A4 - Q2 (Ram Yoogesh Gopu - 20867060)

Reading the data

```
medicalRecords <- read.csv("medicalRecords.csv", header = TRUE)
#medicalRecords
# So there are 1600 rows of data with 5 columns (Age, sex, Treatment, outcome, FReq)
```

multi-way table

```
medicalRecordsTable <- xtabs(Freq ~ Age + Sex + Treatment + Outcome, data = medicalRecords)
dim(medicalRecordsTable)
```

```
## [1] 2 2 2 2
```

```
margin.table(medicalRecordsTable, margin = c(2))
```

```
## Sex
## Female   Male
##   1600   1600
```

```
margin.table(medicalRecordsTable, margin = c(3))
```

```
## Treatment
##    A    B
## 1600 1600
```

```
margin.table(medicalRecordsTable, margin = c(4))
```

```
## Outcome
##      Died Recovered
##      1440      1760
```

```
margin.table(medicalRecordsTable, margin = c(1))
```

```
## Age
## 20-39 40-59
##  1600  1600
```

(A)

(i)

Units -> People in the dataset form the units. Target population - The dataset which contains the data about the outbreak.

(ii)

From the above dataset, the variates are Age, Sex, Treatment, Outcome, Freq. Outcome is considered as a response variate. Explanatory variates are Age, Sex, Treatment, Freq.

(iii)

For the above scenario, i would say Treatment and Age would be of best interest.

(B)

(i)

Dataset which constitutes the data about the patients(outcome status with respect to different treatments that they have been provided with) which is elicited from various countries is called as the study population. The error might increase if the data is been generated only from a specific country. Eg - People in africa have compromised immune systems due to malnutriotion. So they might need a severe treatment , Say B. If we are going to treat people in Canada, who are usually more healthy using B, then it might result in various consequences. SO we need to focus on more generalised dataset.

(ii)

On a general note, sampling plan provides an outline on which the research is conducted. It gives us information on which category we should survey, what should be the sample size, and also how to choose the respondents from the population.

We need to make sure that there is efficient balance in the sampling data. (Focusing only on specific country)As explained in (B)(i), it will increase the sampling bias. So we must focus on balanced dataset. Eg - eliciting the data from the different countries and picking up the samples of equal propotions of men, women and so on.

(iii)

I would say that the above study is observational, because we observe the recovery of people from the disease in different countries when provided with different treatments.

(C)

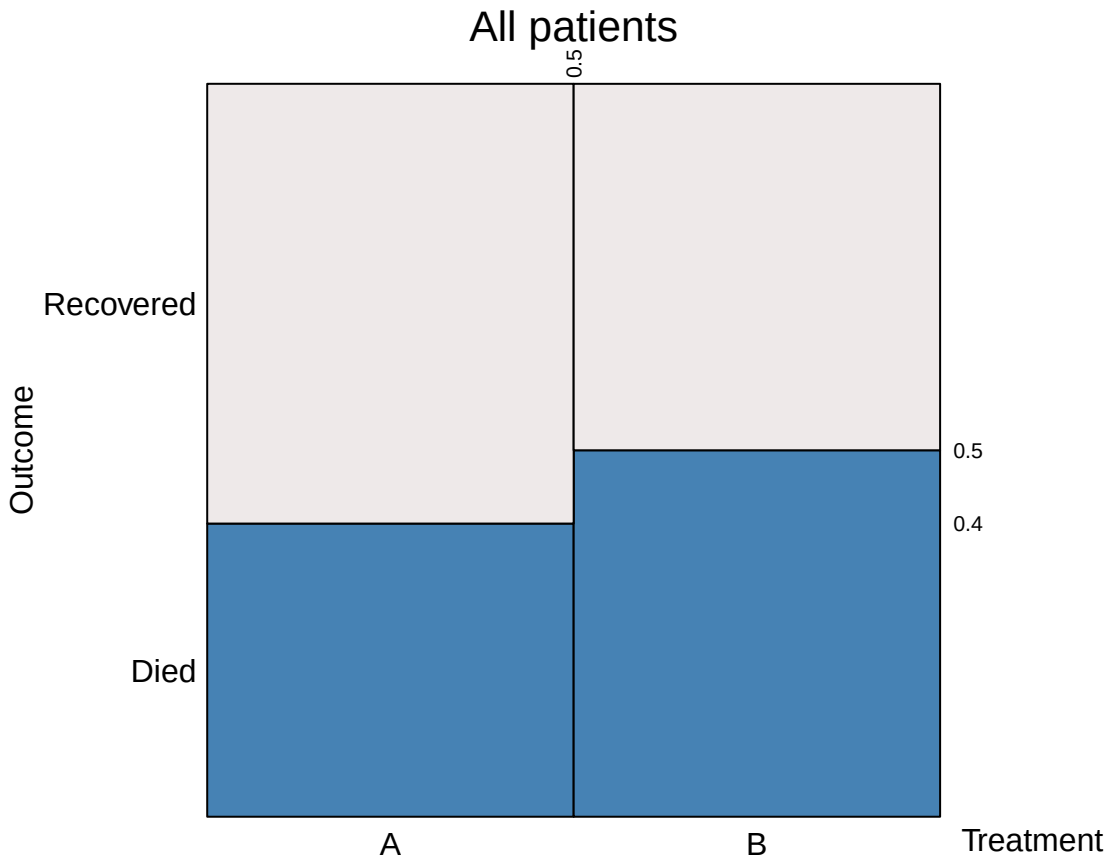
(i)

The above dataset (MedicalRecords) is a sample. This contains an equal rows of men and women (1600 & 1600 observational studies). Error might increase when we lose the balance, and we end up making the wrong analysis.

(d)

(i)

```
library(eikosograms)
eikos("Outcome", "Treatment", data = medicalRecordsTable, main = "All patients")
```



(ii)

```
martable <- margin.table(medicalRecordsTable, margin = c(3,4))
chisq.test(martable)
```

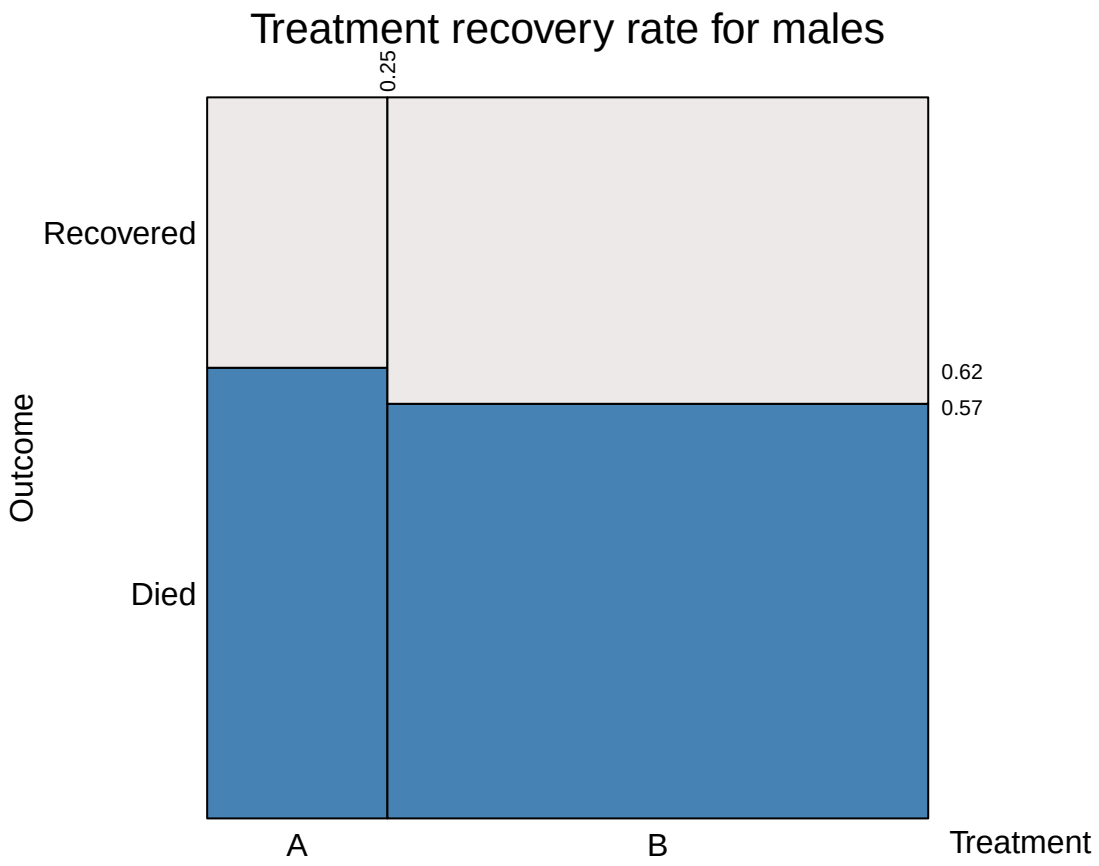
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  martable
## X-squared = 31.92, df = 1, p-value = 1.606e-08
```

```
print(martable)
```

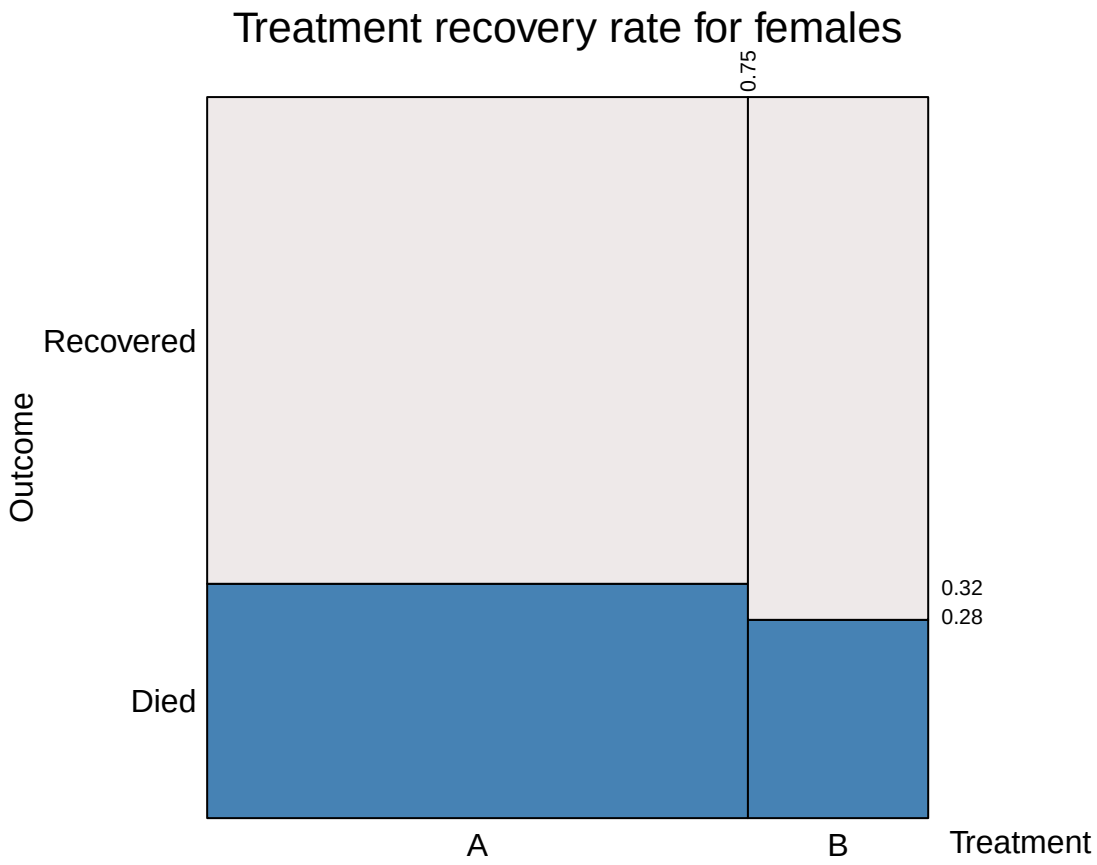
```
##           Outcome
## Treatment Died Recovered
##         A  640       960
##         B  800       800
```

We can conclude that the Outcome is related to the treatment, since the p-value is really low. Also, from the printed value, it is evident that the treatment has few deaths and higher recovery rate. So i would choose Treatment A.

```
# Doing the below harcoding to get the variates data (Age, Sex, Treatment, Outcome, Freq) with respect
maleresponse <- xtabs(Freq ~ Age + Sex + Treatment + Outcome, data = medicalRecords[medicalRecords$Sex =
femalerespose <- xtabs(Freq ~ Age + Sex + Treatment + Outcome, data = medicalRecords[medicalRecords$Sex

eikos("Outcome", "Treatment", data = maleresponse, main = "Treatment recovery rate for males")
```




```
eikos("Outcome", "Treatment",  data = femalerespose ,main="Treatment recovery rate for females")
```

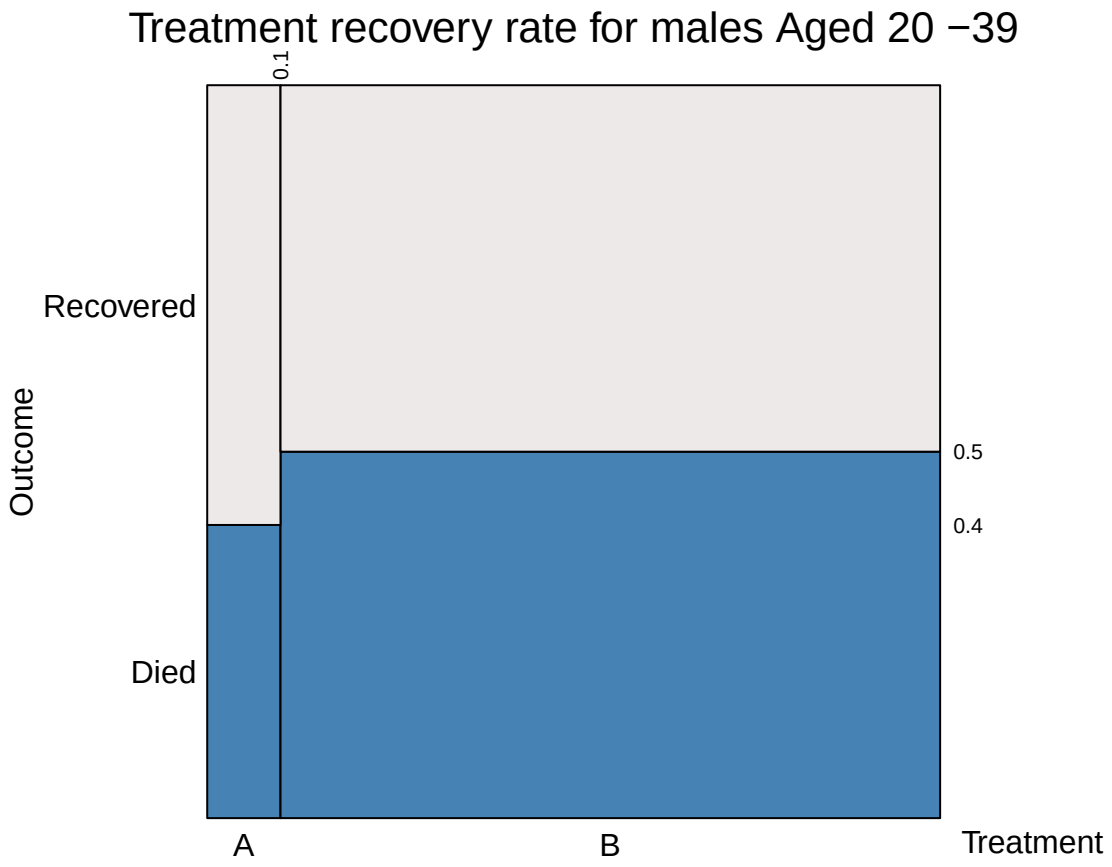


From the graphs, we can conclude that females have a higher rate of recovery.

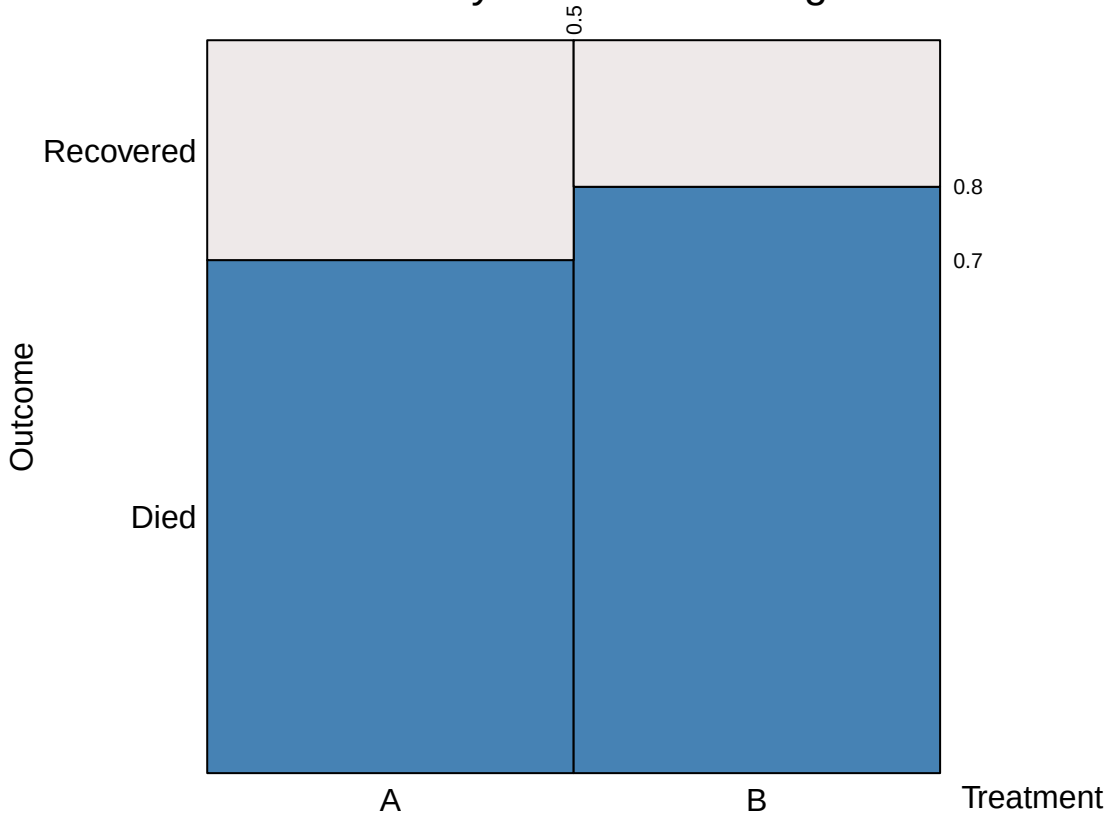
Treatment B is suggested for both males and females since it has higher recovery rate.

(iv)

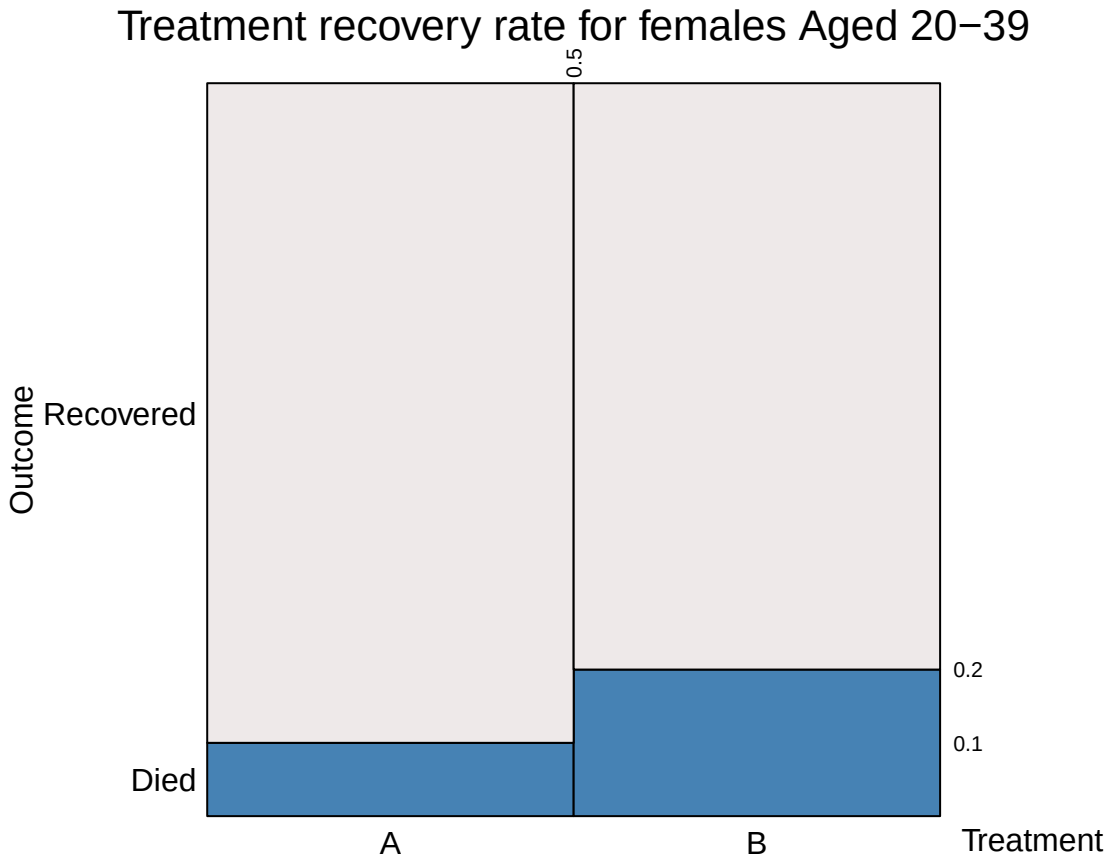
```
## Hardcoding like the previous example in order to plot the ekios

maleres2039 <- xtabs(Freq ~ Age + Sex + Treatment + Outcome, data = medicalRecords[ (medicalRecords$Sex
maleres4059 <- xtabs(Freq ~ Age + Sex + Treatment + Outcome, data = medicalRecords[ (medicalRecords$Sex

femaleres2039 <- xtabs(Freq ~ Age + Sex + Treatment + Outcome, data = medicalRecords[ (medicalRecords$S
femaleres4059 <- xtabs(Freq ~ Age + Sex + Treatment + Outcome, data = medicalRecords[ (medicalRecords$S


eikos("Outcome", "Treatment", data = maleres2039, main = "Treatment recovery rate for males Aged 20 -39
```



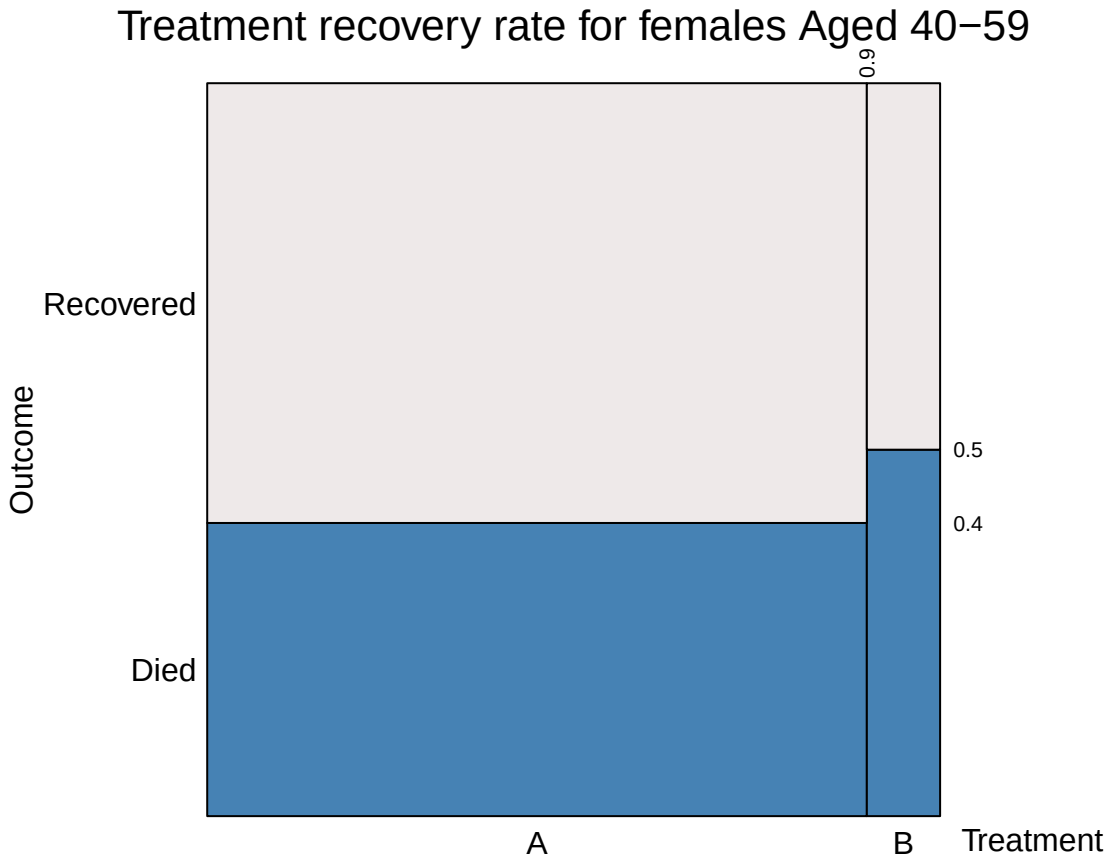

```
eikos("Outcome", "Treatment", data = maleres4059, main = "Treatment recovery rate for males Aged 40 -59
```




```
eikos("Outcome", "Treatment",  data = femaleres2039 ,main="Treatment recovery rate for females Aged 20-3
```




```
eikos("Outcome", "Treatment",  data = femaleres4059 ,main="Treatment recovery rate for females Aged 40-5
```



Males (20-39) (Young Men) - treatment A (It has high recovery rate). Males (40-59)(Older Men) - treatment A (It has high recovery rate). Females (20-39)(Young Women) - treatment A (It has high recovery rate) . Females (40-59)(Older women) - treatment A (It has high recovery rate)

(3)

We can't come up with an binary answer for the above scenario. It is mainly because, we get different asnwers when we perform different methods. In order to come up with more robust conclusions, we need to have more variates (Patients underlying heath conditions such as obesity, heart disease, previous history of diseases, lipids count, RBC & WBC counts and so on). But if its mandatory for me to pick any of the above two mentioned treatments, i would choose the treatment A.