Analysis for clustering cervical cancer survivors symptoms after radiotherapy with different methods and a comparison of the results

Christine Søjberg, Heidi Lunde and Rasmus Hanghøj

31/5-2022

Intro

In this project we look at, which symptoms groups together during the time after ended radiotherapy treatment for cervical cancer. We focus on month 6, 12 and 24 after ended treatment. We want to analyze different cluster methods, to find the method that performs the best. To analyze the data, we have used three different methods: Exploratory Factor Analysis (EFA), Hierarchical Clustering Analysis (HCA) and Principal Component Analysis (PCA). We will compare the methods to see which methods, that perform best.

Data formating

```
# Read Data
Data <- read_excel("DataProject_onlyvariables.xlsx")

Physician <- Data[25:376]

#Making all -1 to NA's
Physician[Physician == -1] <- NA

# unpivot dates
Unpivot_data <- Data %>%
  pivot_longer(
    cols = starts_with("CTCAEA"),
    names_to = "Periods",
    values_to = "Dates"
  )
PivotData_dates <- Unpivot_data %>%
  select(ID_Pat,
         Periods,
         Dates) %>%
  separate(col = Periods, into = c("var", "series"), sep = "_")

#Unpivot all symptoms
PivotData_sideeffects <- Data %>%
  select(-starts_with("CTCAEA")) %>%
  pivot_longer(cols = -c(ID_Pat,CentreID_Pat,
```

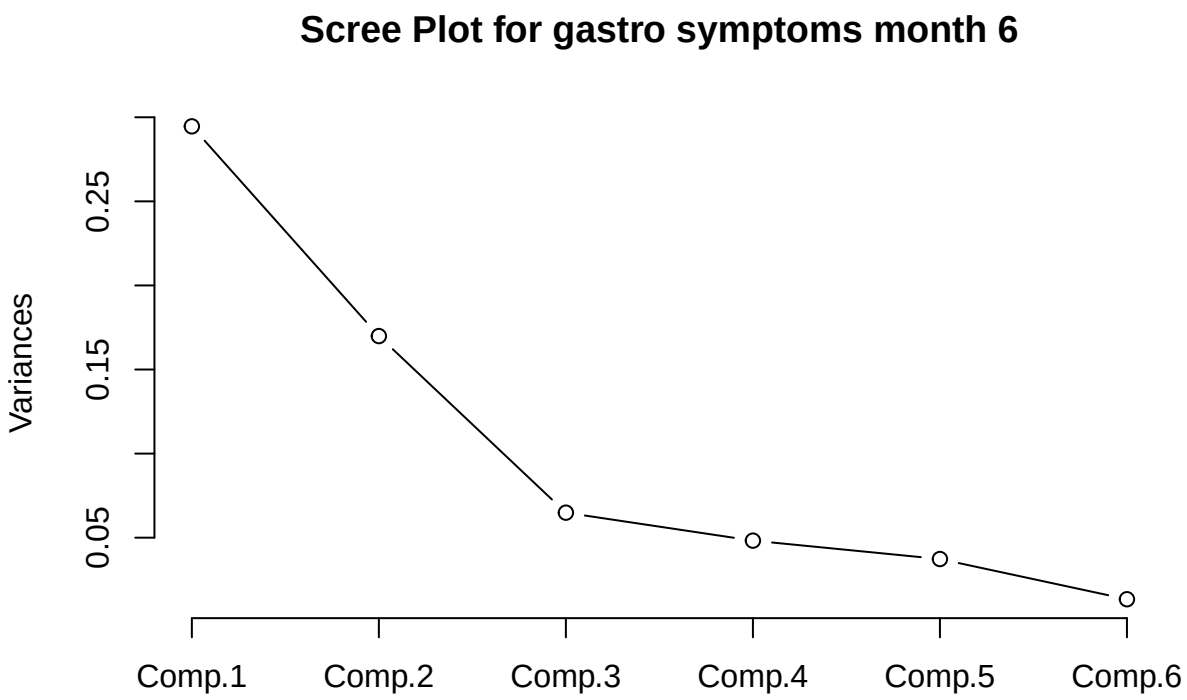
```
                          Age,BMI_categories,
                          Smoking,
                          ChronicDiseases,
                          FIGO_2009,
                          EndTreatmentDate)) %>%
  separate(col = name, into = c("var", "series"), sep = "_") %>%
  pivot_wider(id_cols = c( c(ID_Pat,CentreID_Pat,
                          Age,BMI_categories,
                          Smoking,
                          ChronicDiseases,
                          FIGO_2009,
                          EndTreatmentDate)
                          , series),
              names_from = "var",
              values_from = "value")

#Join the two unpivoted frames
PivotData <- PivotData_dates %>%
  select(-var) %>%
  left_join(PivotData_sideeffects, by = c("ID_Pat", "series")) %>%
  relocate(c(series,Dates), .after = EndTreatmentDate)

#remove -1 and large incorrect valuse for all symptoms coloumns
for(i in 11:77) {
  PivotData[i] <- replace(PivotData[i], PivotData[i]<0, NA)
  PivotData[i] <- replace(PivotData[i], PivotData[i]>8, NA)
}
```

Analysis

EFA - Exploratory Factor Analysis

To make an EFA we use the function factanal(), that calculates the loadings with maximum likelihood method. factanal() requires an estimate of the number of factors, this is a tricky aspect of factor analysis. To suggest a number of factors we use a scree plot (elbow plot). A scree plot that uses a PCA approach, shows the y-axis as the percentage of total variance explained by each individual principal component and the x-axis as the number of components/factors. Below is shown a scree plot of the symptoms for the gastro organ in month 6. We see a difference in level at component 3, which suggests 3 factors for our data. The method is the same for the other organs and suggests 2 or 3 factors, but for all symptoms it suggests 6 factors.

```
gastro_symptoms6m<-Physician[,c(3,19,35,51,67,83)]



fit <- princomp(na.omit(gastro_symptoms6m), cor=FALSE)
plot(fit,type="lines", main = "Scree Plot for gastro symptoms month 6")
```



For our data we use 3 factors for the gastro and bladder organ, because the results makes more sense than with 2 factors. For the vagina organ we use 2 factors because there is not enough symptoms to use 3 factors. Lastly we use 6 factors for all symptoms, based on a scree plot result as described previously. Now we can use the function factanal(). We start off by looking at the symptoms for the gastro organ in month 6.

```
EFA_model_gastro6m <- factanal(na.omit(Physician[,c(3,19,35,51,67,83)]), factors=3)
```

The factor analysis creates a linear combination of factors to abstract the variable's underlying communality. The variability in our data is given by Σ and its estimate $\hat{\Sigma}$, which is given by the factor analysis model:

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Psi}$$

where $\hat{\Lambda}\hat{\Lambda}^T$ is the communality and $\hat{\Psi}$ is the uniqueness/noise. Below the uniqueness is shown.

```
EFA_model_gastro6m$uniquenesses
```

```
##       GastroDiarrhea_6M     GastroFlatulence_6M   GastroIncontinence_6M
##               0.6967676               0.6533089               0.7520814
##      GastroProctitis_6M GastroBleedingRectum_6M  GastroFistulaRectum_6M
##               0.7136840               0.0050000               0.8778073
```

A high uniqueness indicates that the factors does not account well for its variance. In our case 3 factors is the number of factors that gives the lowest uniqueness for the variables.

Now we move on to the loadings, which is the $\hat{\Lambda}$ in the above equation. Variables with a high loading is well explained by the factor. We use the squared loadings to calculate the communality.

```
apply(EFA_model_gastro6m$loadings^2,1,sum)
```

```
##       GastroDiarrhea_6M     GastroFlatulence_6M   GastroIncontinence_6M
##               0.3032322               0.3466912               0.2479186
##      GastroProctitis_6M GastroBleedingRectum_6M  GastroFistulaRectum_6M
##               0.2863162               0.9950001               0.1221927
```

We now calculate the factor analysis model

```
Lambda <- EFA_model_gastro6m$loadings
Psi <- diag(EFA_model_gastro6m$uniquenesses)
S <- EFA_model_gastro6m$correlation
Sigma <- Lambda %*% t(Lambda) + Psi
```

We compare the fitted correlation matrix (Sigma) from the observed correlation matrix (S) and round to 5 digits.

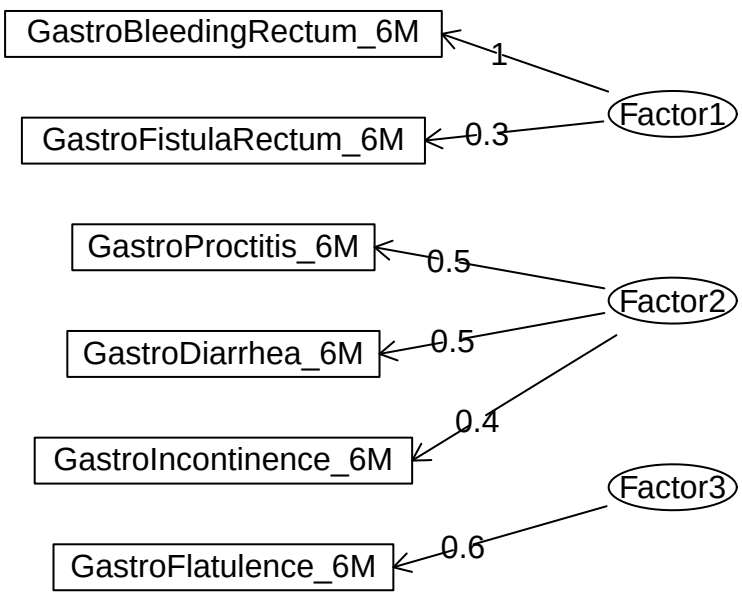
```
round(S - Sigma, 5)
```

```
##                          GastroDiarrhea_6M GastroFlatulence_6M
## GastroDiarrhea_6M                  0.00000            -0.00037
## GastroFlatulence_6M               -0.00037             0.00000
## GastroIncontinence_6M             -0.00010             0.00049
## GastroProctitis_6M                -0.00421             0.00016
## GastroBleedingRectum_6M            0.00005             0.00000
## GastroFistulaRectum_6M            -0.02390             0.00123
##                          GastroIncontinence_6M GastroProctitis_6M
## GastroDiarrhea_6M                      -0.00010           -0.00421
## GastroFlatulence_6M                     0.00049            0.00016
## GastroIncontinence_6M                   0.00000            0.00469
## GastroProctitis_6M                      0.00469            0.00000
## GastroBleedingRectum_6M                -0.00005            0.00000
## GastroFistulaRectum_6M                  0.02575            0.00028
##                          GastroBleedingRectum_6M GastroFistulaRectum_6M
## GastroDiarrhea_6M                          5e-05               -0.02390
## GastroFlatulence_6M                        0e+00                0.00123
## GastroIncontinence_6M                     -5e-05                0.02575
## GastroProctitis_6M                         0e+00                0.00028
## GastroBleedingRectum_6M                    0e+00                0.00002
## GastroFistulaRectum_6M                     2e-05                0.00000
```

The output is called the residual matrix. When numbers are close to 0 it indicates that our factor model is a good representation of the underlying concept.

With the function fa.diagram() the variables correlation to the factors is visualized, and the loadings is used to group the symptoms for the different organs and periods. Below is an example of the organ gastro in month 6.

```
EFA_model_gastro6m <- factanal(na.omit(Physician[,c(3,19,35,51,67,83)]), factors=3)
fa.diagram(EFA_model_gastro6m$loadings, main = "Gastro 6 months after treatment")
```



In this diagram we see that [GastroFistulas, GastroBleeding] is correlated to factor 1. [GastroProctitis, GastroDiarrhea, GastroIncontinence] is correlated to factor 2 and lastly GastroFlatulence is correlated to the third and last factor.

HCA - Hierachical Clustering Analysis

We have choosen the clustering method called hierachical clustering. This method groups the similar symptoms together, and shows the results in a dendogram. In hierachical clustering there is different ways to build the dendogram; 1) Divisive and 2) Agglomerative. Divisive start from the root with all "leaves" together and from that, seperate the leaves into groups. Then the groups will be seperated into smaller groups and so on, until there is only the single "leaf" left in each branch.

Agglomerative starts from the bottom, with all leaves seperated and groups the "leaves" together, until all of them is connected. We have choosen to use the agglomerative method for our case because we want to find the symptoms which occur most frequently together.

Furthermore, we have choosen the linkage method "complete", because we want to separate the symptoms, which are least correlated to each other, so we get clusters of the most similar symptoms.
But first we will look at the math behind the method will be shown.

To prepare the data for the hierachical clusteringt analysis, we first need to make the dissimilarity matrix. We do that by making a distance matrix, because all distance matrices are dissimilarity matrices:

$$A = \begin{pmatrix} 0 & d_{12}^2 & . & . & . \\ d_{21}^2 & 0 & . & . & \mathring{u} \\ d_{31}^2 & .. & .. & & . \\ ... & . & . & 0 & . \\ d_{n1}^2 & . & . & . & 0 \end{pmatrix}$$

Each of the entities will be calculated with the formula below, which is the formula for the euclidean distance:

$$d_{ij} = \sqrt{(x_1 - y_1) + (x_2 - y_2) + ... + (x_n - yn)}$$

Where x is one of the symptoms and y is another symptom and each of them represents each of the data points in the data. We use the function $dist()$ to calculate the distance matrix for each of the chosen time periods of the data. In the below example we find the distance matrix for the gastro symptoms 12 months after ended treatment:

```
Gastro_data_12M <- PivotData %>%
  filter(series == "12M") %>%
  select(starts_with("Gastro"))

#Transposing the data, so the row name are our symptoms
t_Gastro_data_12M<-t(Gastro_data_12M)

#Calculating the disimilarity matrix
dist_t_Gastro_data_12M<-dist(as.matrix(t_Gastro_data_12M), method = "euclidean")

#Printing the result
dist_t_Gastro_data_12M
```

```
##                       GastroDiarrhea GastroFlatulence GastroIncontinence
## GastroFlatulence           20.56040
## GastroIncontinence         19.06877         19.29306
## GastroProctitis            20.74903         21.30425           15.74163
## GastroBleedingRectum       21.06017         22.05199           15.94607
## GastroFistulaRectum        20.60859         21.34669           12.70613
##                       GastroProctitis GastroBleedingRectum
## GastroFlatulence
## GastroIncontinence
## GastroProctitis
## GastroBleedingRectum        14.23174
## GastroFistulaRectum         15.84557             15.95619
```

From the above dissimilarity matrix we can make the hierarchical clustering by connecting the symptoms which have the shortest distance to each other. We see from the above dissimilarity matrix that GastroIncontinence and GastroFistulaRectum has the shortest distance to each other, and will be connected first. After that GastroProctitis and GastroBleedingRectum has the shortest distance and will be connected. It will continue like that, until everything is connected together. From that we get the plot below.

```
plot(hclust(dist_t_Gastro_data_12M, method = "complete"),
     main = "Gastro 12 months after treatment",
     xlab="",
     sub="")
```

The plot shows that [GastroIncontinence, GastroFistulaRectum] cluster together with the [GastroProctitis, GastroBleedingRectum] cluster and the cluster [GastroDiarrhea, GastroFlatulence] is by itself.

PCA - Principal Component Analysis

We want to use PCA for a slightly different subject than EFA and HCA. Since we have to do with a lot of different questions, it would be convenient if it was possible to reduce this to a lower dimension in order to see if better clustering and visualization can be made. We use PCA for this since its a great algorithm to reduce dimensions. The package available for doing PCA is very good in Python, so this part is made in Python code[1].

PCA builds on the concept that we need to keep most of the variance in the data. In general the rule is at least above 90% but the closer to 100 % the better. Standardization of the data is not necessary since our data is on the same scale.

First we have to choose the k eigen vectors. This is the number of components we can reduce the data to. We use the formula:

$$\frac{\sum_{i=1}^{k} \sigma_i}{\sum_{i=1}^{n} \sigma_i} \geq 0.9$$

We use the python PCA package and get it to calculate the number of components to keep 90 % of the variance.
Python output for calculating the number of components:
»Explained Variance Ratio : 0.9073 - With 13 componets

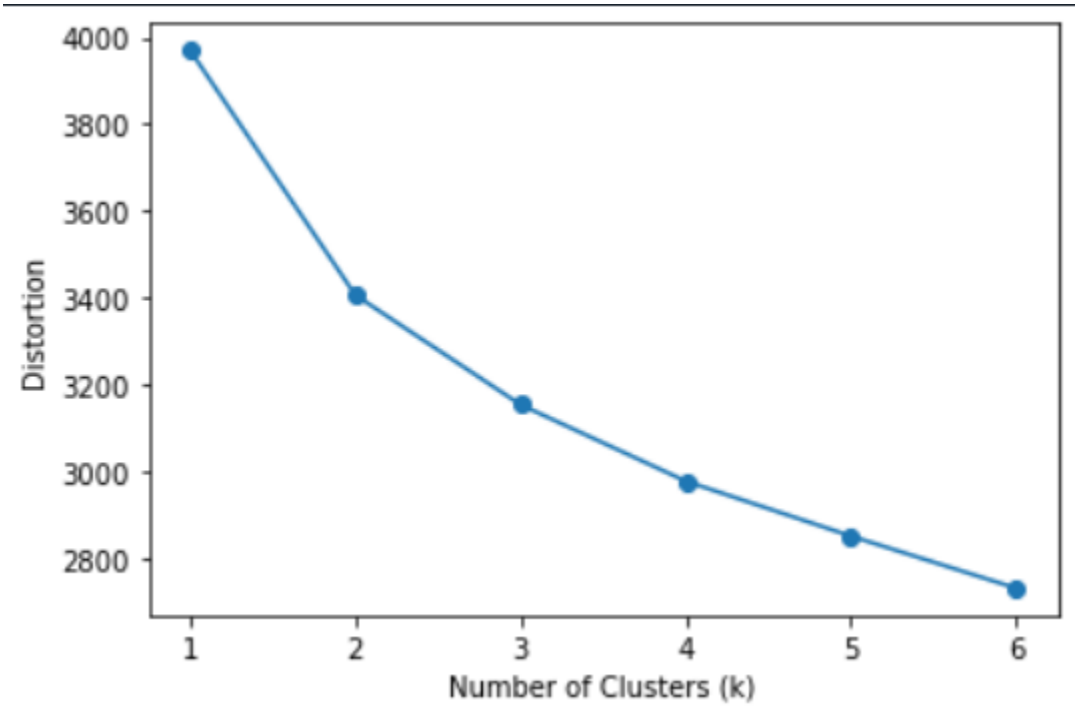
We can see that we need to keep 13 components in order to have above 90 % in variance.

[1]See attached Python file

We transform the data with 13 components using PCA.

We now want to do some clustering. Since we can not make sense of the symptoms due to the reduction of dimensions, we have to try and see if the reduced data is better at clustering the patients. This makes sense since in the future professionals want to understand the different groups of patients and relate them to the symptoms. We use K-means to cluster the patients. We run through 1 to 7 clusters to see what is an appropriate number of clusters. We use a scree plot as explained in the EFA section but on the Y-axis we have the distortion from the K-means.

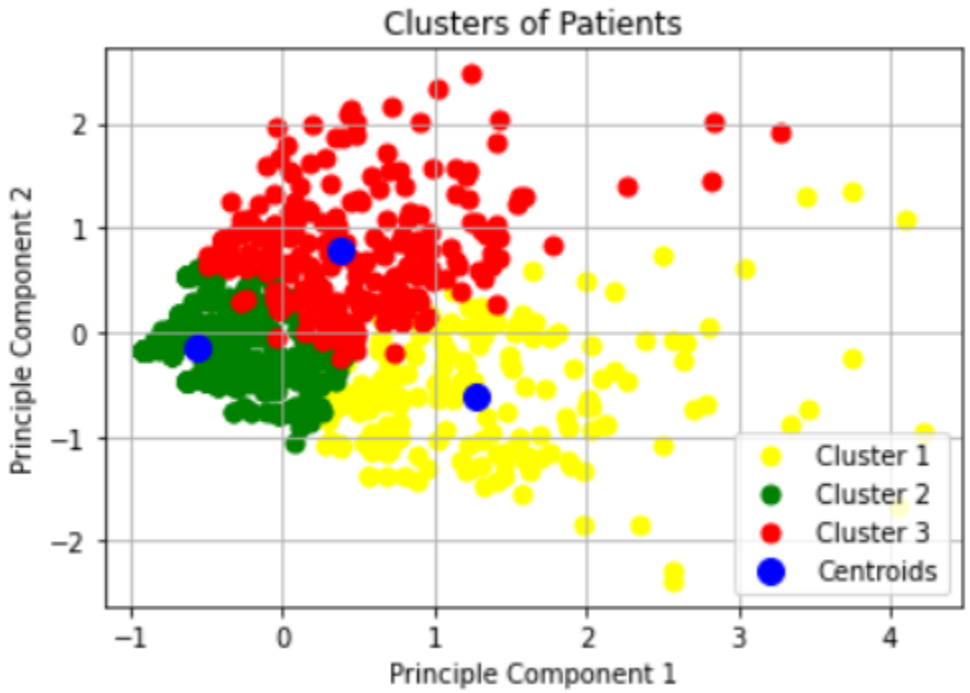
```
knitr::include_graphics("Elbow.png")
```



It looks to be around 2 or 3. We consider what we want to show, and choose 3 clusters. We want to divide the patients into three groups; good, normal and bad patient group.

We use K-means and show the plot of the clustering

```
knitr::include_graphics("Kmeans with pca.png")
```



It is clear from the plot that the patients is closely together. This make sense, when we take a look at how many patients and the average of their symptoms are in each group.
Python output:
»K Means Result :
Counter({0: 190, 1: 613, 2: 272})
Mean cluster 1 : 0.00983981693363844, STD :0.04596831187686466
Mean cluster 2 : 0.9547485637279242, STD :0.18927255232903784
Mean cluster 3 : 1.9291879795396416, STD :0.346494888884988

We see the averege for each cluster is close to each other.

We now do the same clustering without PCA applied to compare if the dimension reduction gives better clustering.

We see the number and average for each group again.
Python output:
»Final K Means Result (no PCA) :
Counter({0: 185, 1: 615, 2: 275})
Mean cluster 1 : 0.0066403162055335965, STD :0.040656230189262024
Mean cluster 2 : 0.9377203290246767, STD :0.22033464459189359
Mean cluster 3 : 1.9034994697773069, STD :0.40163764969541954

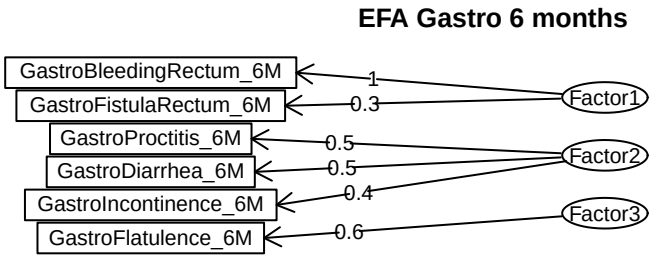
If we look at the counter for each cluster in both PCA and without we see very similar results. Also the means of each cluster is very similar.

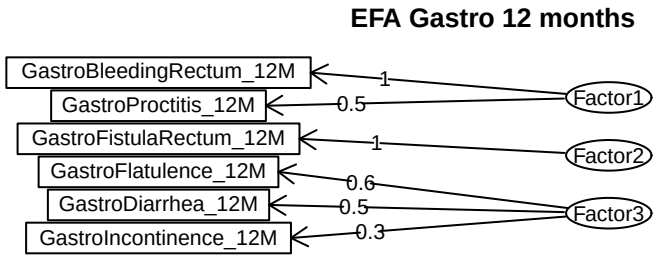
The goal was to see how or if the clustering of patients would change with PCA and whether it brought insights to how it could be related to the clusters of symptoms. The conclusion has to be that PCA does not bring anything great to the table when it comes to better understanding of the data and the relation between symptoms and patients overall well being.

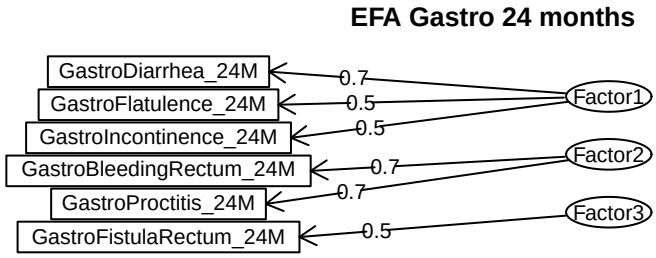
Comparision of EFA AND HCA

Gastro plot

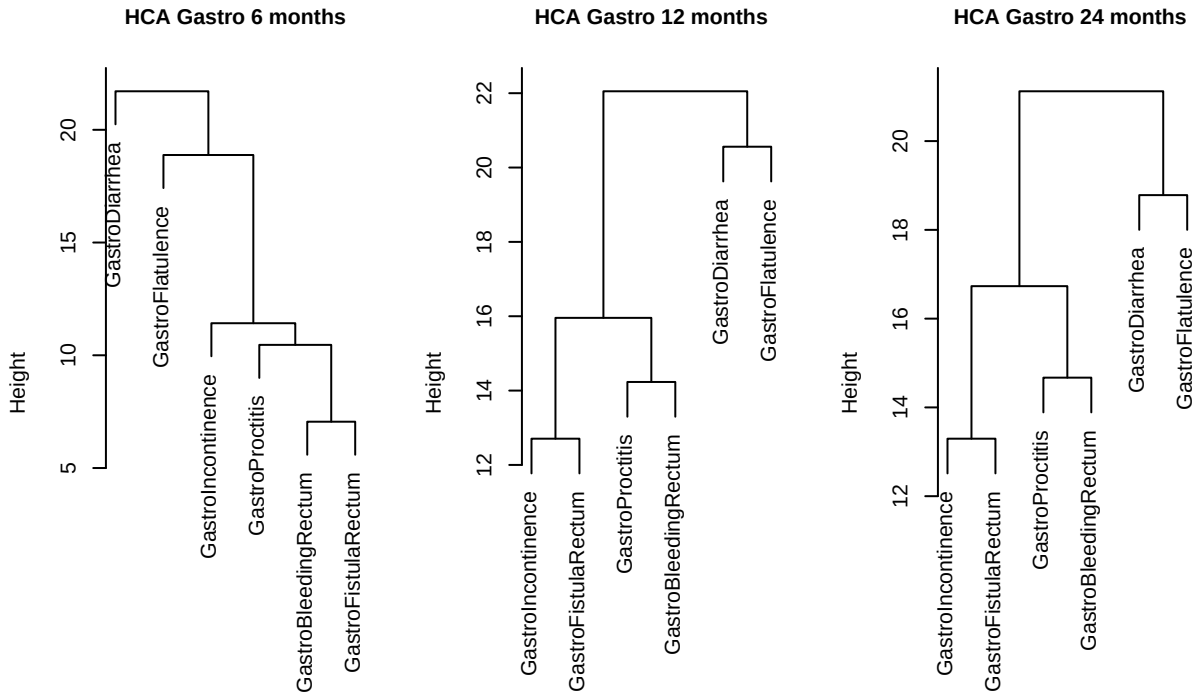
```
#EFA for gastro organ month 6
EFA_model_gastro6m <- factanal(na.omit(Physician[,c(3,19,35,51,67,83)]), factors=3)

#EFA for gastro organ month 12
EFA_model_gastro3m <- factanal(na.omit(Physician[,c(5,21,37,53,69, 85)]), factors=3)

#EFA for gastro organ month 24
EFA_model_gastro9m <- factanal(na.omit(Physician[,c(7,23,39,55,71,87)]), factors=3)


# HCA gastro month 12
#Preparing data for with Gastro symptoms for 6 month
Gastro_data_6M <- PivotData %>%
  filter(series == "6M") %>%
  select(starts_with("Gastro"))
#Tranposing the data
t_Gastro_data_6M<-t(Gastro_data_6M)
#Making the dissimilatiry matrix
dist_t_Gastro_data_6M<-dist(as.matrix(t_Gastro_data_6M), method = "euclidean")


#HCA gastro 12 month
Gastro_data_12M <- PivotData %>%
  filter(series == "12M") %>%
  select(starts_with("Gastro"))
t_Gastro_data_12M<-t(Gastro_data_12M)
dist_t_Gastro_data_12M<-dist(as.matrix(t_Gastro_data_12M), method = "euclidean")


#HCA gastro 24 month
Gastro_data_24M <- PivotData %>%
  filter(series == "24M") %>%
  select(starts_with("Gastro"))
t_Gastro_data_24M<-t(Gastro_data_24M)
dist_t_Gastro_data_24M<-dist(as.matrix(t_Gastro_data_24M), method = "euclidean")


# Plot EFA
par(mfrow=c(3,1))
fa.diagram(EFA_model_gastro6m$loadings, main = "EFA Gastro 6 months")
fa.diagram(EFA_model_gastro3m$loadings, main = "EFA Gastro 12 months")
fa.diagram(EFA_model_gastro9m$loadings, main = "EFA Gastro 24 months")
```








```
#Plot HCA
par(mfrow=c(1,3))
plot(hclust(dist_t_Gastro_data_6M, method = "complete"),
     main = "HCA Gastro 6 months",
     cex.main = 1,
     xlab="",
     sub="")

plot(hclust(dist_t_Gastro_data_12M, method = "complete"),
     main = "HCA Gastro 12 months",
     cex.main = 1,
     xlab="",
     sub="")

plot(hclust(dist_t_Gastro_data_24M, method = "complete"),
     main = " HCA Gastro 24 months",
     cex.main = 1,
     xlab="",
     sub="")
```

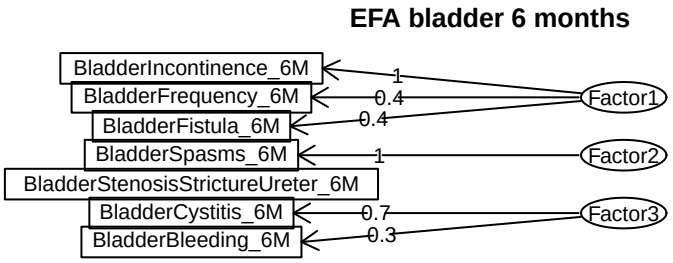


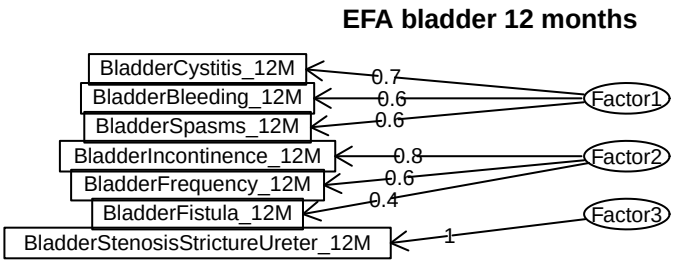
We start by looking at each of the analysis seperate. If we look at the EFA analysis and how it split the symptoms, we see a change in which symptoms that occur together. We see that GastroFistulaRectum, GastroProctitis and GastroFlatulance changes factor from month 6 to month 12. However the symptoms that occur together in month 12 and month 24 are correlated with the same factor, but the degree of correlation changes.

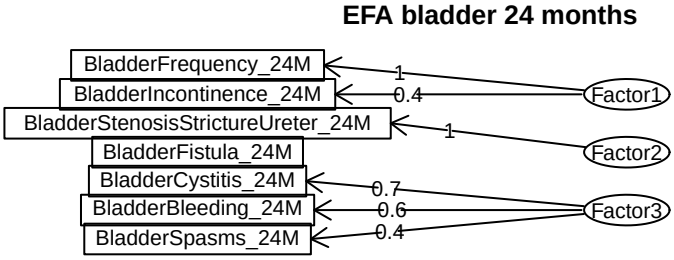
In the HCA we also see a change in which symptoms that occur together from month 6 to month 12. We see that the only pair of symptoms in month 6 is [GastroBleedingRectum,GastroFistulaRectum] where the other symptoms are connected to that pair. In month 12 and 24 they are connected in pairs and after that connected to each other.

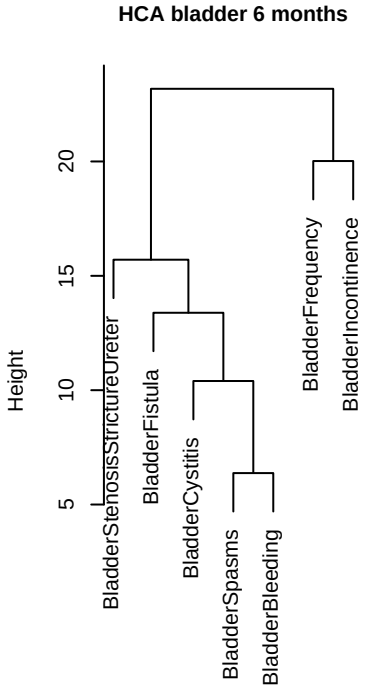
Both in EFA and HCA [GastroProctitis, GastroBleddingRectum] exist together. Furthermore we have [GastroDiarrhea, GastroFlatuence] together in both analysis but in EFA GastroIncontinence also exist together with this group. Medically it makes more sense that [GastroDiarrhea, GastroFlatuence, GastroIncontinence] exist together because GastroFlatuence gives bowel problems which leads to gas in the belly, which can cause Diarrhea and Incontinence. For this organ the EFA will give the most expected result.

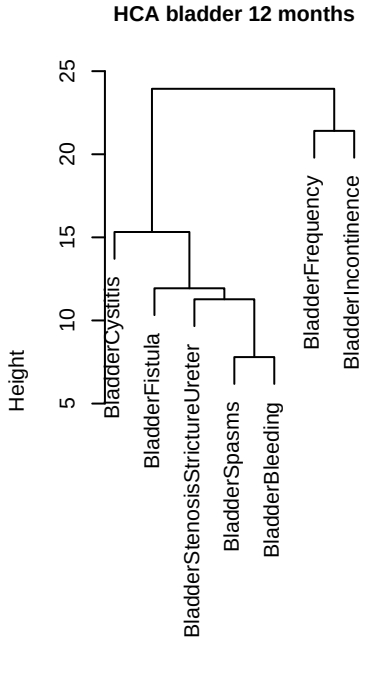
Bladder plots

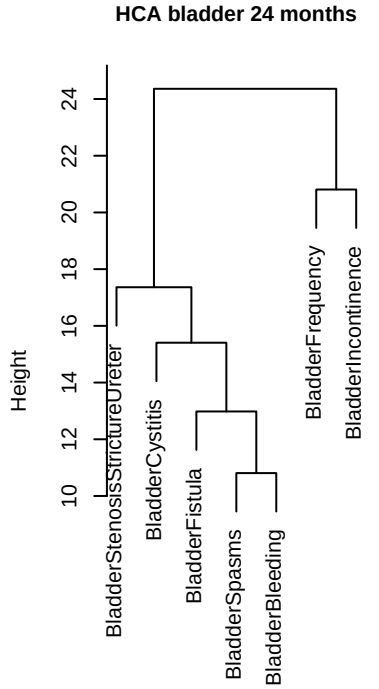










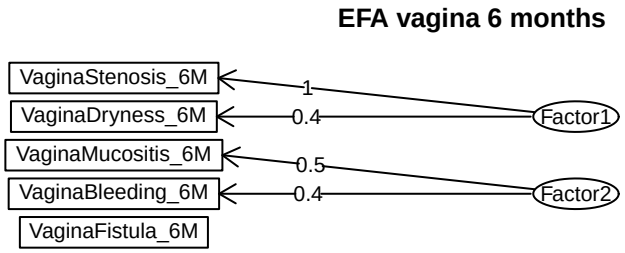


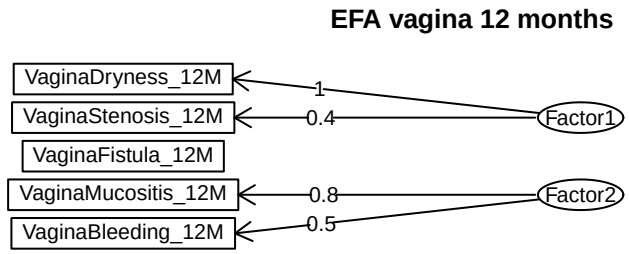
In the EFA for the bladder organ [BladderIncontinence, BladderFrequency] is correlated with the same factor in all tree months. The same applies for the [BladderCystitis, BladderBleeding] and from month 12 the symptom BladderSpasms correlates with the same factor. The last symptoms is BladderStenosisStricturUreter which does not connect with any of the other symptoms.

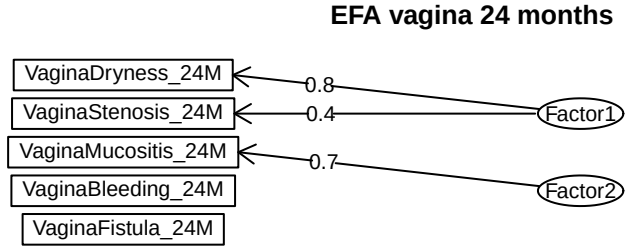
In the HCA [BladderIncontinence, BladderFrequency] also occur together in all three months. The same is applicable to [BladderSpasms, BladderBleeding]. The other three symptoms connect to the [BladderSpasms, BladderBleeding] cluster in different order ind the different months.

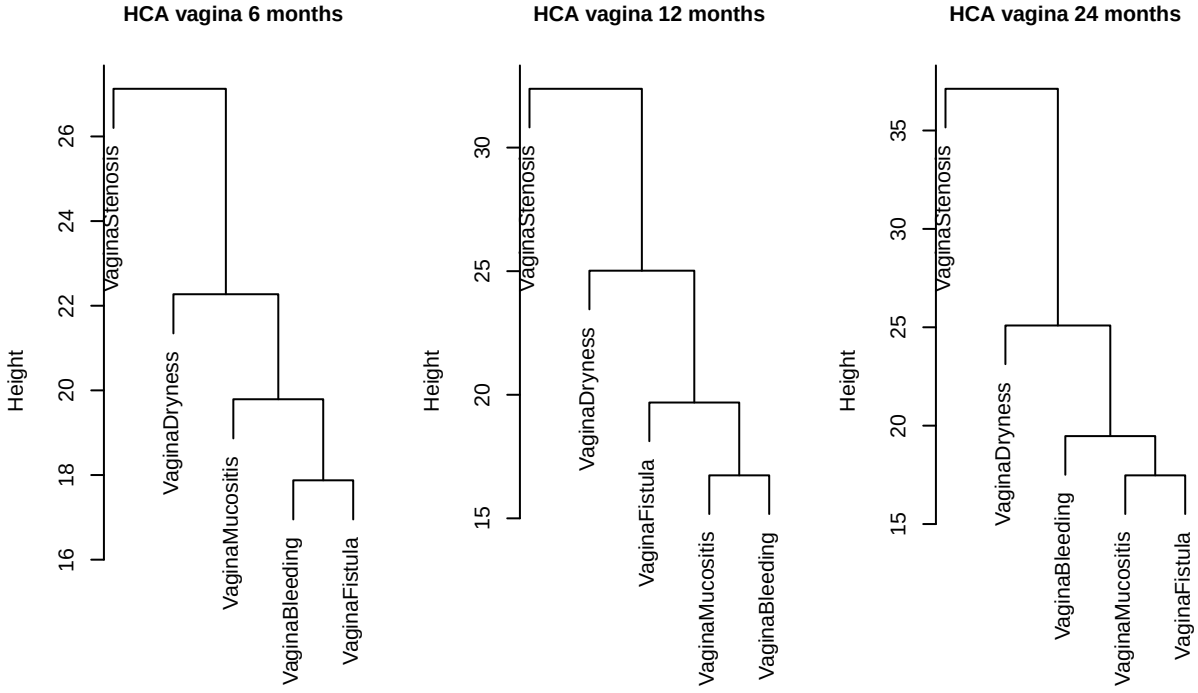
Both analysis agree on that [BladderIncontinence, BladderFrequency] cluster together in all time periods, which make sense medically, because BladderIncontinence leads to more often toilet visits. In the HCA the other symptoms connect differently for every time period. This might be because there is few observations of BladderFistula, BladderBleeding and BladderSpasms, which make it difficult to compare the symptoms. In the EFA we see that the symptom clusters are more consistent during time, except for BladderFistula. An explanation to, that the BladderFistula does not connect to BladderIncontinence and BladderFrequency in month 24 is because usually Fistula requires a surgery intervention, so in the next physician follow ups, it will not be reported anymore. The HCA will probably give better results if we had more cases of BladderFistula, BladderBleeding and BladderSpasms. Regardless of the number of observations, the EFA will perform more consistent.

Vagina plots







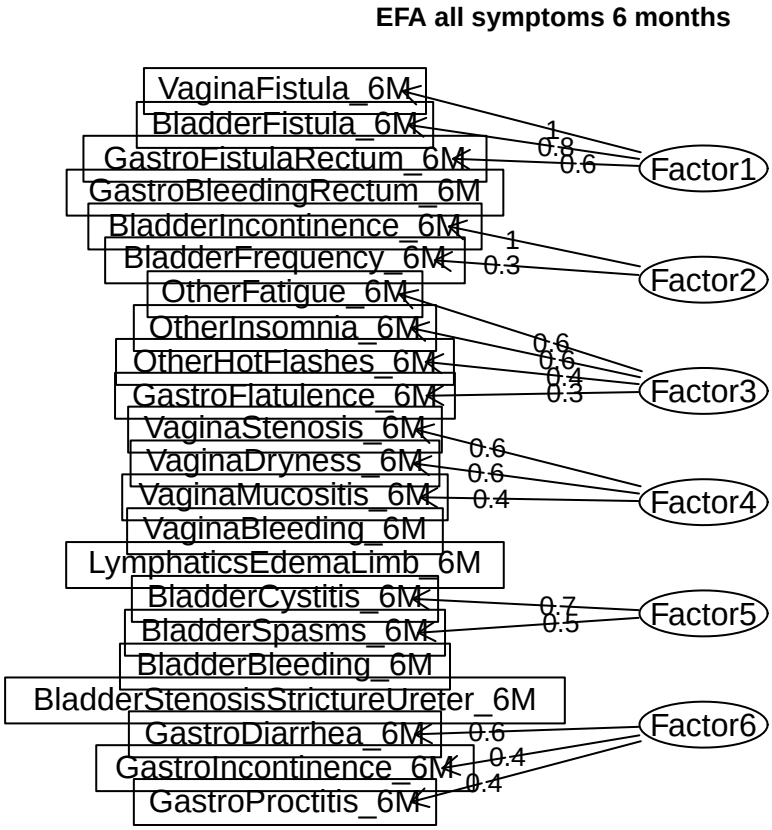


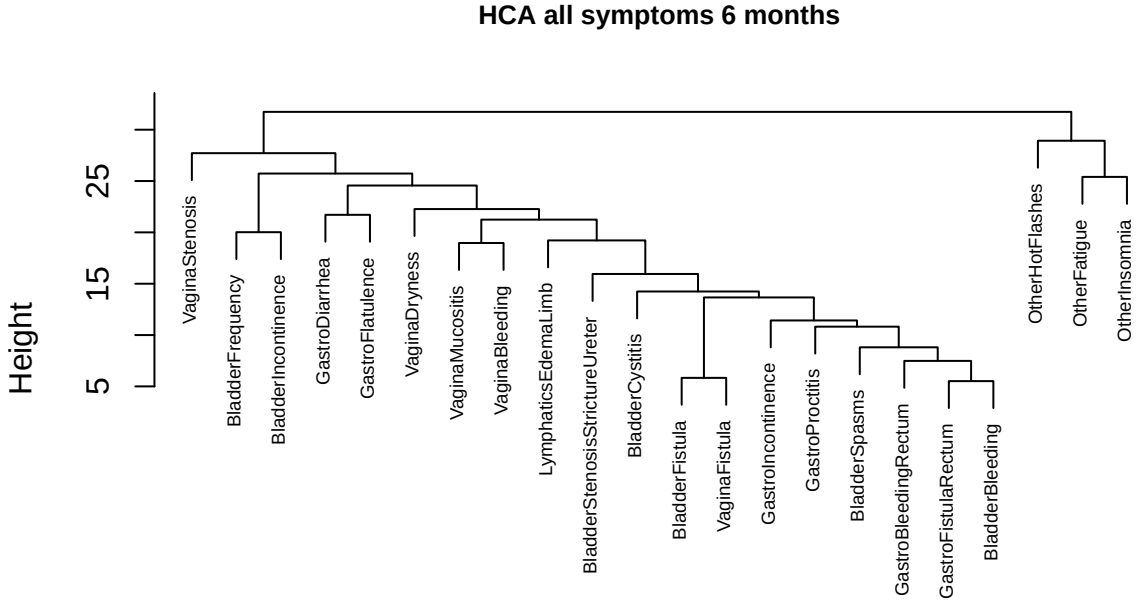
In the EFA [VaginaStenois, VaginaDryness] correlates with the same factor in all three analyzed months. The symptoms [VaginaMucositis, VaginaBleeding] is group with the same factor in month 6 and 12 but not in month 24. VaginaFistula does not correlated with any factor in any of the time periods.

In the HCA we see that the three symptoms [VaginaBleeding, VaginaMucositis, VaginaFistula] all is connected to each other in all three time periods but it changes a bit which of them that are correlated the most.

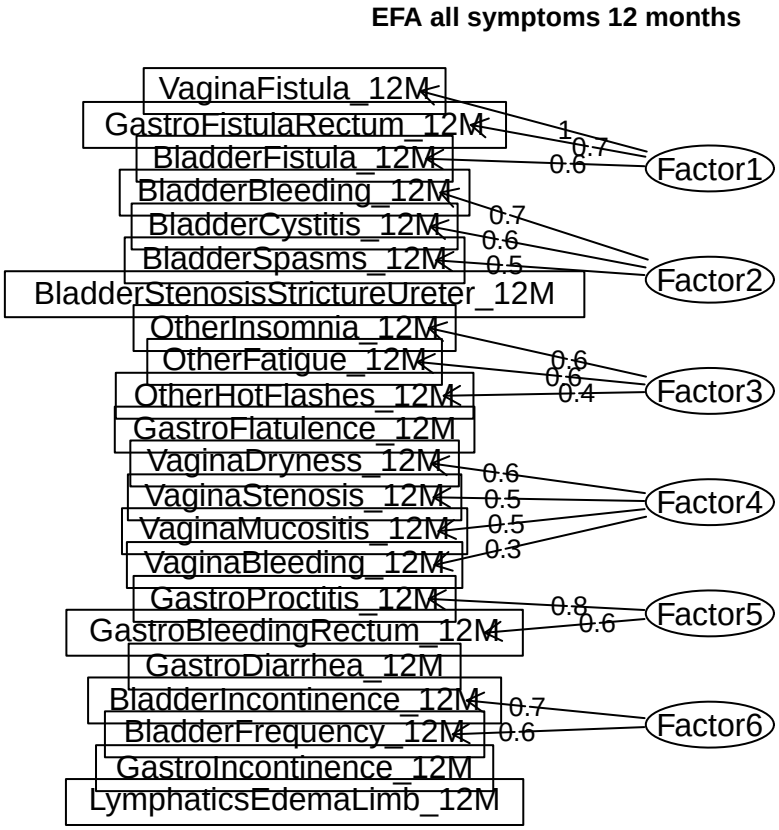
The only symptoms that both analysis group together is [VaginaBleeding, VaginaMucositis], which makes sense since VaginaMucositis is causing inflammation, which leads to bleeding. The reason why VaginaBleeding does not cluster with VaginaMucositis in EFA in month 24 is because VaginaMucositis usually sort itself out with time. Generally, we get the same output for both HCA and EFA. However the HCA forces all symptoms to be connected in some way, where the EFA loose some information about the connections for the symptoms with low correlation

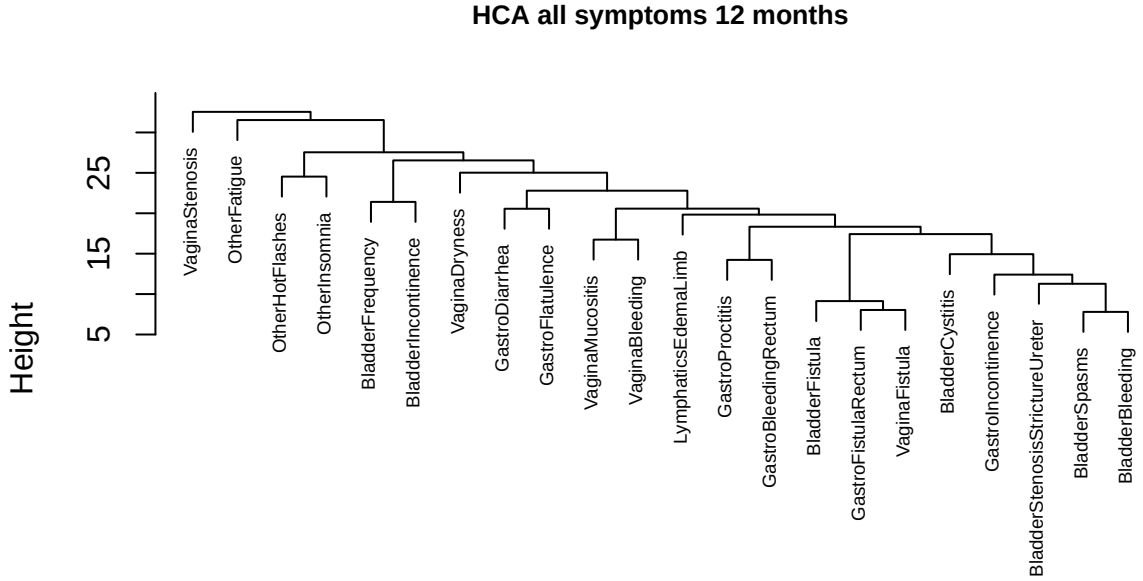
Plots all symptoms 6 months





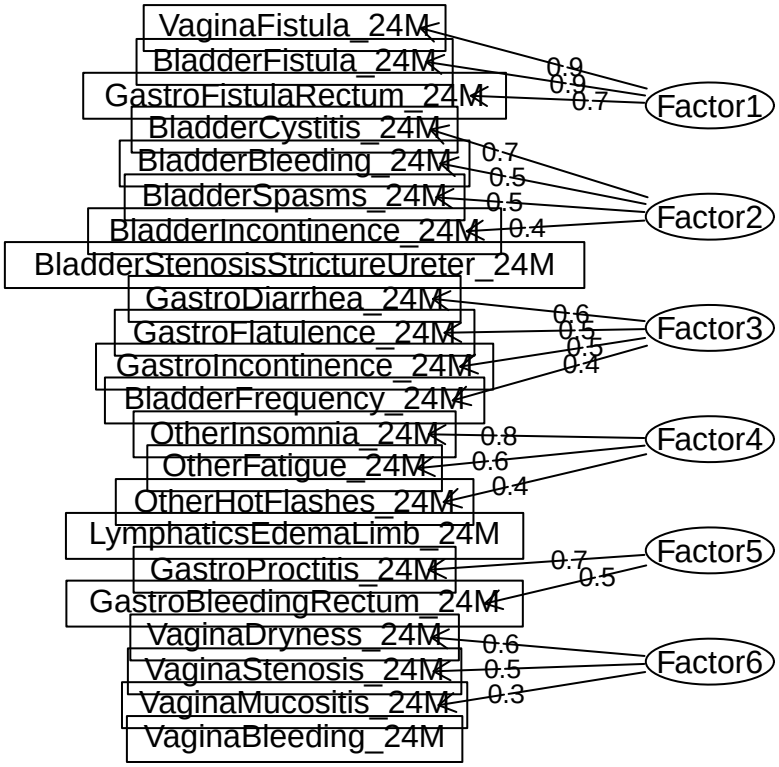
Plots all symptoms 12 months



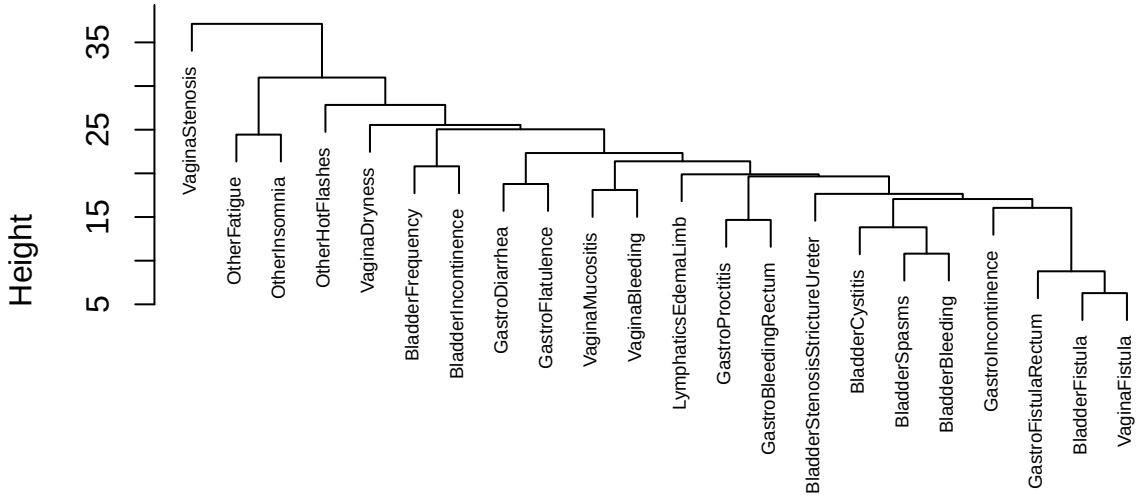


Plots all symptoms 24 months

EFA all symptoms 24 months



HCA all symptoms 24 months



In the EFA all the three fistulas group together in one factor. Furthermore, each organs symptoms more or less cluster together, which makes sense, since they often are correlated.

In the HCA we see that all the Fistula is close to each other in all three time periods. The same is applicable for the "Other" symptoms which also are placed close to each other.

Generally, the plot for month 12 in both analysis differs from month 6 and month 24. It can be caused by fewer observations for some of the symptoms.

In both analysis we find a cluster with the symptoms [VaginaFistula, BladderFistula, GastroFistulaRectum] and another cluster with the symptoms [OtherInsomnia, OtherFatigue, OtherHotFlashes]. Overall the EFA give a better output and it makes sense how the symptoms are distributed in the clusters.