Statistic_Assagnment _02

Done by Emmanuel NDAHIMANA and Zirhumanana BALIKE Dieudonne

October 31, 2018

Introduction

In this is study we are going to analyze the dataset contained in the file DonneesAnorexie.txt. The data we consider are from the persons suffering from anorexia and that had the different medical treatments.Some patients did a family therapy (FT), others a therapy called Cognitive Behavioural Treatment (CBT), and the remaining patients did not have any treatment and we refer to them as the control group (Cont). For each patient the table contains the type of therapy, and his/her weight before and after the therapy (weight is given in pounds).

Data Analysis

In this part we import date called DonneesAnorexie.txt with including two more columns for changing weight from pounds to kg.

We use function **head** to just take the first six samples, but in analysis does not change anything.

This is a table which contains the weight in pound and kg of the patients before and after getting therapy.

```
therapy<- read.table(file = 'DonneesAnorexie.txt',header = TRUE )
therapy$pre_in_kg<-0.453*therapy[,2]
therapy$post_in_kg<-0.453*therapy[,3]
head(therapy)
```

```
##   treatment  pre post pre_in_kg post_in_kg
## 1      Cont 80.7 80.2   36.5571    36.3306
## 2      Cont 89.4 80.1   40.4982    36.2853
## 3      Cont 91.8 86.4   41.5854    39.1392
## 4      Cont 74.0 86.3   33.5220    39.0939
## 5      Cont 78.1 76.1   35.3793    34.4733
## 6      Cont 88.3 78.1   39.9999    35.3793
```

```
attach(therapy)
```

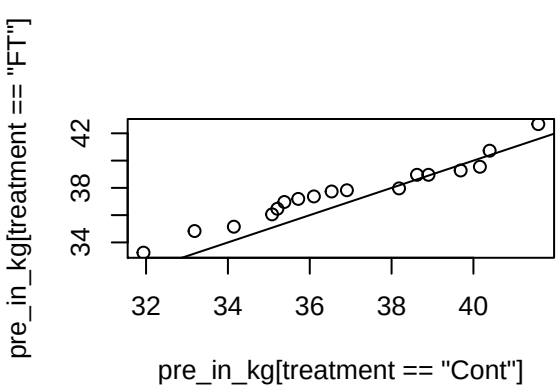
We want to check if the composition of the three groups of patients for the different treatments is done correctly. First of all, we compare the sample size of the three groups and their distribution. Here below, one can see the sample size and the graphs of the **QQ-plot** related to the three groups. In case our purpose is to verify if the the data of the three groups come from the same distribution.

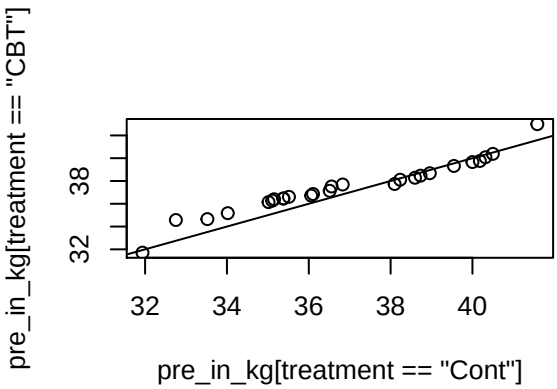
The figures below show that the three group have the same distribution since the points are around the straight line **x=y**.

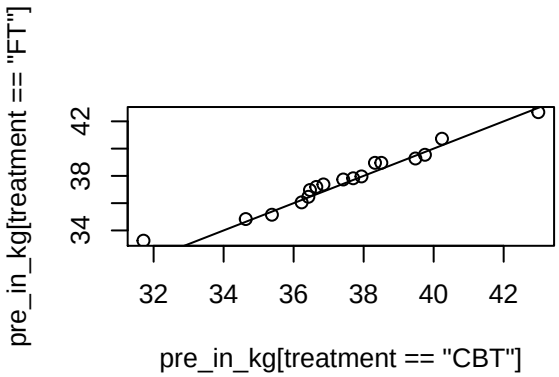
```
table(therapy$treatment)
```

```
##
##  CBT Cont   FT
##   29   26   17
```

```
par(mfrow=c(2,2))
qqplot(pre_in_kg[treatment=='Cont'],pre_in_kg[treatment=='FT'])
abline(0,1)
qqplot(pre_in_kg[treatment=='Cont'],pre_in_kg[treatment=='CBT'])
abline(0,1)
qqplot(pre_in_kg[treatment=='CBT'],pre_in_kg[treatment=='FT'])
abline(0,1)
```







To evaluate the success of a treatment, it is more relevant to analyze the variations of the weight during the entire study by analyzing the variations of the weight before and after treatment. The variation of the weight must be understood as difference between the weight of the patients after treatment and before treatment. The data below provide the difference of weight per groups and their relative means.

```
diff1<-post_in_kg[treatment=='Cont']-pre_in_kg[treatment=='Cont']
diff1
```

```
##  [1] -0.2265 -4.2129 -2.4462  5.5719 -0.9060 -4.6206 -5.5266  5.2548
##  [9] -3.2163  2.8086 -0.0906 -4.1676  3.7599  1.4949  5.1189  0.0000
## [17] -0.4530 -4.8018 -2.0838 -3.0351  1.2684  0.1359  0.8154  1.6761
## [25]  7.2027 -4.6206
```

```
mean(diff1)
```

```
## [1] -0.20385
```

```
diff2<-post_in_kg[treatment=='FT']-pre_in_kg[treatment=='FT']
diff2
```

```
##  [1]  5.1642  4.9830  2.4915  4.2582  6.1608 -1.3137 -0.0453  3.3522
##  [9]  9.7395 -2.4009 -1.7214  6.0702  5.9343  4.0770  1.7667  2.5821
## [17]  4.8471
```

```
mean(diff2)
```

```
## [1] 3.290912
```

```
diff3<-post_in_kg[treatment=='CBT']-pre_in_kg[treatment=='CBT']
diff3
```

```
##  [1]  0.7701  0.3171 -0.0453 -0.3171 -1.5855  6.7497  1.5855  7.7463
```

```
##  [9] -3.4428  0.7248  5.3001  2.7633  0.4983 -1.8120  9.4677 -4.1223
## [17]  0.9513 -0.6342  0.6342 -0.1359 -1.6761 -0.3624  1.0872  5.7078
## [25]  0.8607  1.7667  0.0453  6.9762 -0.3171
```

```
mean(diff3)
```

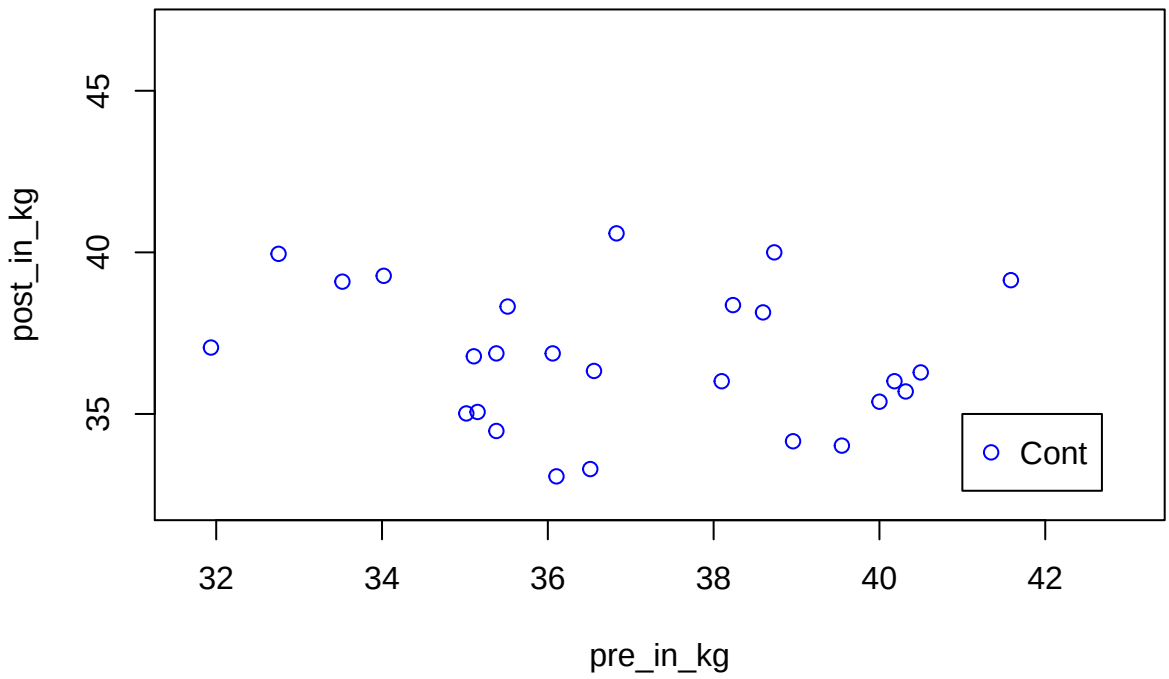
```
## [1] 1.362124
```

The control group ('Cont') is composed of patients that have not received any particular treatment or therapy,Now we are going to look how the weights of these patients evolve over time. The control group has small variation of weight and when we look in its summary we get negative change in first and third quatile,mean and maximum. Only minimum value of control group has positive change. The weights of these patients in control group has negative change over time.

```
vec1<-therapy$treatment=='Cont'
vec2<-therapy[vec1,(1:5)]
summary(vec2)
```

```
##  treatment      pre             post          pre_in_kg       post_in_kg
##  CBT : 0   Min.   :70.50   Min.   :73.00   Min.   :31.94   Min.   :33.07
##  Cont:26   1st Qu.:77.72   1st Qu.:77.58   1st Qu.:35.21   1st Qu.:35.14
##  FT  : 0   Median :80.65   Median :80.70   Median :36.53   Median :36.56
##            Mean   :81.56   Mean   :81.11   Mean   :36.95   Mean   :36.74
##            3rd Qu.:85.88   3rd Qu.:84.67   3rd Qu.:38.90   3rd Qu.:38.36
##            Max.   :91.80   Max.   :89.60   Max.   :41.59   Max.   :40.59
```

```
plot(pre_in_kg[treatment=='Cont'],post_in_kg[treatment=='Cont'],col = 'blue',type = 'p',xlab='pre_in_kg
legend(41,35,'Cont',col='blue',pch = 1)
```



To know the impact of the different therapies on the weight of the patients, we use the summary for every therapy and we use plotting of all therapies in the same plot and we look which ones has the big positive difference in weight from after and before the treatment.

```
vec1<-therapy$treatment=='Cont'
vec2<-therapy[vec1,(1:5)]
summary(vec2)
```

```
##  treatment      pre             post          pre_in_kg       post_in_kg
##  CBT : 0   Min.   :70.50   Min.   :73.00   Min.   :31.94   Min.   :33.07
##  Cont:26   1st Qu.:77.72   1st Qu.:77.58   1st Qu.:35.21   1st Qu.:35.14
##  FT  : 0   Median :80.65   Median :80.70   Median :36.53   Median :36.56
##            Mean   :81.56   Mean   :81.11   Mean   :36.95   Mean   :36.74
##            3rd Qu.:85.88   3rd Qu.:84.67   3rd Qu.:38.90   3rd Qu.:38.36
##            Max.   :91.80   Max.   :89.60   Max.   :41.59   Max.   :40.59
```

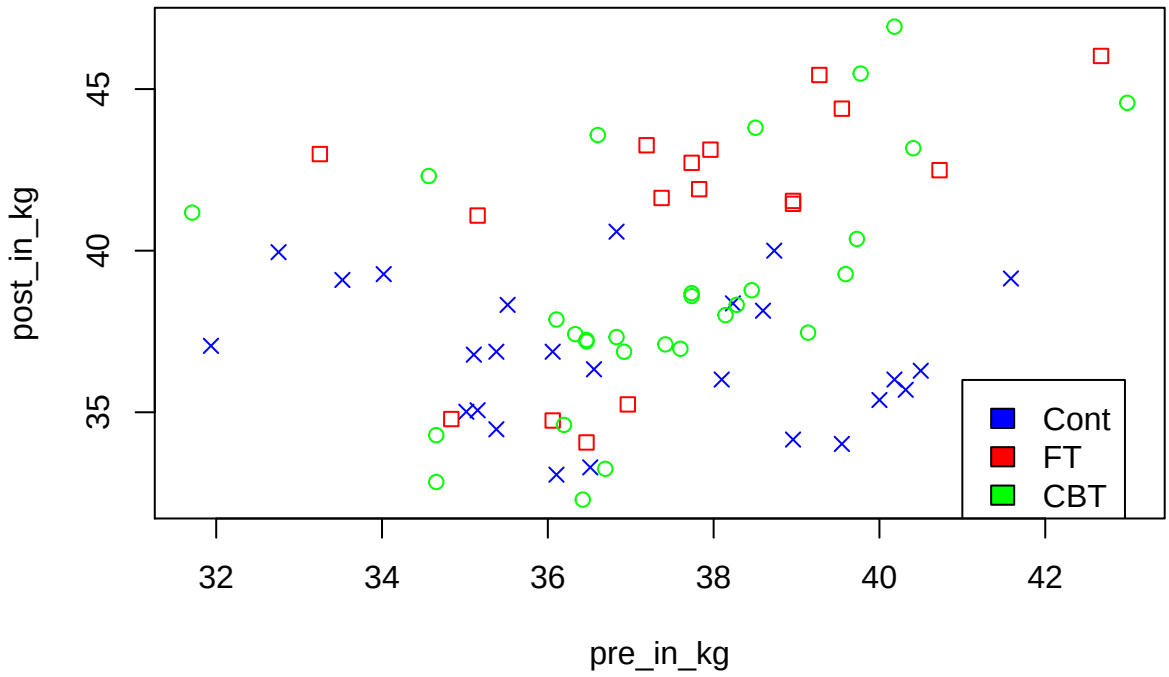
```
vec3<-therapy$treatment=='CBT'
vec4<-therapy[vec3,(1:5)]
summary(vec4)
```

```
##  treatment      pre             post          pre_in_kg       post_in_kg
##  CBT :29   Min.   :70.00   Min.   : 71.3   Min.   :31.71   Min.   :32.30
##  Cont: 0   1st Qu.:80.40   1st Qu.: 81.9   1st Qu.:36.42   1st Qu.:37.10
##  FT  : 0   Median :82.60   Median : 83.9   Median :37.42   Median :38.01
##            Mean   :82.69   Mean   : 85.7   Mean   :37.46   Mean   :38.82
##            3rd Qu.:85.00   3rd Qu.: 90.9   3rd Qu.:38.51   3rd Qu.:41.18
##            Max.   :94.90   Max.   :103.6   Max.   :42.99   Max.   :46.93
```

```
vec5<-therapy$treatment=='FT'
vec6<-therapy[vec5,(1:5)]
summary(vec6)
```

```
##  treatment      pre             post           pre_in_kg
##  CBT : 0   Min.   :73.40   Min.   : 75.20   Min.   :33.25
##  Cont: 0   1st Qu.:80.50   1st Qu.: 90.70   1st Qu.:36.47
##  FT  :17   Median :83.30   Median : 92.50   Median :37.73
##            Mean   :83.23   Mean   : 90.49   Mean   :37.70
##            3rd Qu.:86.00   3rd Qu.: 95.20   3rd Qu.:38.96
##            Max.   :94.20   Max.   :101.60   Max.   :42.67
##    post_in_kg
##  Min.   :34.07
##  1st Qu.:41.09
##  Median :41.90
##  Mean   :40.99
##  3rd Qu.:43.13
##  Max.   :46.02
```

```
plot(pre_in_kg[treatment=='Cont'],post_in_kg[treatment=='Cont'],col = 'blue',type = 'p',pch=4,xlab='pre
points(pre_in_kg[treatment=='FT'],post_in_kg[treatment=='FT'],pch=0,col = 'red',type = 'p',xlab='pre_in
points(pre_in_kg[treatment=='CBT'],post_in_kg[treatment=='CBT'],col = 'green',type = 'p')
legend(41,36,c('Cont','FT','CBT'),c(col='blue',col='red',col='green'))
```

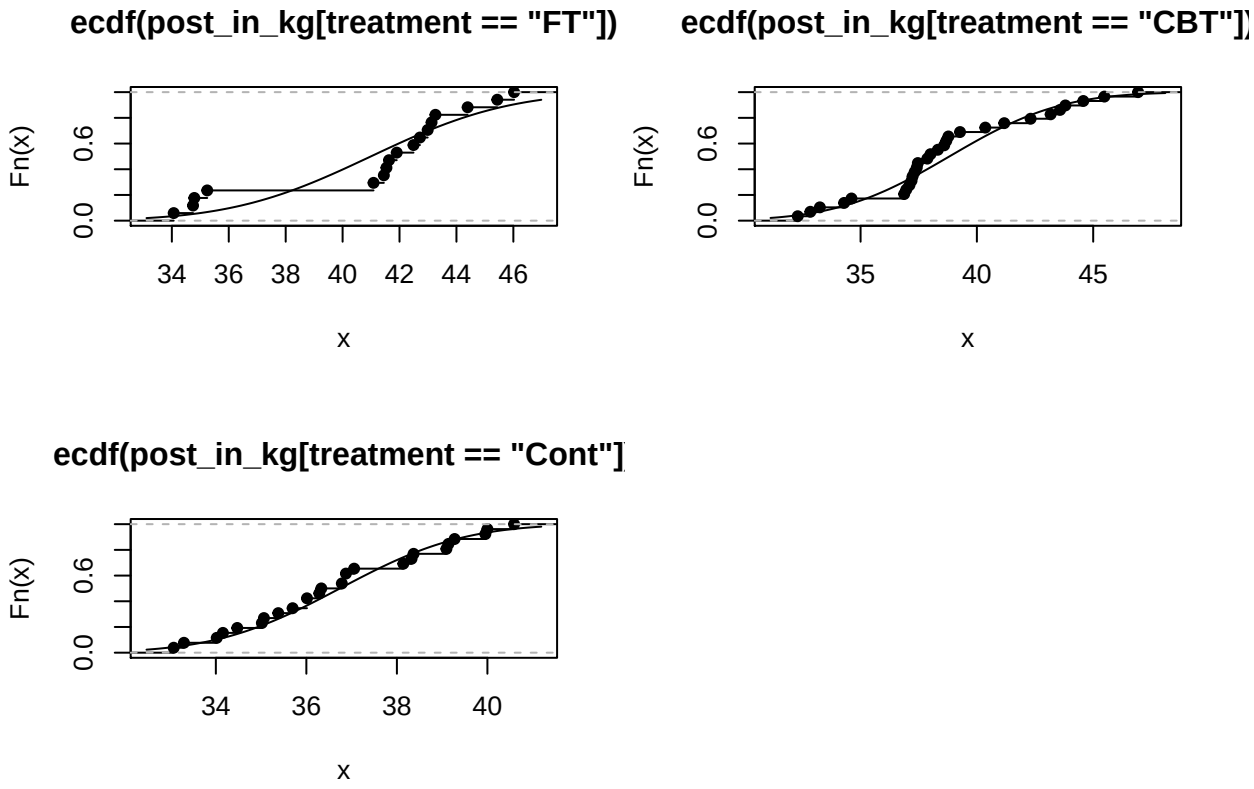


From the observations provided by the summary of each group situation and the plot of therapies, the treatment seems to be the best is family therapy **FT**.

We have to analyze the distribution of the different variables related to the weight per type of treatment.

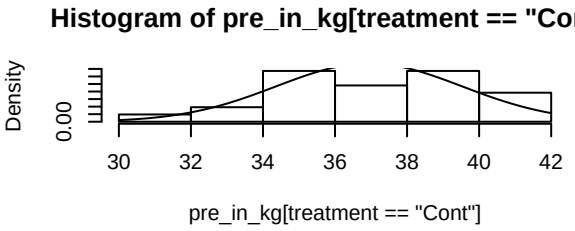
First of all we need to check if our datasets made of groups of patients are continuous or discrete. Below there are the empirical cummulative distribution function *ecdf* and one remarks that our distributions are continuous since the jumps are small.

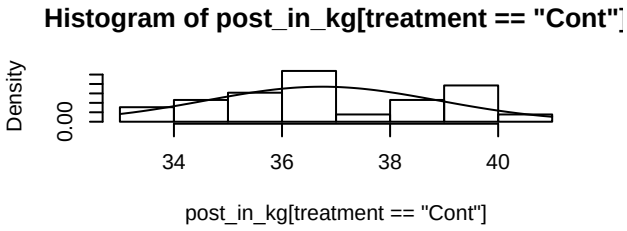
```
par(mfrow=c(2,2))
dist1<-ecdf(post_in_kg[treatment=='FT'])
plot(dist1)
curve(pnorm(x,mean(post_in_kg[treatment=='FT']),sd(post_in_kg[treatment=='FT'])), add=TRUE)
dist2<-ecdf(post_in_kg[treatment=='CBT'])
plot(dist2)
curve(pnorm(x,mean(post_in_kg[treatment=='CBT']),sd(post_in_kg[treatment=='CBT'])), add=TRUE)
dist3<-ecdf(post_in_kg[treatment=='Cont'])
plot(dist3)
curve(pnorm(x,mean(post_in_kg[treatment=='Cont']),sd(post_in_kg[treatment=='Cont'])), add=TRUE)
```

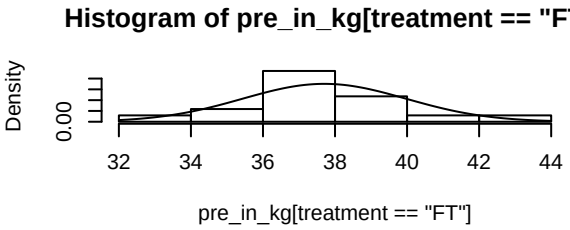


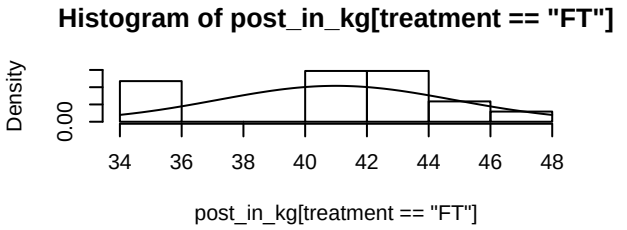
Since our data are continuous, we can use histogram to check if the data are normally distributed.

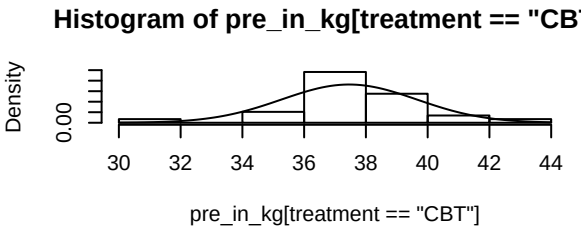
```
par(mfrow=c(3,2))
hist(pre_in_kg[treatment=='Cont'],freq = FALSE)
curve(dnorm(x,mean(pre_in_kg[treatment=='Cont']),sd(pre_in_kg[treatment=='Cont'])),add=TRUE)
hist(post_in_kg[treatment=='Cont'],freq = FALSE)
curve(dnorm(x,mean(post_in_kg[treatment=='Cont']),sd(post_in_kg[treatment=='Cont'])),add=TRUE)
hist(pre_in_kg[treatment=='FT'],freq = FALSE)
curve(dnorm(x,mean(pre_in_kg[treatment=='FT']),sd(pre_in_kg[treatment=='FT'])),add=TRUE)
hist(post_in_kg[treatment=='FT'],freq = FALSE)
curve(dnorm(x,mean(post_in_kg[treatment=='FT']),sd(post_in_kg[treatment=='FT'])),add=TRUE)
hist(pre_in_kg[treatment=='CBT'],freq = FALSE)
curve(dnorm(x,mean(pre_in_kg[treatment=='CBT']),sd(pre_in_kg[treatment=='CBT'])),add=TRUE)
hist(post_in_kg[treatment=='CBT'],freq = FALSE)
curve(dnorm(x,mean(post_in_kg[treatment=='CBT']),sd(post_in_kg[treatment=='CBT'])),add=TRUE)
```

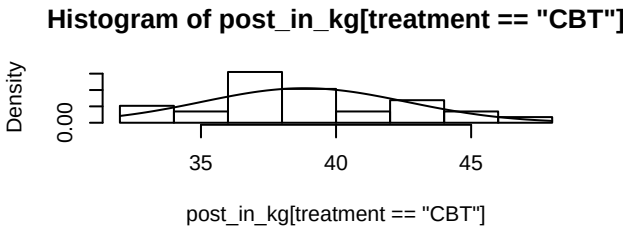












These are the boxplots of different variables related to the weight per type of treatment and they show that our data are not symmetric. This proves that it is not reasonable to model these variables by normal distributions

```
par(mfrow=c(2,2))
plot(pre_in_kg~treatment)
plot(post_in_kg~treatment)
```