HW9

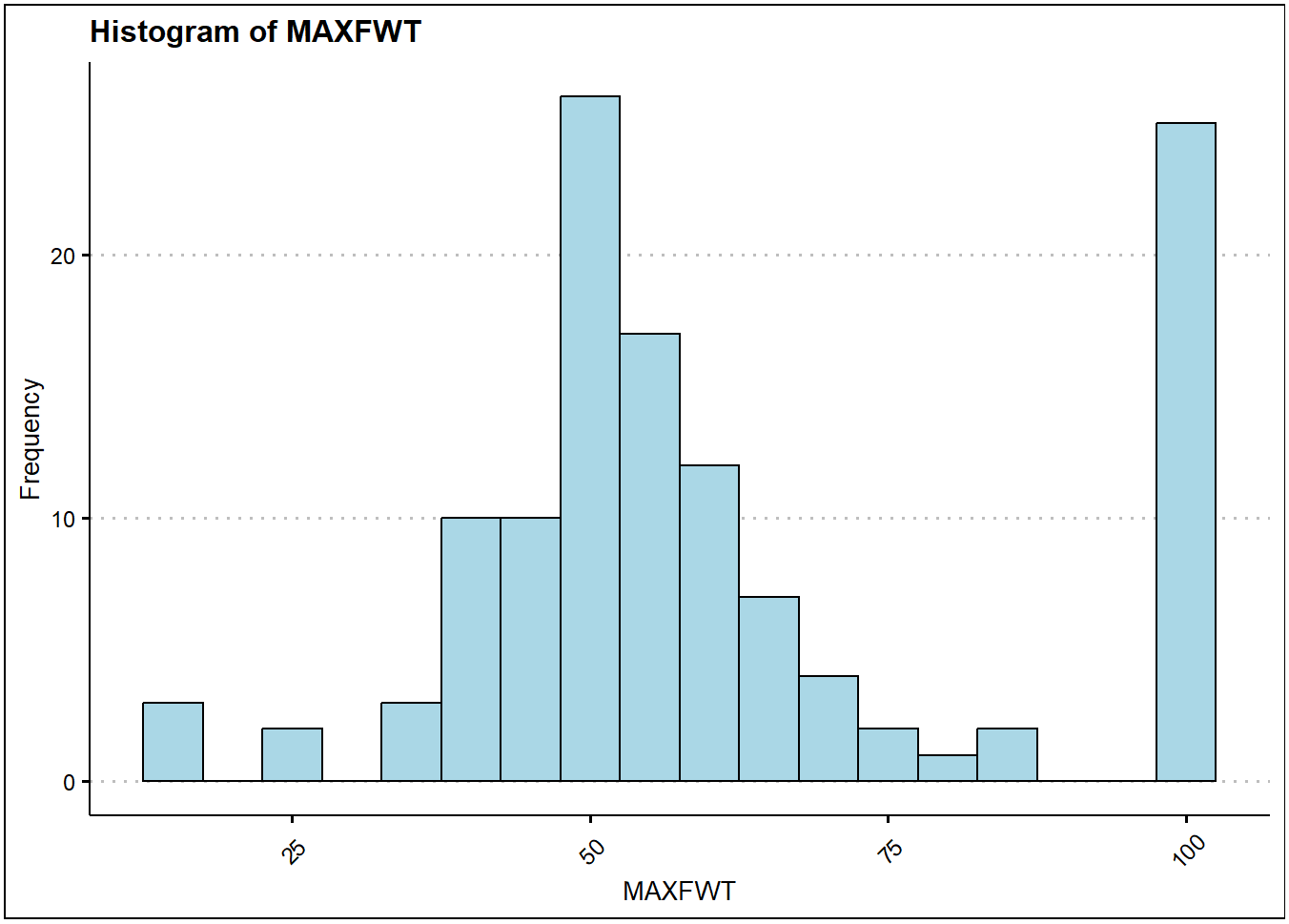
Yue Zhang

2024-11-19

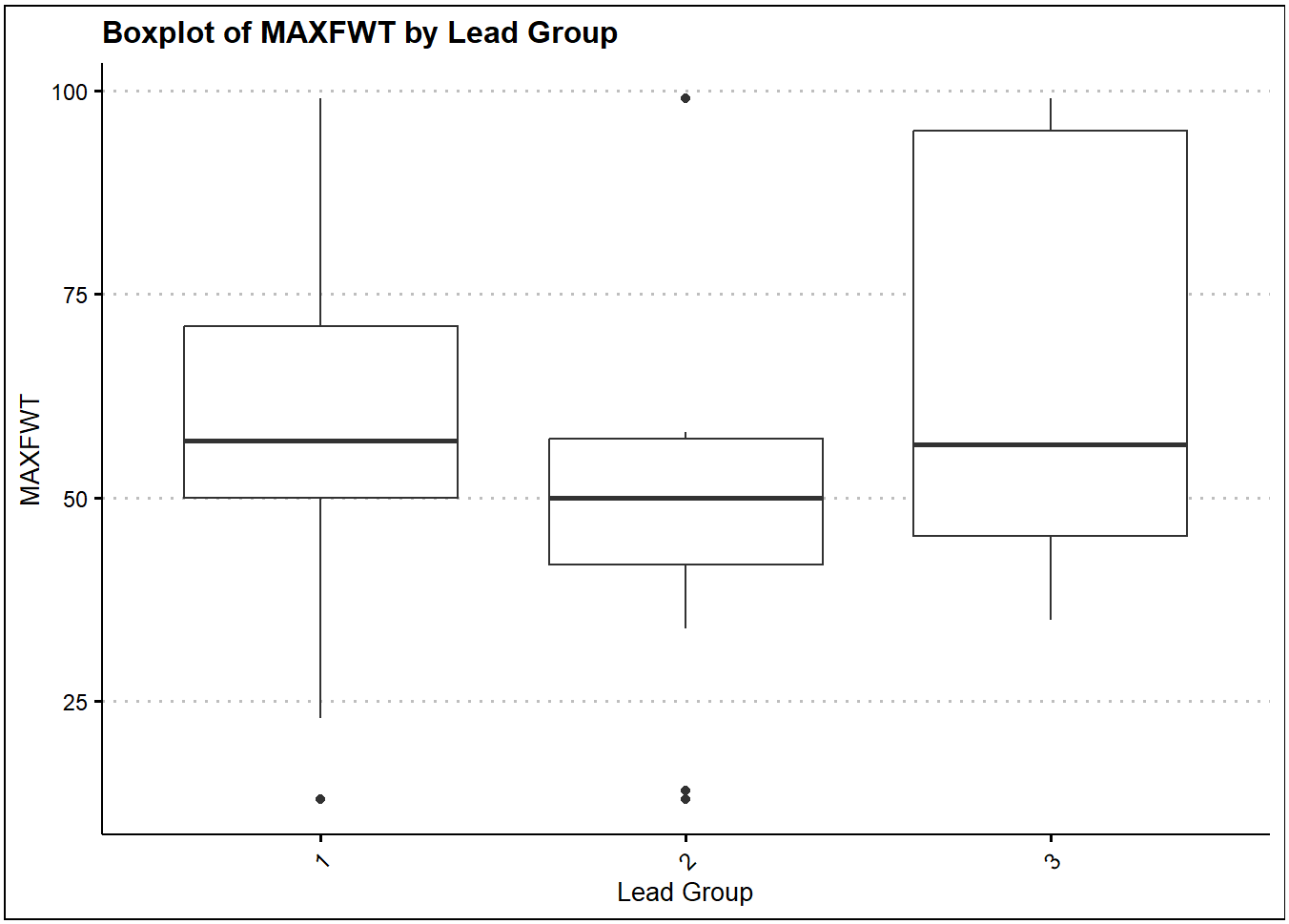
#12.44

```
lead = read_dta("E:/Biostat/Biostatistics/PHL_1700/Data/Raw/LEAD-1.DAT.dta")

# Plot histogram of MAXFWT
ggplot(lead, aes(x = maxfwt)) + geom_histogram(binwidth = 5,
    color = "black", fill = "lightblue") + labs(title = "Histogram of MAXFWT",
    x = "MAXFWT", y = "Frequency")
```




```
# Boxplot of MAXFWT by exposure group (LEAD_GRP)
ggplot(lead, aes(x = as.factor(lead_grp), y = maxfwt)) + geom_boxplot() +
    labs(title = "Boxplot of MAXFWT by Lead Group", x = "Lead Group",
        y = "MAXFWT")
```




```
# Shapiro-Wilk test for normality Null hypothesis: the
# MAXFWT is normally distributed Alternative hypothesis:
# the MAXFWT is not normally distributed
shapiro.test(lead$maxfwt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lead$maxfwt
## W = 0.88286, p-value = 1.911e-08
```

The histogram is skewed and it appears to be not normally distributed and symmetric. The boxplot shows that there are few outliers. Also, the p-value for Shapiro-Wilk test is less than 0.05, indicating the null hypothesis is rejected, thus the distribution of MAXFWT is not normal and non-parametric methods are needed.

```
kruskal_test = kruskal.test(maxfwt ~ as.factor(lead_grp), data = lead)
print(kruskal_test)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  maxfwt by as.factor(lead_grp)
## Kruskal-Wallis chi-squared = 4.2895, df = 2, p-value = 0.1171
```

Null hypothesis: there's no difference in the medians of MAXFWT across different lead groups Alternative hypothesis: there's difference in the medians of MAXFWT across different lead groups Since the p-value is 0.117 which is larger than 0.05, we fail to reject the null hypothesis and there's insufficient evidence to conclude that there's difference across each lead group.

#Required Additional Problems ##1 ###(a)

```
# ANOVA
anova_result = aov(iqf ~ as.factor(lead_grp), data = lead)
summary(anova_result)
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(lead_grp)   2    711   355.4   1.734  0.181
## Residuals           121  24808   205.0
```

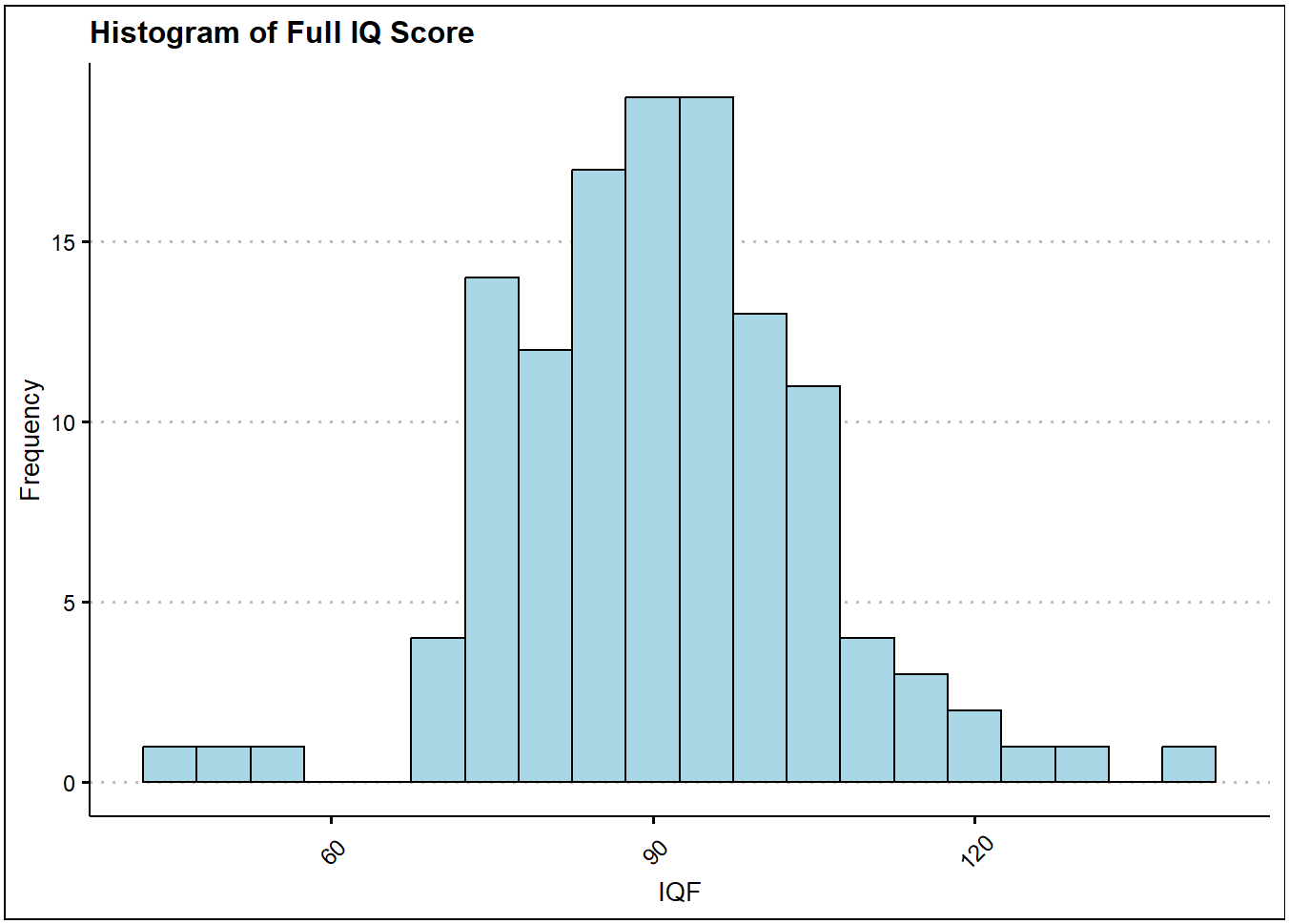
```
# Kruskal-Wallis
kruskal_result = kruskal.test(iqf ~ as.factor(lead_grp), data = lead)
print(kruskal_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  iqf by as.factor(lead_grp)
## Kruskal-Wallis chi-squared = 3.3578, df = 2, p-value = 0.1866
```

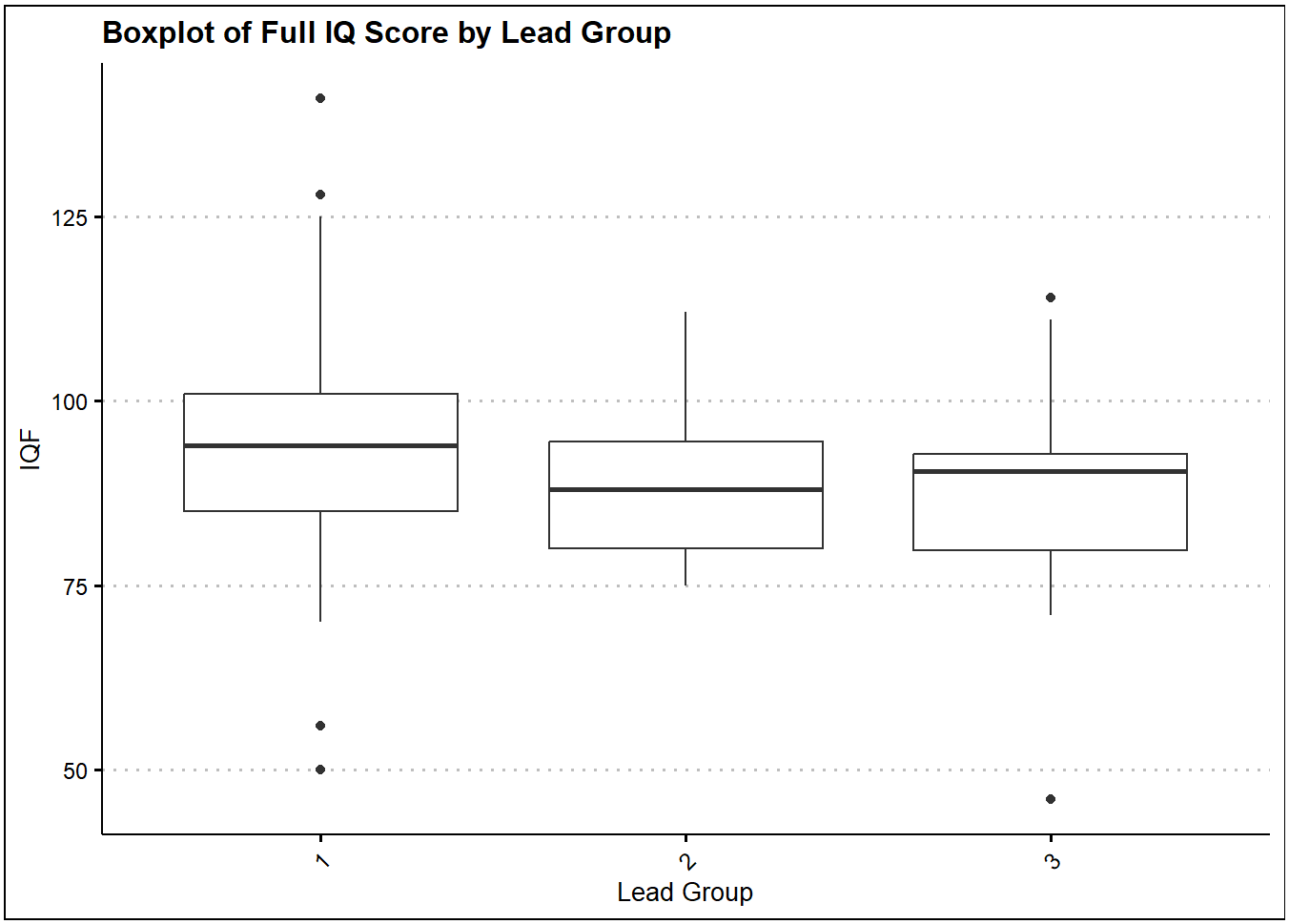
Null hypothesis: there's no difference in full IQ score across the lead groups Alternative hypothesis: there's difference in full IQ score across the lead groups For one-way ANOVA method, the F value is 1.734 with p-value equal to 0.181 which is higher than 0.05. Therefore, we fail to reject the null hypothesis. For the non-parametric method, the chi-squared is 3.3578 with p-value equal to 0.1866 which is higher than 0.05. Therefore, we also fail to reject the hypothesis. Thus, there's insufficient evidence to conclude that there's difference in full IQ score across each lead group.

###(b)

```
# Plot histogram of full IQ score
ggplot(lead, aes(x = iqf)) + geom_histogram(binwidth = 5, color = "black",
    fill = "lightblue") + labs(title = "Histogram of Full IQ Score",
    x = "IQF", y = "Frequency")
```




```
# Boxplot of MAXFWT by exposure group (LEAD_GRP)
ggplot(lead, aes(x = as.factor(lead_grp), y = iqf)) + geom_boxplot() +
    labs(title = "Boxplot of Full IQ Score by Lead Group", x = "Lead Group",
        y = "IQF")
```




```
# Shapiro-Wilk test for normality Null hypothesis: the iqf
# is normally distributed Alternative hypothesis: the iqf
# is not normally distributed
shapiro.test(lead$iqf)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lead$iqf
## W = 0.97722, p-value = 0.03412
```

The p-value of Shapiro-Wilk test is 0.034 which is less than 0.05, indicating we can reject the null hypothesis. This suggests that full IQ score is not normally distributed. Therefore, the non-parametric method would be more preferable.

###(c) Since the p-value of the Kruskal-Wallis test is larger than 0.05, we do not need to perform the pairwise comparisons because the test did not find significant differences across each group.

##2 ###(a)

```
icu = read.csv("E:/Biostat/Biostatistics/PHL_1700/Data/Raw/icu-1.csv")
icu_modle = glm(STA ~ AGE + SEX + CPR, family = binomial(link = "logit"),
    data = icu)
summary(icu_modle)
```

```
##
## Call:
## glm(formula = STA ~ AGE + SEX + CPR, family = binomial(link = "logit"),
##     data = icu)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.33063    0.75092  -4.435 9.19e-06 ***
## AGE          0.03026    0.01132   2.672  0.00754 **
## SEX         -0.16957    0.38848  -0.436  0.66248
## CPR          1.82940    0.61847   2.958  0.00310 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 183.76  on 196  degrees of freedom
## AIC: 191.76
##
## Number of Fisher Scoring iterations: 5
```

The prediction equation is:
$logit(P(STA = 1)/(1 - P(STA = 0))) = -3.331 + 0.030 * age - 0.170 * sex + 1.829 * cpr$

###(b)

```
icu_modle0 = glm(STA ~ 1, family = binomial(link = "logit"),
    data = icu)
lrtest.default(icu_modle, icu_modle0)
```

```
## Likelihood ratio test
##
## Model 1: STA ~ AGE + SEX + CPR
## Model 2: STA ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4  -91.88
## 2   1 -100.08 -3 16.401  0.0009385 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis: there's no relationship between the predictors and mortality in the ICU Alternative hypothesis: at least one of the predictor is significantly related to the mortality in the ICU The chi-squared is 16.401 with p-value less than 0.05. Therefore, we can reject the null hypothesis indicating the model is significant.

###(c)

```
exp(0.03026)
```

```
## [1] 1.030722
```

Null hypothesis: age has no effect on the mortality in ICU ($\beta_{age} = 0$) Alternative hypothesis: age has an effect on the mortality in ICU ($\beta_{age} \neq 0$) The z value is 2.672 with a p-value equals to 0.0075 which is less than 0.05. Therefore, we can reject the null hypothesis indicating age is a significant risk factor. For every year increases in age, the risk of dying in the ICU will increase by 3.07%.

###(d)

```
exp(-0.16957)
```

```
## [1] 0.8440277
```

Null hypothesis: sex has no effect on the mortality in ICU ($\beta_{sex} = 0$) Alternative hypothesis: sex has an effect on the mortality in ICU ($\beta_{sex} \neq 0$) The z value is -0.436 with a p-value equals to 0.6625 which is larger than 0.05. Therefore, we cannot reject the null hypothesis indicating sex is not a significant risk factor. The dying risk in the ICU for females is 0.844 times than men.

###(e)

```
exp(1.8294)
```

```
## [1] 6.230147
```

Null hypothesis: CPR has no effect on the mortality in ICU ($\beta_{CPR} = 0$) Alternative hypothesis: CPR has an effect on the mortality in ICU ($\beta_{CPR} \neq 0$) The z value is 2.958 with a p-value equals to 0.0031 which is less than 0.05. Therefore, we can reject the null hypothesis indicating CPR is a significant risk factor. The dying risk in the ICU for people who have CPR prior to ICU entry is 5.23 times than people who don’t have CPR prior to ICU.