##  Mean   : 84.58   Mean   :20.80    Mean   :36.68   Mean   : 96.27   Mean   :1899
##  3rd Qu.: 95.91   3rd Qu.:23.39    3rd Qu.:37.02   3rd Qu.: 97.92   3rd Qu.:2500
##  Max.   :135.71   Max.   :40.90    Max.   :39.13   Max.   :100.00   Max.   :8820
##  NA's   :12       NA's   :12       NA's   :18      NA's   :12       NA's   :35
##    data.m.a.p
##  Min.   : 51.16
##  1st Qu.: 71.37
##  Median : 77.30
##  Mean   : 79.02
##  3rd Qu.: 85.18
##  Max.   :129.01
##  NA's   :15
```

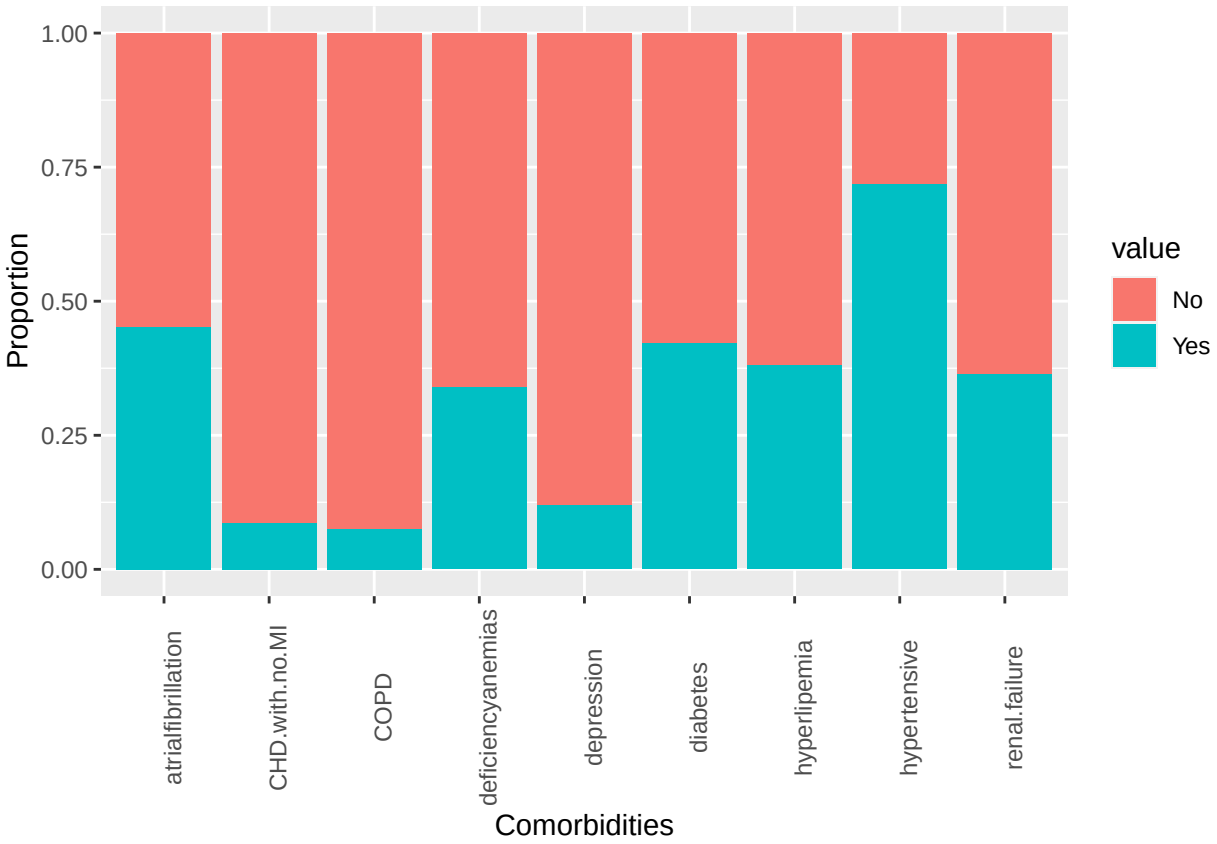
All vital signs present from 12 to 18 missing values except urine.output that has 35.



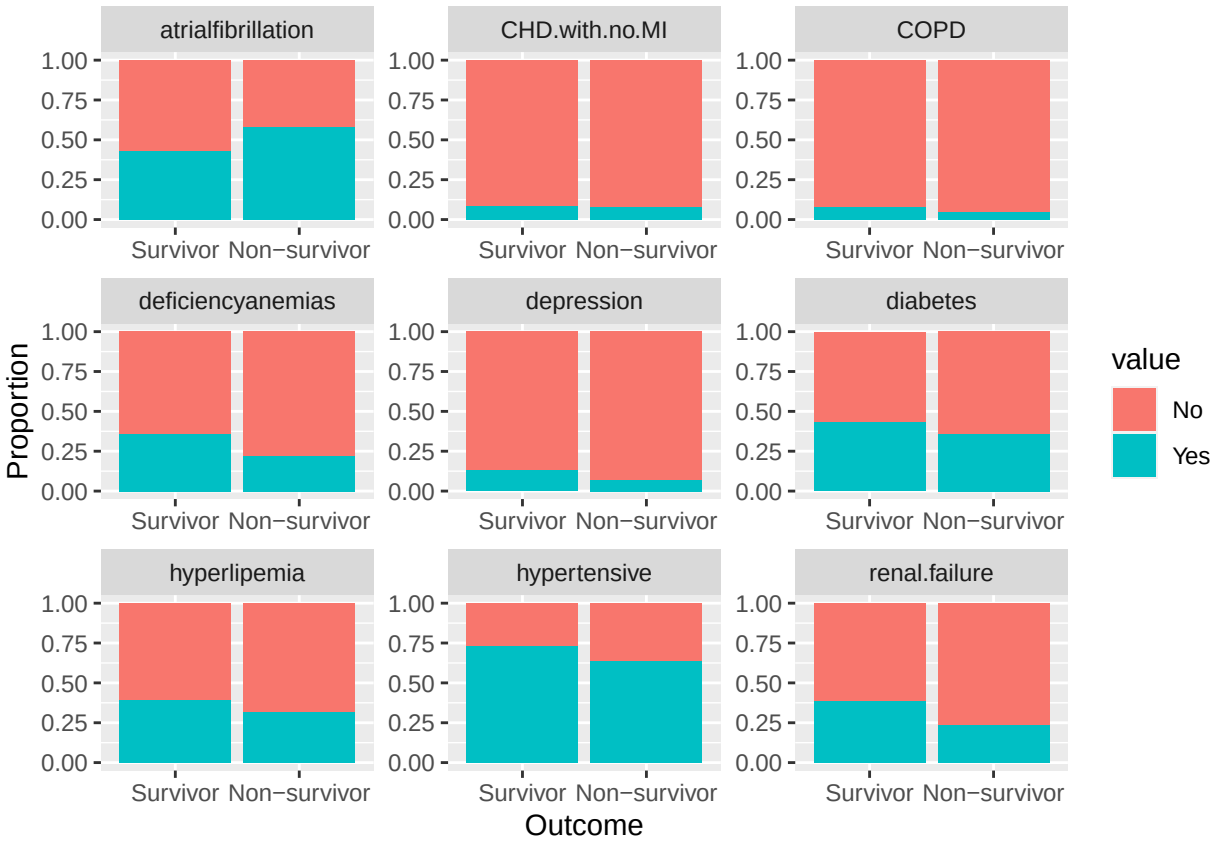
The boxplots indicate that the group of non-survivors has an elevated heart and respiratory rate and lower mean arterial pressures and urine outputs compared to survivors. In the case of temperature and SP.O2, there isn't seem to be an important difference between the two groups indicating that those two variables are probably not good predictors.

Exploring comorbidities

```
##  hypertensive atrialfibrillation CHD.with.no.MI diabetes  deficiencyanemias depression
##  No :332      No :645            No :1075       No :681   No :777           No :1036
##  Yes:844      Yes:531            Yes: 101       Yes:495   Yes:399           Yes: 140
##  hyperlipemia renal.failure  COPD
##  No :729      No :747        No :1087
##  Yes:447      Yes:429        Yes:  89
```



The most prevalent comorbidity is hypertension which is found in almost 75% of the patients. Atrial fibrillation, deficiency anemias, diabetes, hyperlipemia and renal failure are present between 30 to 50% of the patients. Finally, depression, COPD and CHD with no MI are present in less the 10% of the patients.



The relation between the presence or absence of comorbidities between survivors and non-survivors is similar for all variables except for atrial fibrillation. In general, the presence of comorbidities is smaller among non-survivors than survivors. In the case of atrial fibrillation is the opposite, there are more patients with atrial fibrillation among non-survivors than among survivors. This indicates that the presence of atrial fibrillation could be an important outcome predictor.

Exploring laboratory variables

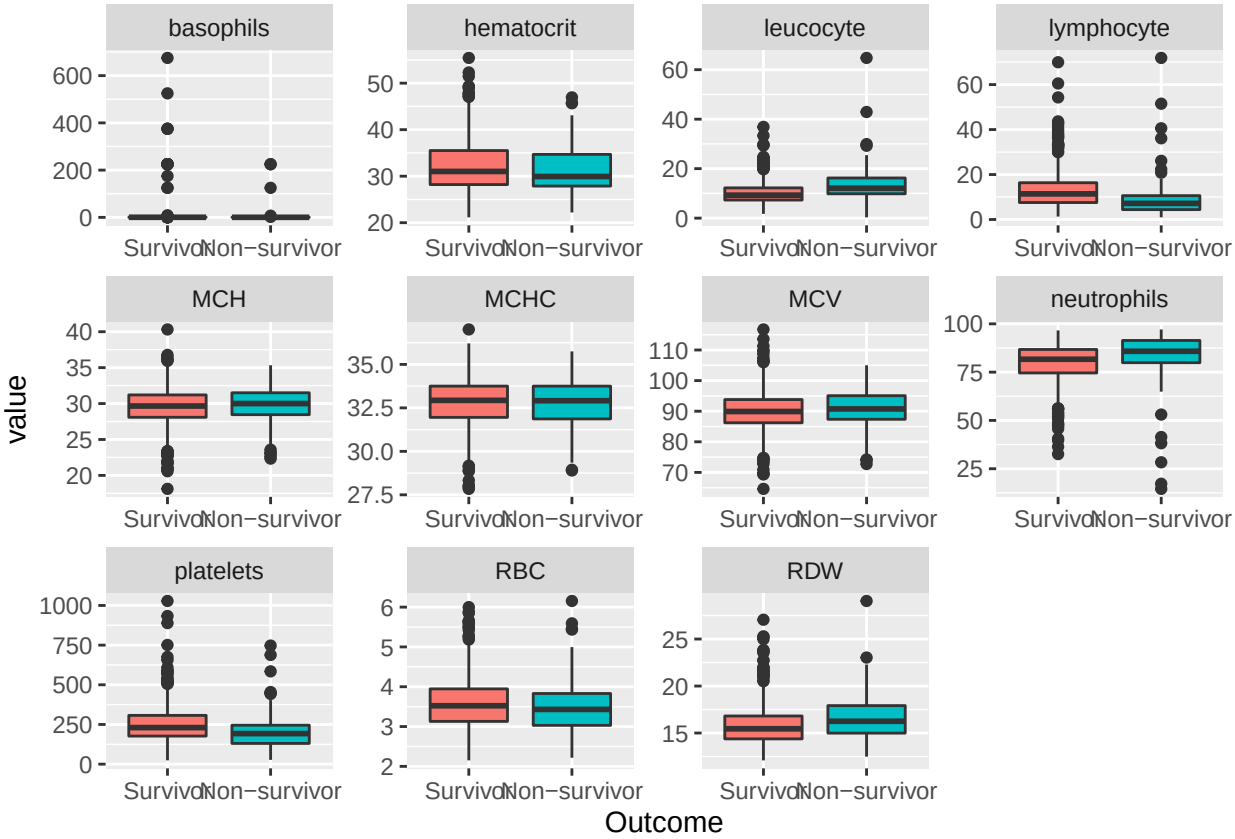
The laboratory variables can be divided into five groups: blood count, coagulation factors, chemistry, venous blood and heart specific.

Blood count

```
##    hematocrit          RBC             MCH             MCHC            MCV
##  Min.   :20.31   Min.   :2.030   Min.   :18.12   Min.   :27.82   Min.   : 62.60
##  1st Qu.:28.15   1st Qu.:3.120   1st Qu.:28.25   1st Qu.:32.01   1st Qu.: 86.25
##  Median :30.80   Median :3.489   Median :29.75   Median :32.99   Median : 90.00
##  Mean   :31.91   Mean   :3.575   Mean   :29.54   Mean   :32.86   Mean   : 89.90
##  3rd Qu.:35.00   3rd Qu.:3.900   3rd Qu.:31.24   3rd Qu.:33.83   3rd Qu.: 93.86
##  Max.   :55.42   Max.   :6.575   Max.   :40.31   Max.   :37.01   Max.   :116.71
##
##       RDW          leucocyte        platelets         neutrophils      basophils
##  Min.   :12.09   Min.   : 0.100   Min.   :   9.571   Min.   : 5.00   Min.   :  0.100
##  1st Qu.:14.46   1st Qu.: 7.436   1st Qu.: 168.904   1st Qu.:74.77   1st Qu.:  0.200
##  Median :15.51   Median : 9.684   Median : 222.667   Median :82.47   Median :  0.300
##  Mean   :15.95   Mean   :10.715   Mean   : 241.518   Mean   :80.12   Mean   :  6.234
```

```
##  3rd Qu.:16.94   3rd Qu.:12.744   3rd Qu.: 304.278   3rd Qu.:87.46   3rd Qu.:  0.500
##  Max.   :29.05   Max.   :64.750   Max.   :1028.200   Max.   :98.00   Max.   :675.000
##                                                     NA's   :144     NA's   :259
##    lymphocyte
##  Min.   : 0.9667
##  1st Qu.: 6.6333
##  Median :10.4667
##  Mean   :12.2327
##  3rd Qu.:15.4750
##  Max.   :83.5000
##  NA's   :145
```

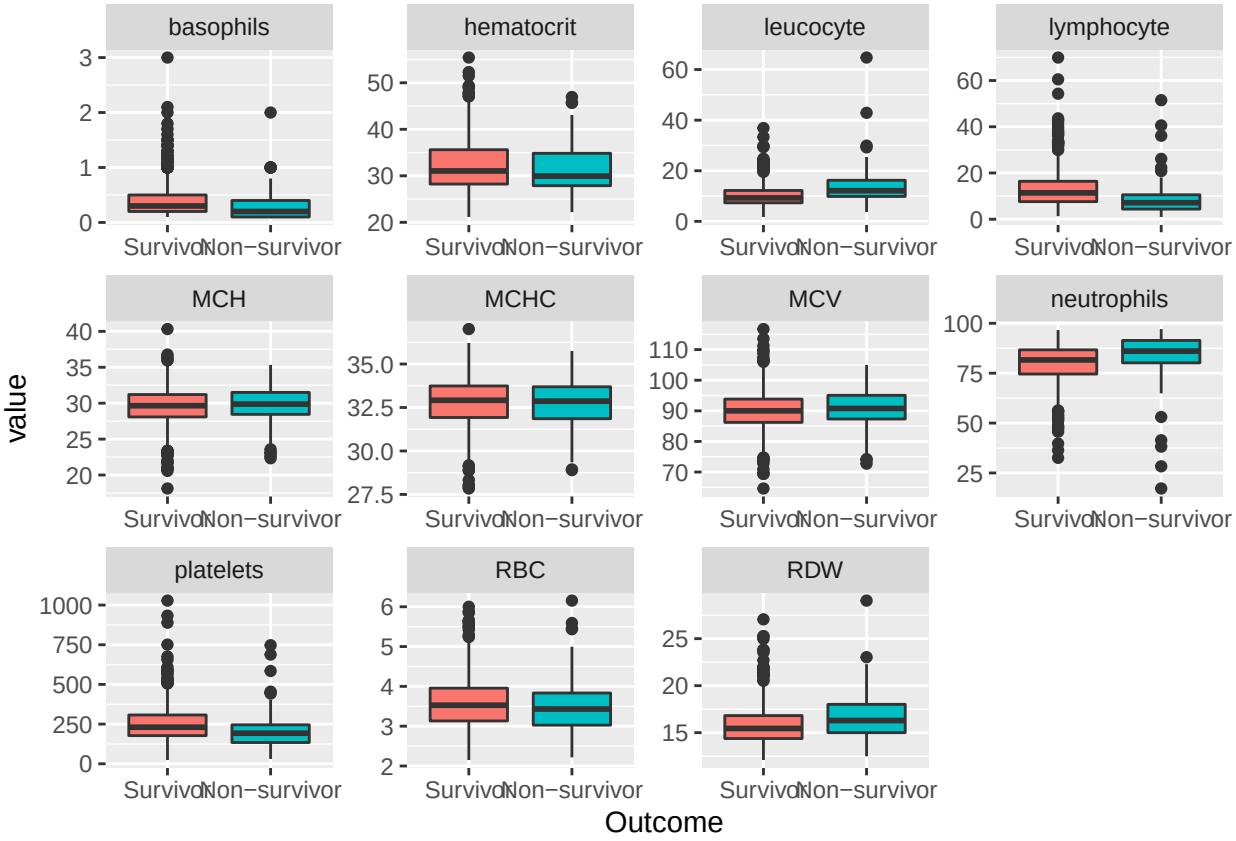
Three variables (neutrophils, basophils and lymphocytes) present between 144 and 259 missing values.



The boxplot of basophils seems to have outliers. Values above 6 are replaced by NA:

```
##    hematocrit          RBC              MCH             MCHC            MCV
##  Min.   :20.31   Min.   :2.030   Min.   :18.12   Min.   :27.82   Min.   : 62.60
##  1st Qu.:28.15   1st Qu.:3.120   1st Qu.:28.25   1st Qu.:32.01   1st Qu.: 86.25
##  Median :30.80   Median :3.489   Median :29.75   Median :32.99   Median : 90.00
##  Mean   :31.91   Mean   :3.575   Mean   :29.54   Mean   :32.86   Mean   : 89.90
##  3rd Qu.:35.00   3rd Qu.:3.900   3rd Qu.:31.24   3rd Qu.:33.83   3rd Qu.: 93.86
##  Max.   :55.42   Max.   :6.575   Max.   :40.31   Max.   :37.01   Max.   :116.71
##
##       RDW          leucocyte        platelets         neutrophils      basophils
##  Min.   :12.09   Min.   : 0.100   Min.   :   9.571   Min.   : 5.00   Min.   :0.100
##  1st Qu.:14.46   1st Qu.: 7.436   1st Qu.: 168.904   1st Qu.:74.77   1st Qu.:0.200
##  Median :15.51   Median : 9.684   Median : 222.667   Median :82.47   Median :0.300
```

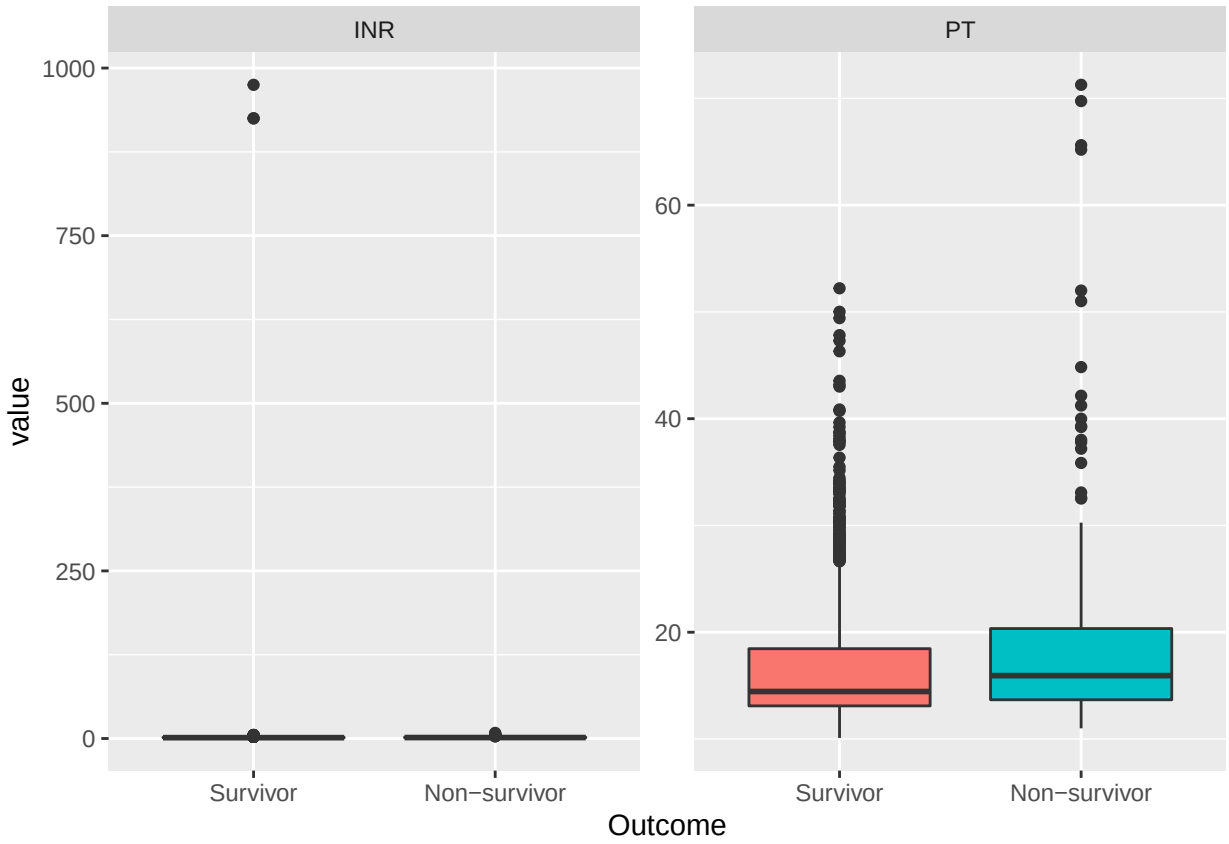
```
##  Mean   :15.95   Mean   :10.715   Mean   : 241.518   Mean   :80.12   Mean   :0.392
##  3rd Qu.:16.94   3rd Qu.:12.744   3rd Qu.: 304.278   3rd Qu.:87.46   3rd Qu.:0.500
##  Max.   :29.05   Max.   :64.750   Max.   :1028.200   Max.   :98.00   Max.   :3.000
##                                                      NA's   :144     NA's   :281
##    lymphocyte
##  Min.   : 0.9667
##  1st Qu.: 6.6333
##  Median :10.4667
##  Mean   :12.2327
##  3rd Qu.:15.4750
##  Max.   :83.5000
##  NA's   :145
```



Coagulation factors

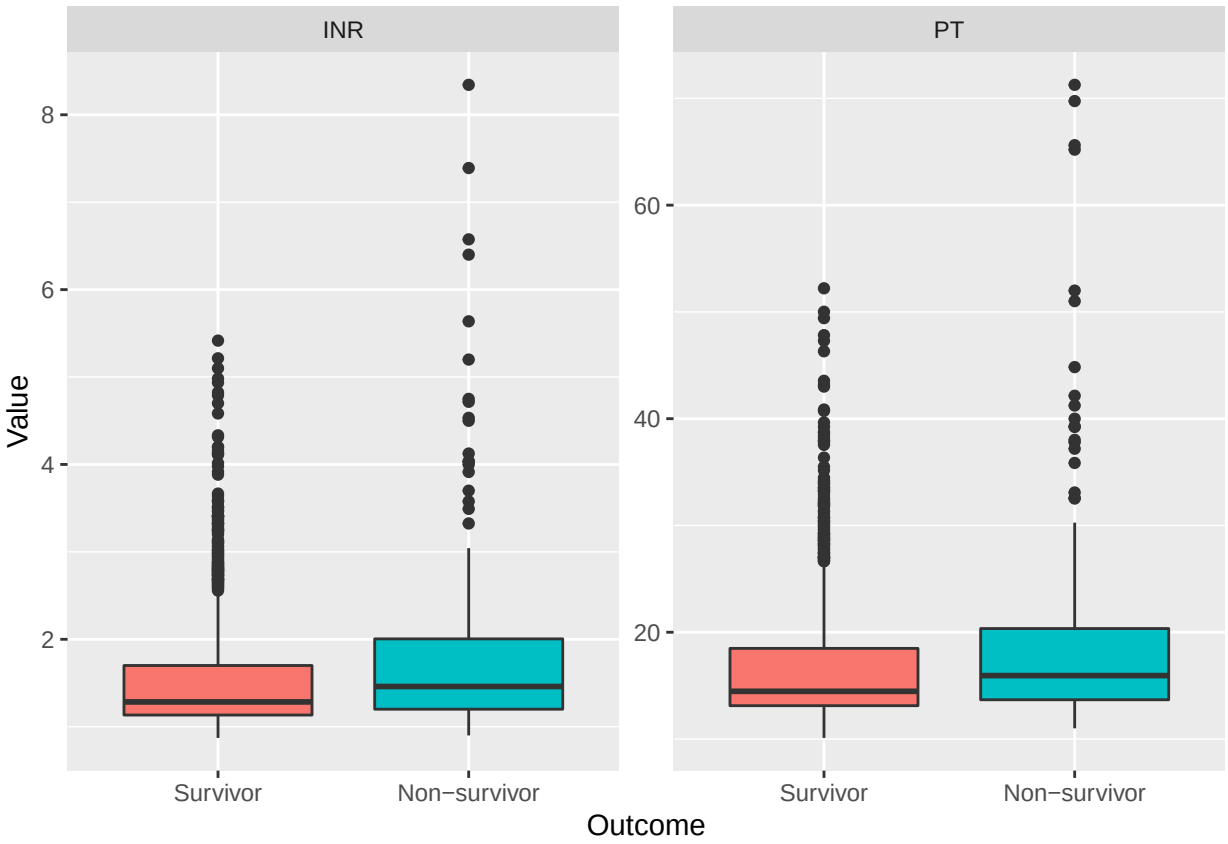
```
##        PT              INR
##  Min.   :10.10   Min.   :  0.8714
##  1st Qu.:13.16   1st Qu.:  1.1400
##  Median :14.64   Median :  1.3000
##  Mean   :17.49   Mean   :  4.0674
##  3rd Qu.:18.80   3rd Qu.:  1.7433
##  Max.   :71.27   Max.   :975.0000
##  NA's   :20      NA's   :20
```

Both variables present 20 missing values.



The variable INR seems to have a few outliers. They will be replaced for missing values.

```
##        PT              INR
##  Min.   :10.10   Min.   :0.8714
##  1st Qu.:13.16   1st Qu.:1.1400
##  Median :14.64   Median :1.3000
##  Mean   :17.49   Mean   :1.6278
##  3rd Qu.:18.80   3rd Qu.:1.7364
##  Max.   :71.27   Max.   :8.3429
##  NA's   :20      NA's   :23
```

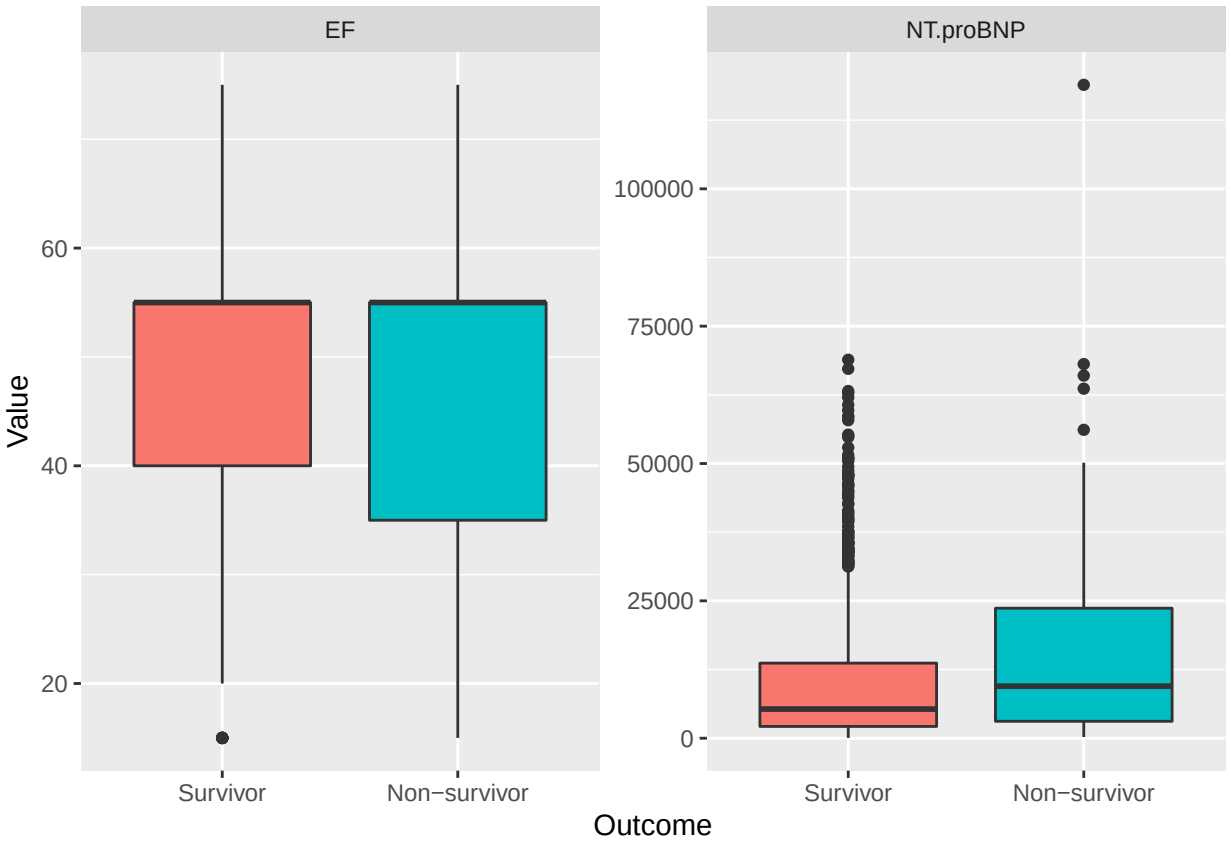


In both variables the median is higher for survivors than non-survivors and the data is right-skewed.

Heart specifics factors

```
##    NT.proBNP            EF
##  Min.   :    50   Min.   :15.00
##  1st Qu.:  2250   1st Qu.:40.00
##  Median :  5838   Median :55.00
##  Mean   : 11011   Mean   :48.71
##  3rd Qu.: 14981   3rd Qu.:55.00
##  Max.   :118928   Max.   :75.00
```

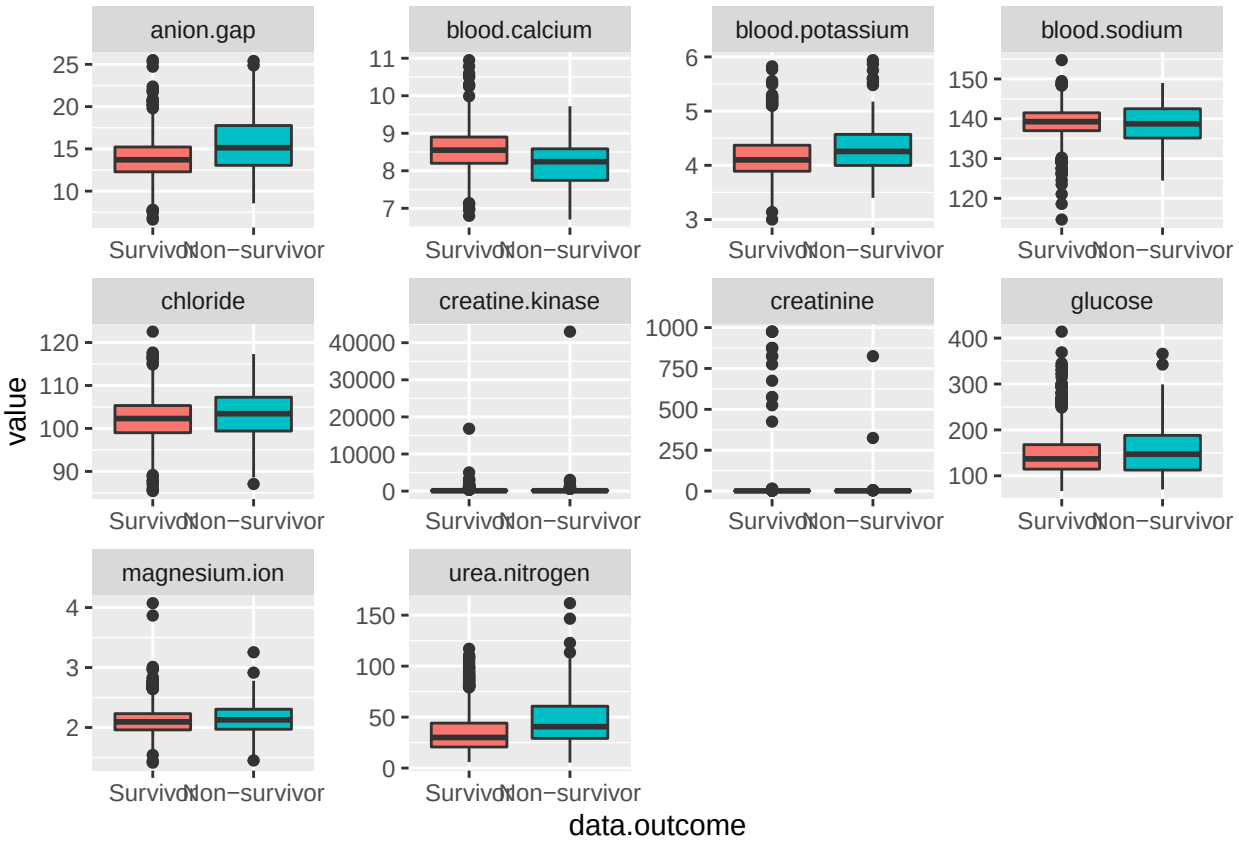
Heart specific factors have all values.



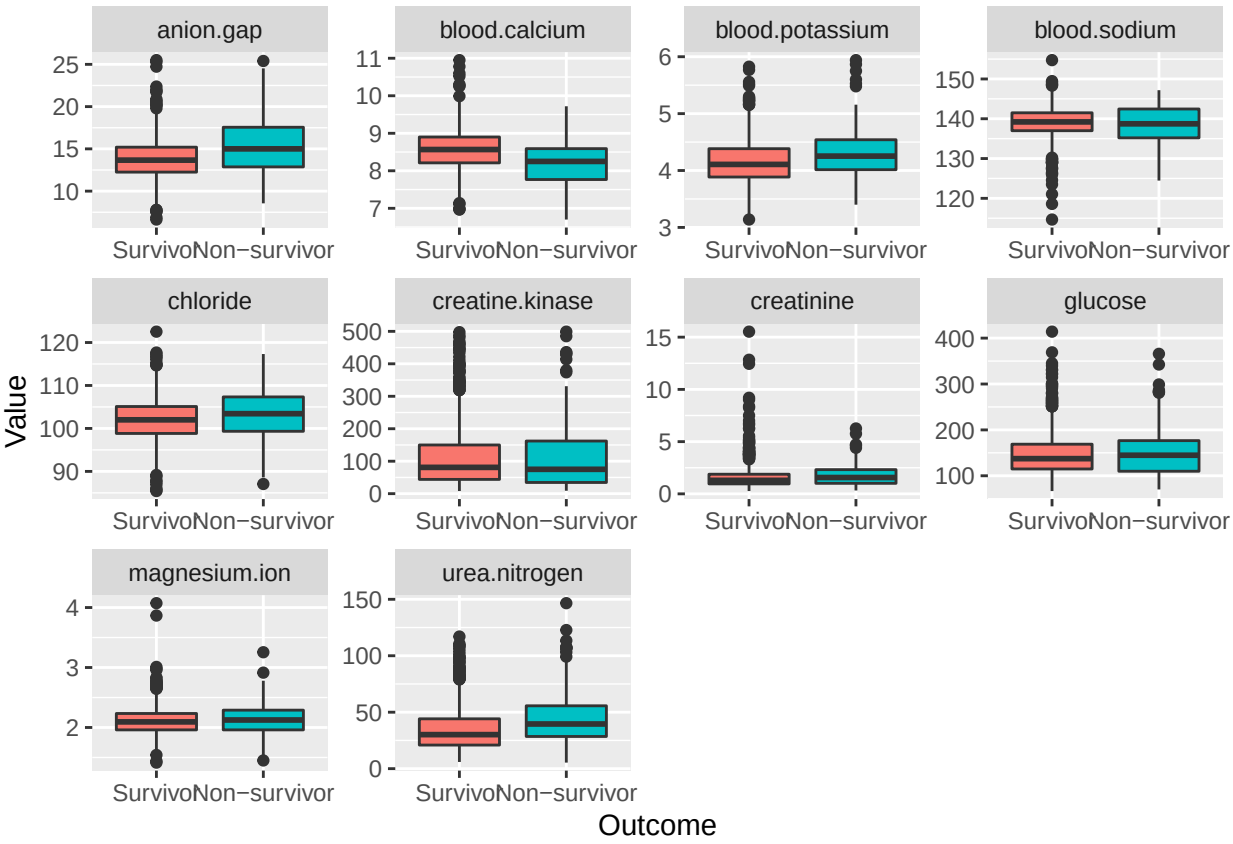
Chemistry

```
##  creatine.kinase     creatinine       urea.nitrogen        glucose       blood.potassium
##  Min.   :    8.0   Min.   :  0.2667   Min.   :  5.357   Min.   : 66.67   Min.   :3.000
##  1st Qu.:   46.0   1st Qu.:  0.9600   1st Qu.: 20.833   1st Qu.:113.94   1st Qu.:3.900
##  Median :   89.5   Median :  1.3279   Median : 30.611   Median :136.40   Median :4.115
##  Mean   :  246.9   Mean   : 15.9970   Mean   : 36.294   Mean   :148.80   Mean   :4.176
##  3rd Qu.:  185.4   3rd Qu.:  1.9682   3rd Qu.: 45.256   3rd Qu.:169.50   3rd Qu.:4.400
##  Max.   :42987.5   Max.   :975.0000   Max.   :161.750   Max.   :414.10   Max.   :6.567
##  NA's   :165                                            NA's   :17
##   blood.sodium   blood.calcium       chloride        anion.gap      magnesium.ion
##  Min.   :114.7   Min.   : 6.700   Min.   : 80.27   Min.   : 6.636   Min.   :1.400
##  1st Qu.:136.7   1st Qu.: 8.150   1st Qu.: 99.00   1st Qu.:12.250   1st Qu.:1.956
##  Median :139.2   Median : 8.500   Median :102.52   Median :13.667   Median :2.093
##  Mean   :138.9   Mean   : 8.502   Mean   :102.29   Mean   :13.924   Mean   :2.120
##  3rd Qu.:141.6   3rd Qu.: 8.869   3rd Qu.:105.57   3rd Qu.:15.404   3rd Qu.:2.242
##  Max.   :154.7   Max.   :10.950   Max.   :122.53   Max.   :25.500   Max.   :4.073
##                  NA's   :1
```

Blood calcium, glucose and creatine kinase present missing values.



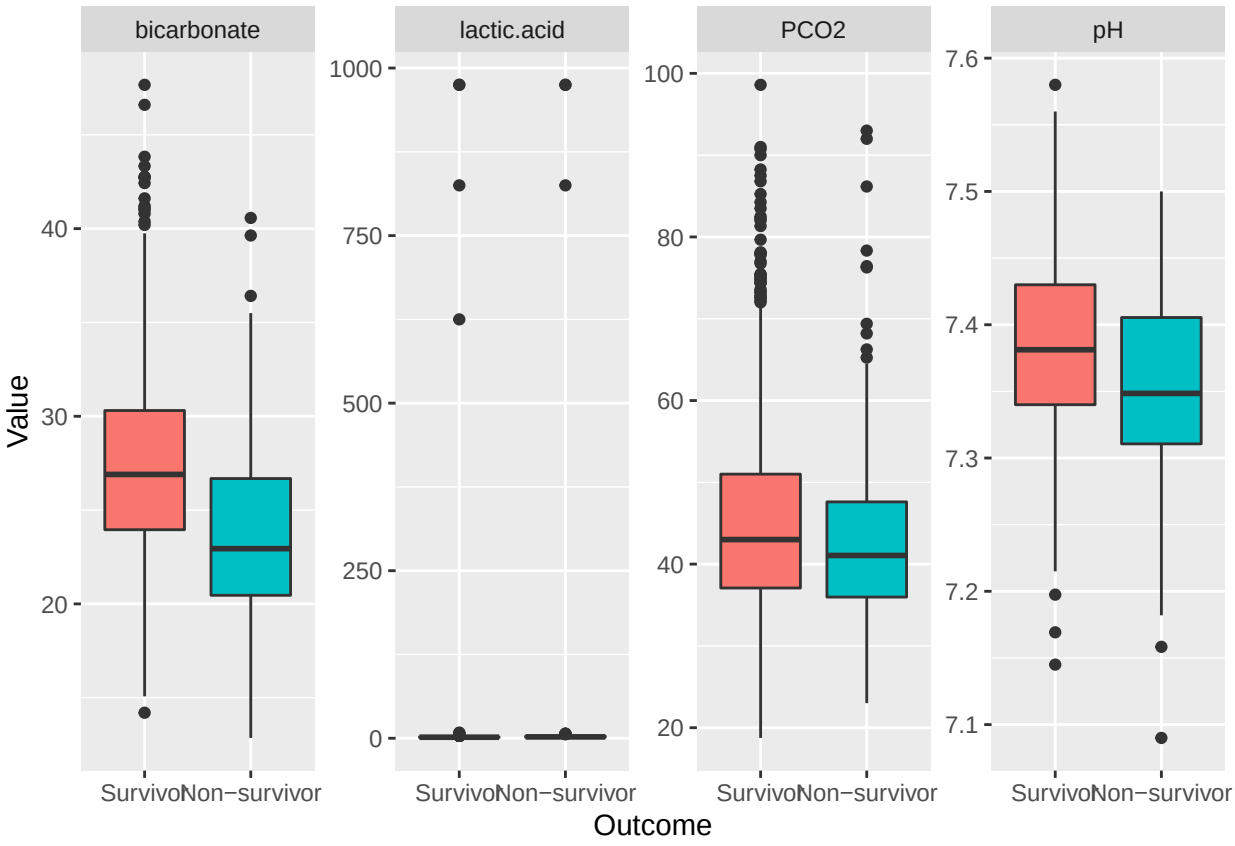

```
##  creatine.kinase    creatinine      urea.nitrogen        glucose       blood.potassium
##  Min.   :  8.00   Min.   : 0.2667   Min.   :  5.357   Min.   : 66.67   Min.   :3.000
##  1st Qu.: 43.15   1st Qu.: 0.9514   1st Qu.: 20.833   1st Qu.:113.94   1st Qu.:3.900
##  Median : 80.50   Median : 1.3077   Median : 30.611   Median :136.40   Median :4.115
##  Mean   :113.91   Mean   : 1.6573   Mean   : 36.294   Mean   :148.80   Mean   :4.176
##  3rd Qu.:152.12   3rd Qu.: 1.9047   3rd Qu.: 45.256   3rd Qu.:169.50   3rd Qu.:4.400
##  Max.   :499.00   Max.   :15.5273   Max.   :161.750   Max.   :414.10   Max.   :6.567
##  NA's   :248      NA's   :22                          NA's   :17
##   blood.sodium   blood.calcium       chloride        anion.gap      magnesium.ion
##  Min.   :114.7   Min.   : 6.700   Min.   : 80.27   Min.   : 6.636   Min.   :1.400
##  1st Qu.:136.7   1st Qu.: 8.150   1st Qu.: 99.00   1st Qu.:12.250   1st Qu.:1.956
##  Median :139.2   Median : 8.500   Median :102.52   Median :13.667   Median :2.093
##  Mean   :138.9   Mean   : 8.502   Mean   :102.29   Mean   :13.924   Mean   :2.120
##  3rd Qu.:141.6   3rd Qu.: 8.869   3rd Qu.:105.57   3rd Qu.:15.404   3rd Qu.:2.242
##  Max.   :154.7   Max.   :10.950   Max.   :122.53   Max.   :25.500   Max.   :4.073
##                  NA's   :1
```



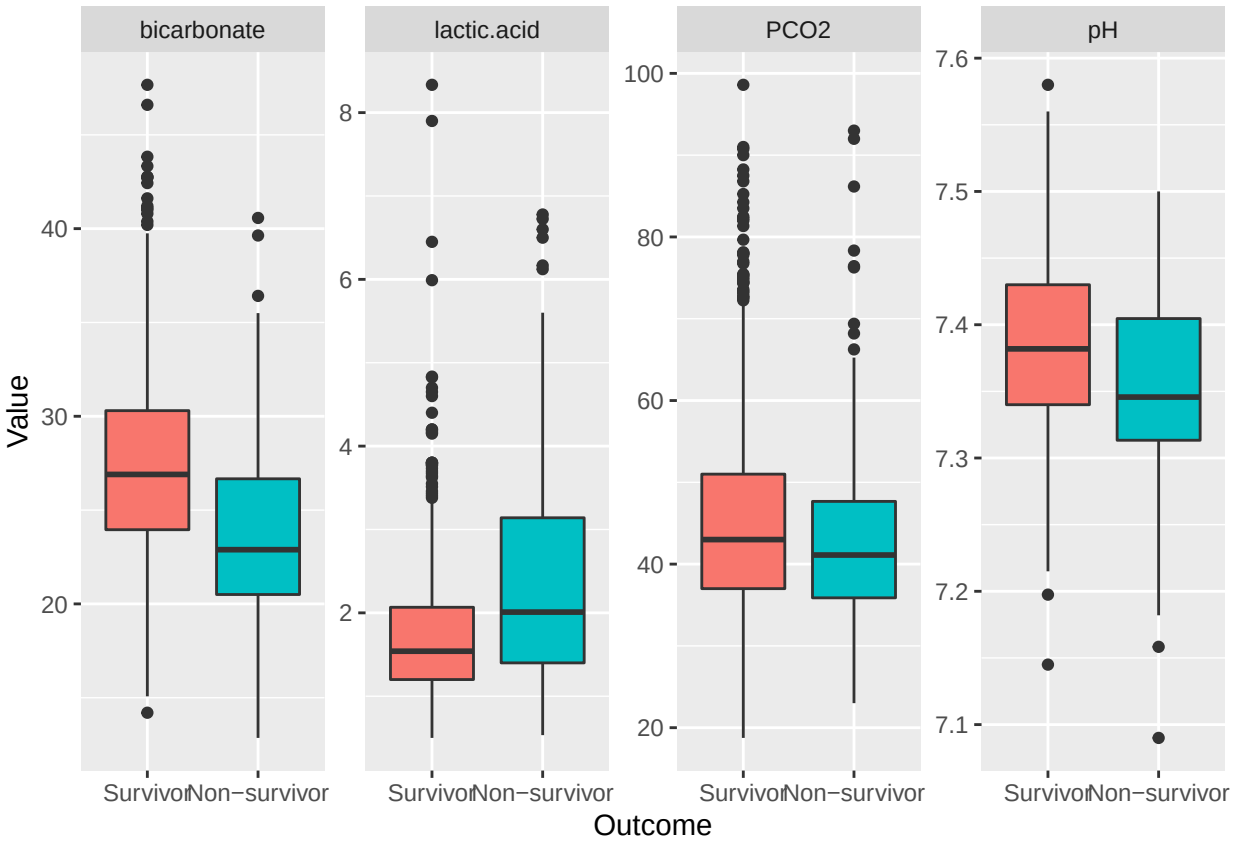
Venous blood factors

```
##        pH          bicarbonate     lactic.acid          PCO2
##  Min.   :7.090   Min.   :12.86   Min.   :  0.500   Min.   :18.75
##  1st Qu.:7.335   1st Qu.:23.45   1st Qu.:  1.200   1st Qu.:37.04
##  Median :7.380   Median :26.50   Median :  1.620   Median :43.00
##  Mean   :7.379   Mean   :26.91   Mean   :  8.361   Mean   :45.54
##  3rd Qu.:7.430   3rd Qu.:29.88   3rd Qu.:  2.200   3rd Qu.:50.59
##  Max.   :7.580   Max.   :47.67   Max.   :975.000   Max.   :98.60
##  NA's   :291                     NA's   :228       NA's   :293
```

pH, PCO2 and lactic acid present between 229 to 294 missing values.




```
##        pH          bicarbonate     lactic.acid        PCO2
##  Min.   :7.090   Min.   :12.86   Min.   :0.500   Min.   :18.75
##  1st Qu.:7.335   1st Qu.:23.45   1st Qu.:1.200   1st Qu.:37.04
##  Median :7.380   Median :26.50   Median :1.614   Median :43.00
##  Mean   :7.379   Mean   :26.91   Mean   :1.861   Mean   :45.54
##  3rd Qu.:7.430   3rd Qu.:29.88   3rd Qu.:2.200   3rd Qu.:50.59
##  Max.   :7.580   Max.   :47.67   Max.   :8.333   Max.   :98.60
##  NA's   :291                     NA's   :235     NA's   :293
```



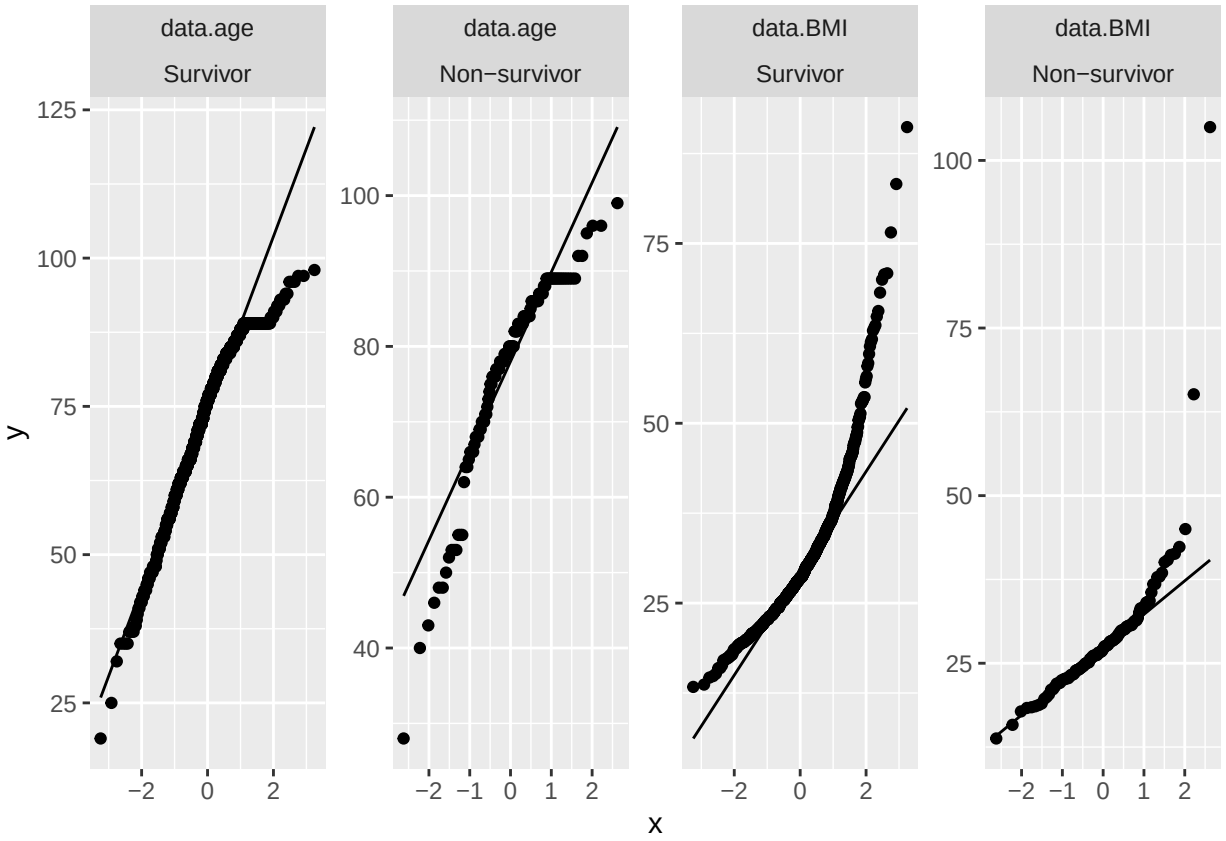
According to the boxplots, the group of non-survivors has higher RDW, leucocytes, neutrophils, PT, NT.proBNP, urea, potassium, chloride, anion gap, magnesium and lactic acid compared to the survivors. They also ha lower platelet, lymphocytes, sodium, calcium pH, PCO2 and bicarbonate.

Step 4 - Bivariate analysis

As far as I know, in order to assess group comparisons, Li 2021 used T-test and Mann-Whitney-Wilcoxon Test for continuous variables and Chi-squared or Fisher's exact tests for categorical variables. Therefore they did no apply multivariate testing.

Outcome group comparison for demographic variables

Age and BMI



numerical_variables	variable	statistic	p
data.age	Values	0.9341760	0
data.BMI	Values	0.8431138	0

Neither variable presents a normal distribution.

```
## data.age ~ data.outcome data.BMI ~ data.outcome
##              0.01650458              0.01732153
```

As the p-value is significant in both cases, we conclude that age and BMI are different for survivors and non-survivors. Non-survivors had lower BMI and different age incidence compared to the survivors.

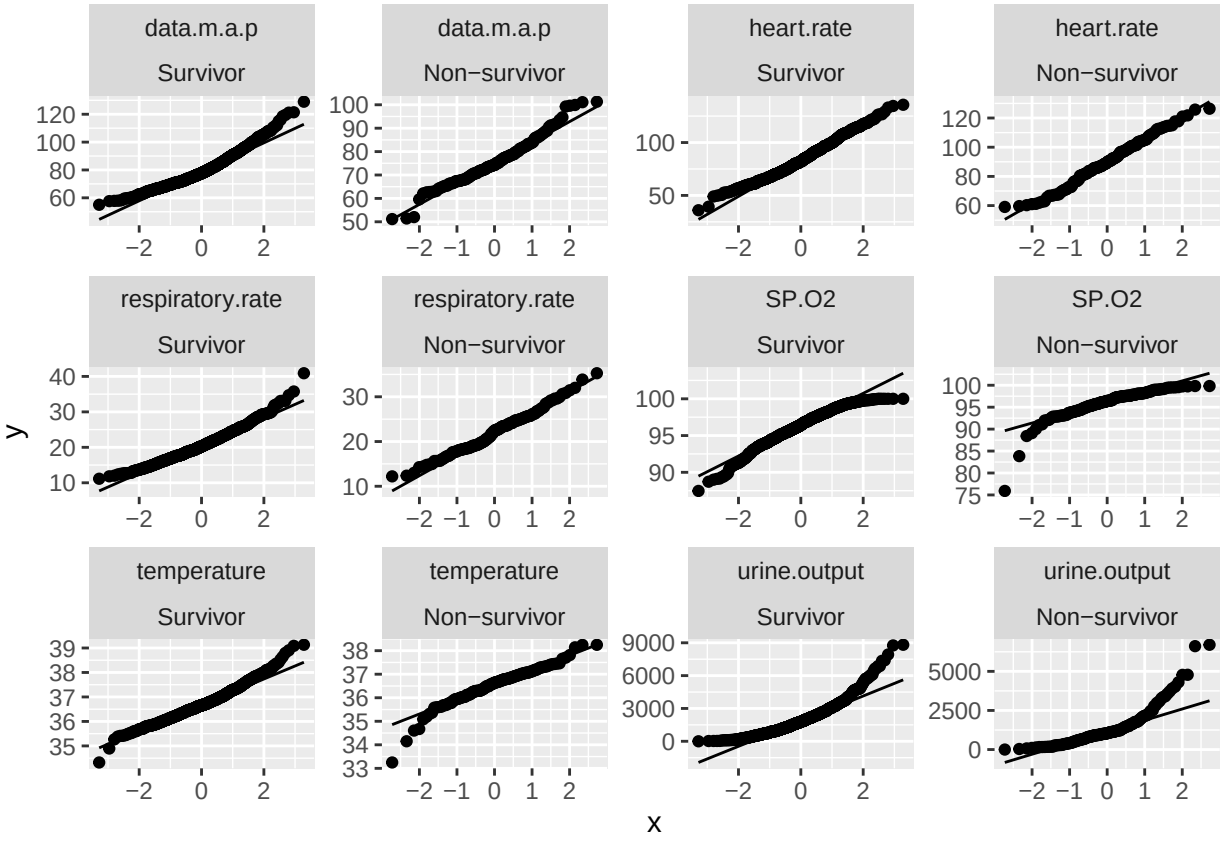
Gender

outcome	gender	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	M	478	1017	47.00	0.88	0.62	1.24	0.44
Survivor	F	539	1017	53.00	0.88	0.62	1.24	0.44
Non-survivor	M	80	159	50.31	0.88	0.62	1.24	0.44

outcome	gender	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Non-survivor	F	79	159	49.69	0.88	0.62	1.24	0.44

The p-value is no significant, because of that, gender got discarded as outcome predictor.

Outcome group comparison for vital signs



numerical_variables	variable	statistic	p
data.m.a.p	Values	0.9644124	0e+00
heart.rate	Values	0.9900906	5e-07
respiratory.rate	Values	0.9834534	0e+00
SP.O2	Values	0.9302654	0e+00
temperature	Values	0.9827231	0e+00
urine.output	Values	0.9081972	0e+00

All vital signs variables do not present a normal distribution.

```
##       heart.rate ~ data.outcome respiratory.rate ~ data.outcome
##                    4.214480e-06                    1.788116e-04
##      temperature ~ data.outcome            SP.O2 ~ data.outcome
##                    5.465815e-02                    2.614293e-01
##     urine.output ~ data.outcome       data.m.a.p ~ data.outcome
##                    3.819006e-13                    2.574004e-05
```

All the p-values resulting from the Wilcoxon-rank test are significant indicating that all vital signs are different between the survivors and non-survivors.

Outcome group comparison of comorbidities

outcome	hypertensive	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	274	1017	26.94	0.64	0.45	0.93	0.02
Survivor	Yes	743	1017	73.06	0.64	0.45	0.93	0.02
Non-survivor	No	58	159	36.48	0.64	0.45	0.93	0.02
Non-survivor	Yes	101	159	63.52	0.64	0.45	0.93	0.02

outcome	atrialfibrillation	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	578	1017	56.83	1.81	1.27	2.58	0
Survivor	Yes	439	1017	43.17	1.81	1.27	2.58	0
Non-survivor	No	67	159	42.14	1.81	1.27	2.58	0
Non-survivor	Yes	92	159	57.86	1.81	1.27	2.58	0

outcome	CHD.with.no.MI	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	928	1017	91.25	0.85	0.41	1.61	0.76
Survivor	Yes	89	1017	8.75	0.85	0.41	1.61	0.76
Non-survivor	No	147	159	92.45	0.85	0.41	1.61	0.76
Non-survivor	Yes	12	159	7.55	0.85	0.41	1.61	0.76

outcome	diabetes	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	579	1017	56.93	0.74	0.51	1.06	0.1
Survivor	Yes	438	1017	43.07	0.74	0.51	1.06	0.1
Non-survivor	No	102	159	64.15	0.74	0.51	1.06	0.1
Non-survivor	Yes	57	159	35.85	0.74	0.51	1.06	0.1

outcome	deficiencyanemias	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	653	1017	64.21	0.51	0.33	0.76	0
Survivor	Yes	364	1017	35.79	0.51	0.33	0.76	0
Non-survivor	No	124	159	77.99	0.51	0.33	0.76	0
Non-survivor	Yes	35	159	22.01	0.51	0.33	0.76	0

outcome	depression	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	888	1017	87.32	0.51	0.24	0.98	0.04
Survivor	Yes	129	1017	12.68	0.51	0.24	0.98	0.04
Non-survivor	No	148	159	93.08	0.51	0.24	0.98	0.04
Non-survivor	Yes	11	159	6.92	0.51	0.24	0.98	0.04

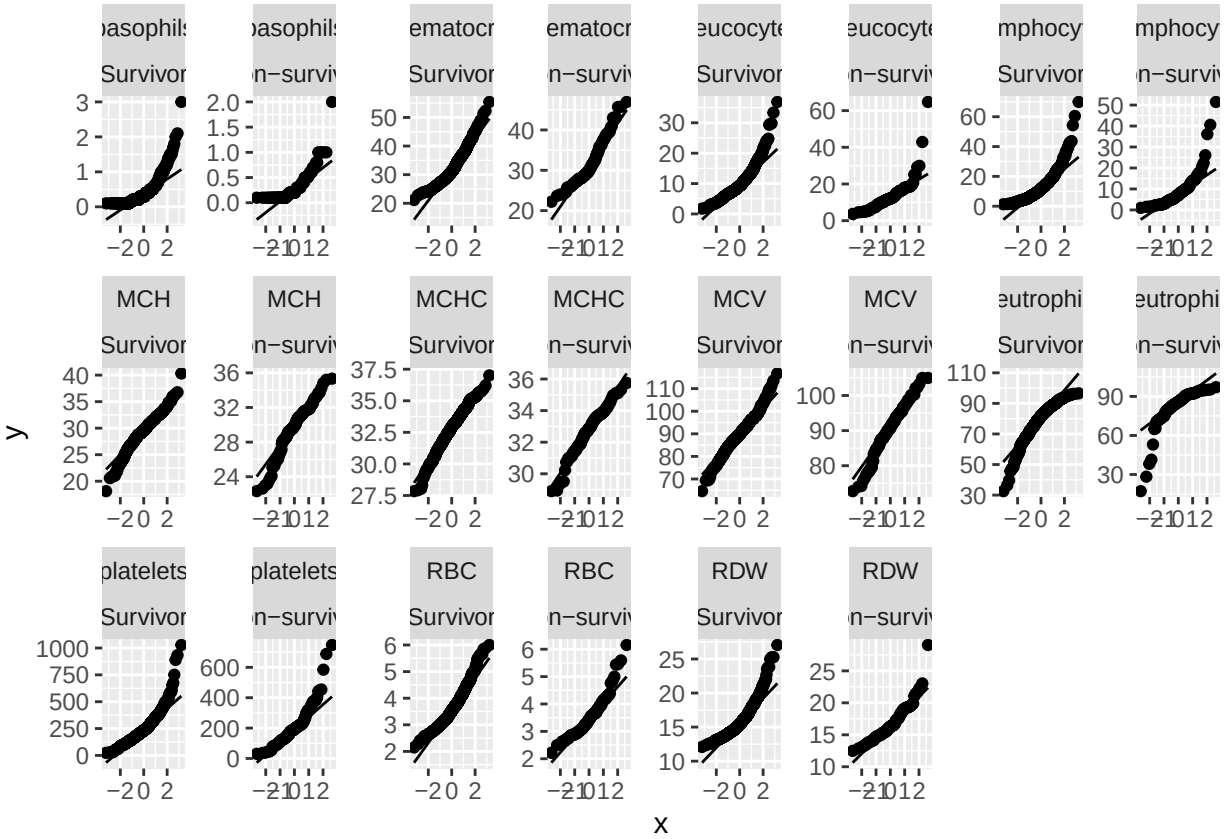
outcome	hyperlipemia	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	620	1017	60.96	0.72	0.49	1.04	0.08
Survivor	Yes	397	1017	39.04	0.72	0.49	1.04	0.08
Non-survivor	No	109	159	68.55	0.72	0.49	1.04	0.08
Non-survivor	Yes	50	159	31.45	0.72	0.49	1.04	0.08

outcome	renal.failure	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	625	1017	61.46	0.48	0.32	0.72	0
Survivor	Yes	392	1017	38.54	0.48	0.32	0.72	0
Non-survivor	No	122	159	76.73	0.48	0.32	0.72	0
Non-survivor	Yes	37	159	23.27	0.48	0.32	0.72	0

outcome	COPD	n	TOTAL	Freq	OR	OR_low	OR_upp	pvalor
Survivor	No	935	1017	91.94	0.53	0.2	1.16	0.14
Survivor	Yes	82	1017	8.06	0.53	0.2	1.16	0.14
Non-survivor	No	152	159	95.60	0.53	0.2	1.16	0.14
Non-survivor	Yes	7	159	4.40	0.53	0.2	1.16	0.14

The p-value is significant for the comorbidities of hypertension, atrial fibrillation, deficiency anemias, depression and renal failure.

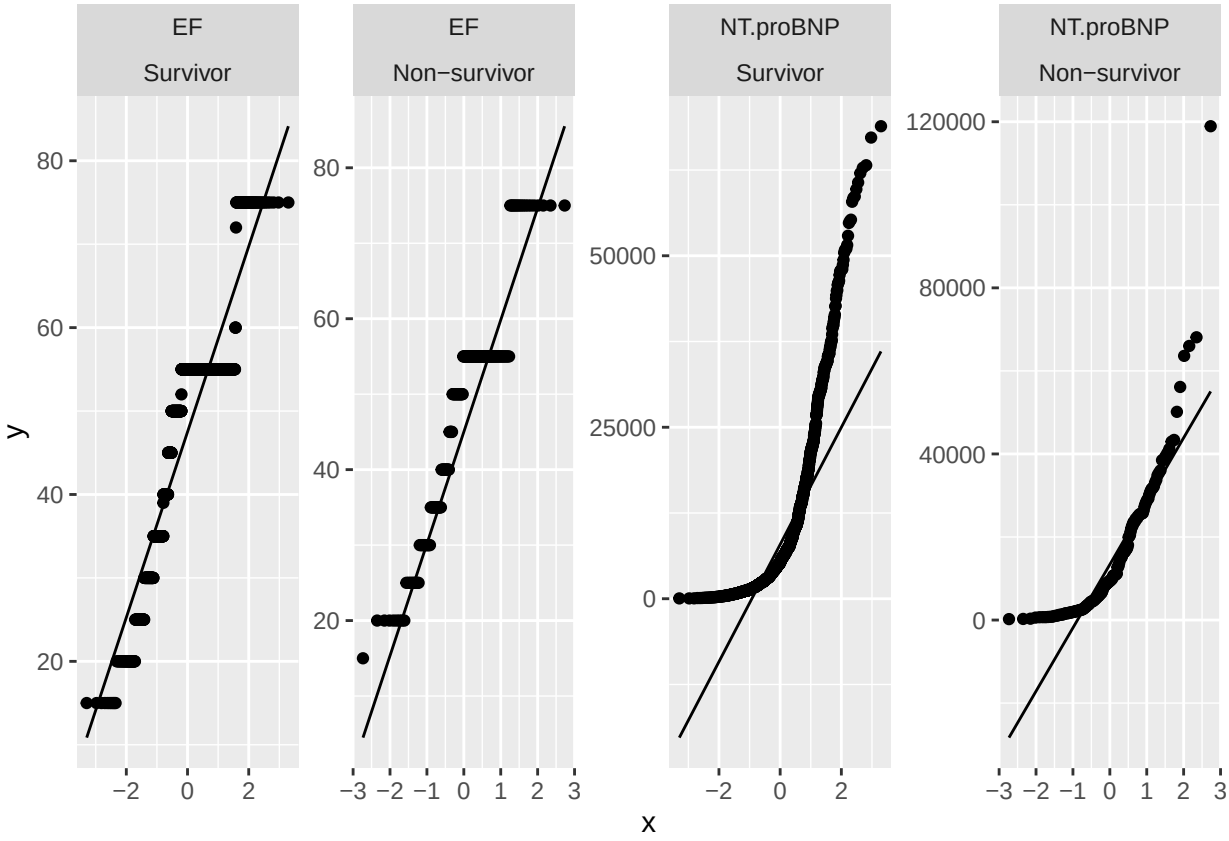
Outcome group comparison for lab variables



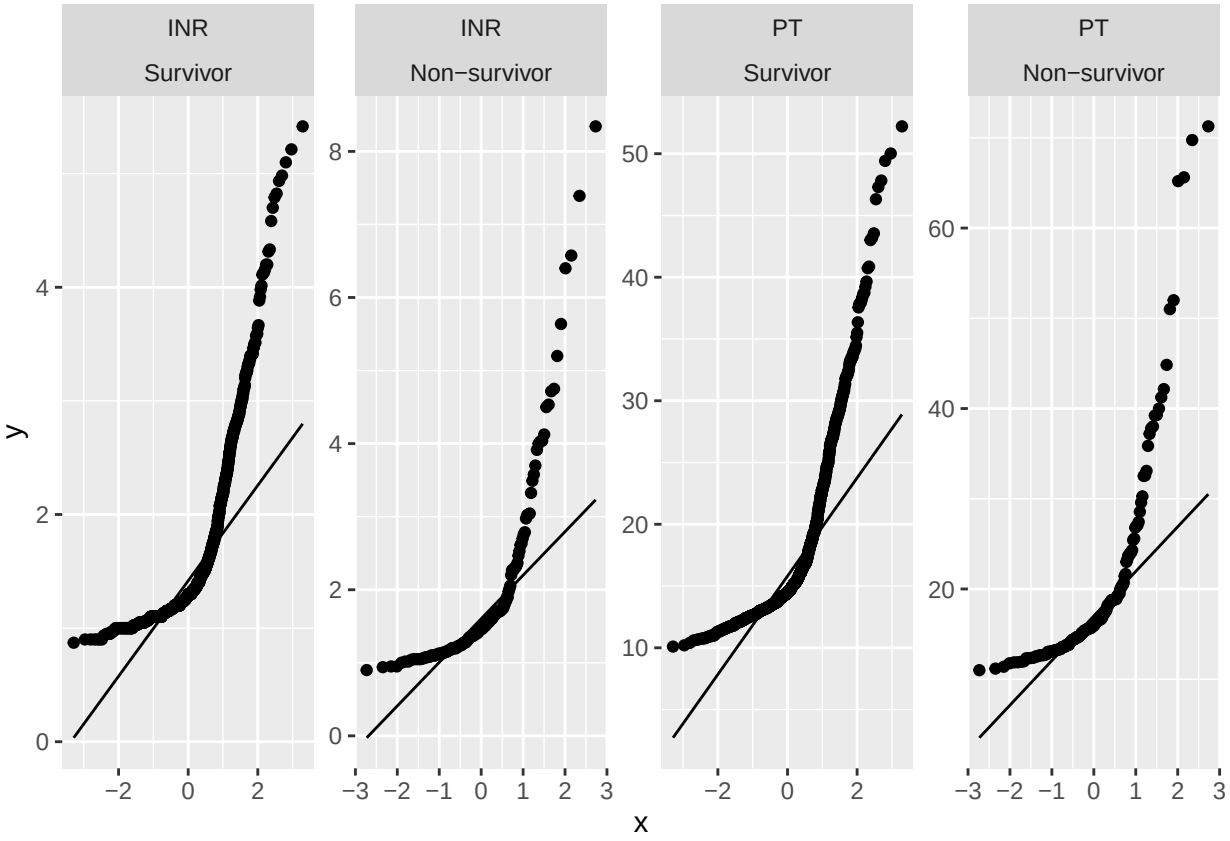
blood_variables	variable	statistic	p
basophils	Values	0.7855873	0
hematocrit	Values	0.9474574	0
leucocyte	Values	0.8411816	0
lymphocyte	Values	0.8203446	0
MCH	Values	0.9718948	0
MCHC	Values	0.9876165	0
MCV	Values	0.9881094	0
neutrophils	Values	0.8635688	0
platelets	Values	0.9183727	0
RBC	Values	0.9541083	0
RDW	Values	0.9130230	0

Shapiro's test p values are all significant so they do not follow a normal distribution.

```
##  hematocrit ~ data.outcome         RBC ~ data.outcome         MCH ~ data.outcome
##               5.103751e-01                2.201564e-01               2.868608e-01
##        MCHC ~ data.outcome         MCV ~ data.outcome         RDW ~ data.outcome
##               3.880542e-01                8.334707e-02               7.221857e-08
##   leucocyte ~ data.outcome   platelets ~ data.outcome neutrophils ~ data.outcome
##               1.115906e-11                4.285584e-05               1.187011e-06
##   basophils ~ data.outcome  lymphocyte ~ data.outcome
##               7.728385e-05                9.848251e-13
```



heart_variables	variable	statistic	p
EF	Values	0.8516930	0
NT.proBNP	Values	0.7536759	0

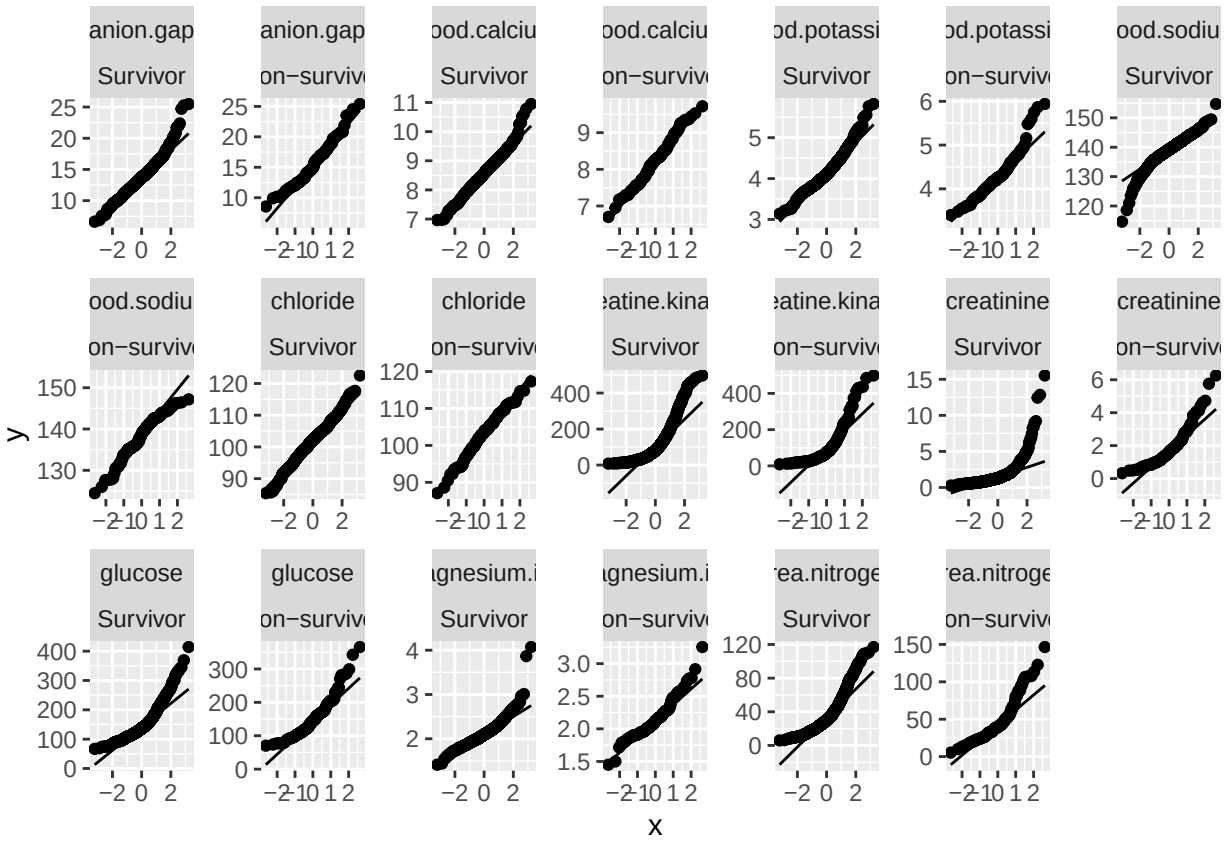


coagulation_variables	variable	statistic	p
INR	Values	0.6949326	0
PT	Values	0.7118099	0

Both heart specific factors and coagulation factors do not present a normal distribution.

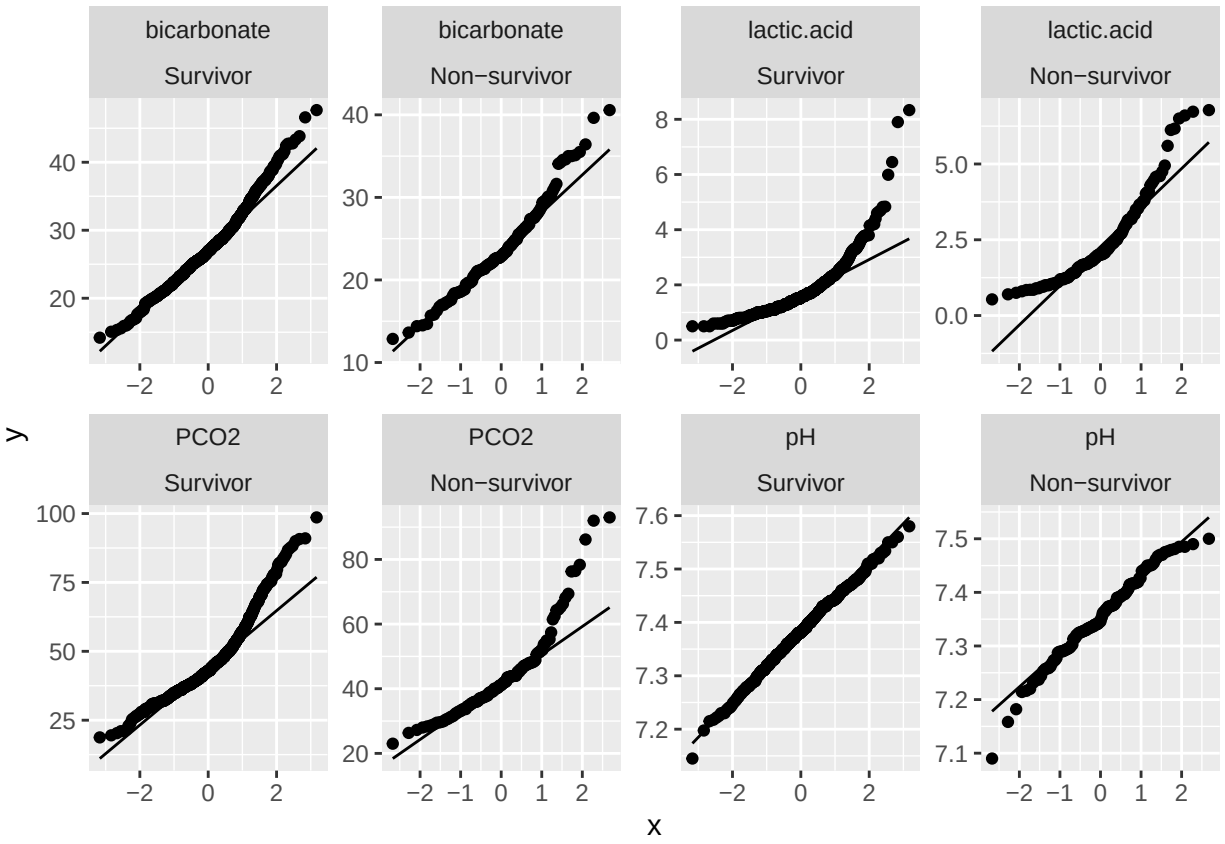
```
## NT.proBNP ~ data.outcome        EF ~ data.outcome
##             6.654393e-05             2.427507e-01

##  PT ~ data.outcome INR ~ data.outcome
##       8.865229e-05       1.973914e-04
```



chemistry_variables	variable	statistic	p
anion.gap	Values	0.9657477	0.0000000
blood.calcium	Values	0.9927618	0.0000165
blood.potassium	Values	0.9567158	0.0000000
blood.sodium	Values	0.9654099	0.0000000
chloride	Values	0.9973323	0.0479430
creatine.kinase	Values	0.8277572	0.0000000
creatinine	Values	0.6673312	0.0000000
glucose	Values	0.8858069	0.0000000
magnesium.ion	Values	0.9354107	0.0000000
urea.nitrogen	Values	0.8768406	0.0000000

```
## creatine.kinase ~ data.outcome      creatinine ~ data.outcome
##                   2.276969e-01                   7.341392e-03
##   urea.nitrogen ~ data.outcome         glucose ~ data.outcome
##                   2.831815e-10                   5.800849e-01
## blood.potassium ~ data.outcome    blood.sodium ~ data.outcome
##                   2.419953e-04                   8.588213e-02
##   blood.calcium ~ data.outcome        chloride ~ data.outcome
##                   2.433626e-09                   2.326432e-02
##       anion.gap ~ data.outcome   magnesium.ion ~ data.outcome
##                   3.340753e-09                   4.176436e-02
```



venous_variables	variable	statistic	p
bicarbonate	Values	0.9852710	0.0000000
lactic.acid	Values	0.8196621	0.0000000
PCO2	Values	0.9177035	0.0000000
pH	Values	0.9946401	0.0031947

```
##          pH ~ data.outcome bicarbonate ~ data.outcome lactic.acid ~ data.outcome
##                4.518966e-05                3.303179e-15                1.036033e-07
##        PCO2 ~ data.outcome
##                5.708430e-02
```

Neither of the laboratory variables presents a normal distribution and all the p-values resulting from the Mann-Whitney-Wilcoxon test are significant.

In the bivariate analysis, there is not enough statistical evidence that there are differences between outcome and gender, hyperlipemia, diabetes, CHD and COPD. Therefore, those variables will be removed when modeling.