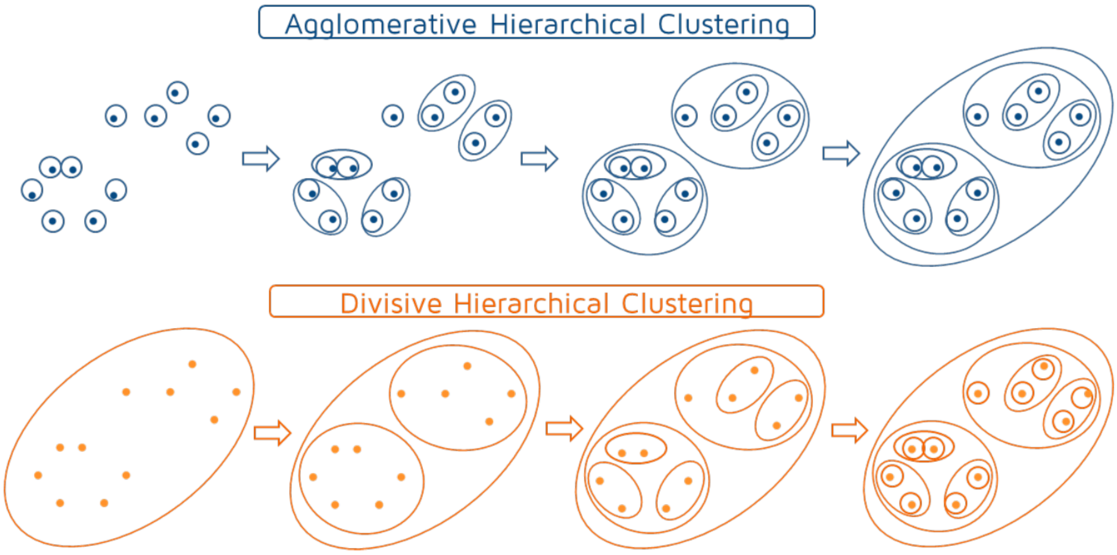
EHR Analysis – EMRBots

1 Exercise – What do we want to achieve?

Choose your data from the 100-Patients dataset and start to investigate. Ask yourself questions: “What would you like to do?”, “What do you think makes sense?”, “How could you make usage from the data without spending too much time on it?”

1.1 Cluster the patients

Let us take the patient data and try to cluster the patient. Compare the results from the two options of hierarchical clustering → Agglomerative Hierarchical Clustering & Divisive Hierarchical Clustering


Agglomerative Hierarchical Clustering
Divisive Hierarchical Clustering

1.2 Investigate possible comorbidity

Comorbidity is the co-existence of two or more diseases in the same patient. EHR for data mining offers the opportunity to discover disease associations and comorbidity patterns from the clinical history of patients gathered during routine medical care. With the help of the Text-Mining-Library in R, an approach should be prepared to set the base for further investigation on possible patterns.

2 Cluster the patients

2.1 Load & transform the data

- Load the data with the code from script-part {r read}
- We would like to transform the data, so we will be able to visualize a hierarchical clustering of the patients. The first thing coming to mind after recognizing there is a column “*PatientDateOfBirth*” is, that we would like to add a new column with the computed current age of each patient as a numerical value. Maybe the age of the patients will be needed later.
- Within the output above, you may recognize, that “*PatientID*” is a very large string, which will make it difficult for us to visualize each patient within a dendrogram. Therefore, let us create another column “*PatientNumber*” which contains consecutive numbers being assigned to the patients.
- This means we can now visualize the patient with the *ID == FB2ABB23-C9D0-4D…* as patient with the number “1” within the upcoming dendrogram. The statements within the part {r reorder} moves the columns into the wished order, as we’d like to have now “*PatientNumber*” on the first position.

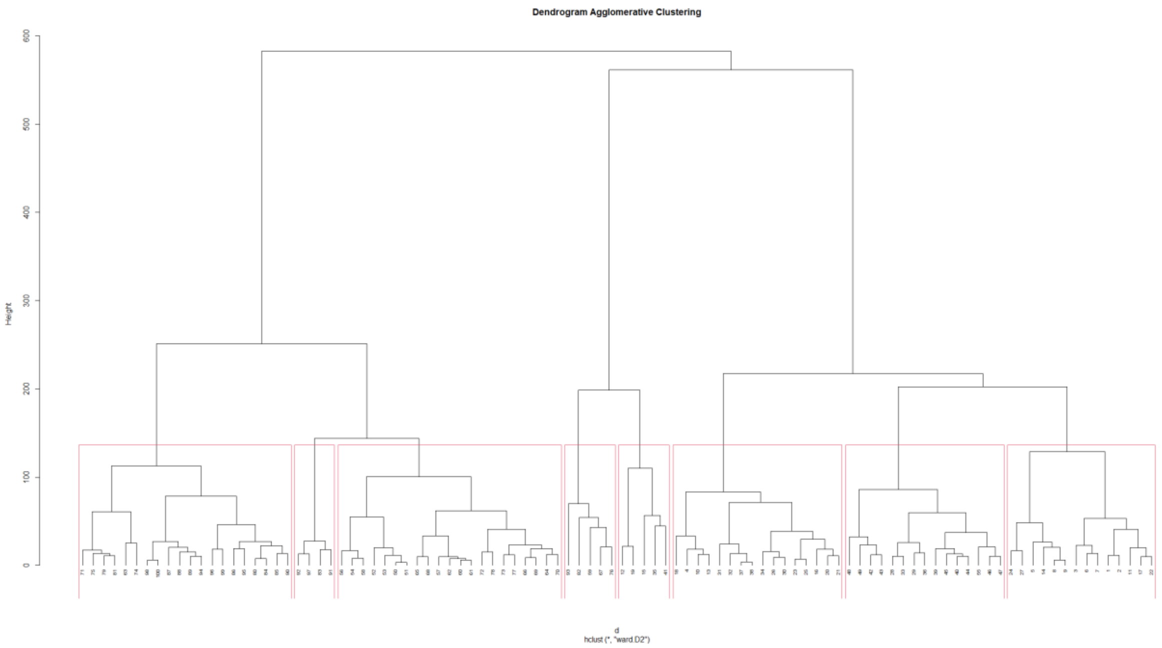
2.2 Apply Hierarchical Clustering

Hierarchical Clustering is a technique to club similar data points into one group and separates dissimilar observations into different groups or clusters. In Hierarchical Clustering, this hierarchy of clusters can either be created from top to bottom or vice-versa. Hence, it is two types namely – Divisive and Agglomerative

2.2.1 Agglomerative Hierarchical Clustering

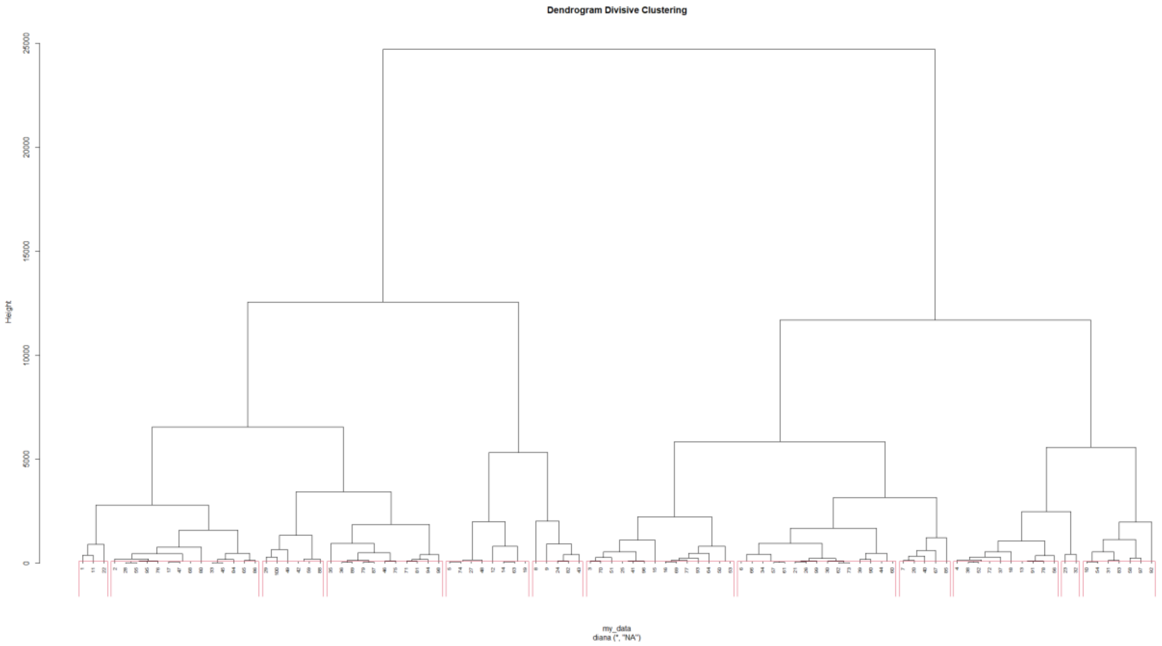
It’s also known as Hierarchical Agglomerative Clustering (HAC) {r agglomerative} or AGNES (acronym for Agglomerative Nesting → {r agnes}). In this method, each observation is assigned to its own cluster. Then, the similarity (or distance) between each of the clusters is computed and the two most similar clusters are merged into one. Finally, steps 2 and 3 are repeated until there is only one cluster left.

We’re applying this approach with the above statements and receive the following plots of dendrograms where the amount of generated clusters for our patients might be recognized.



2.2.2 Divisive Hierarchical Clustering

In the Divisive method, we assume that the observations belong to a single cluster and then divide the cluster into two least similar clusters. This is repeated recursively on each cluster until there is one cluster for each observation. This technique is also called “DIANA”, which is an acronym for “Divisive Analysis”.



3 Text Processing Approach

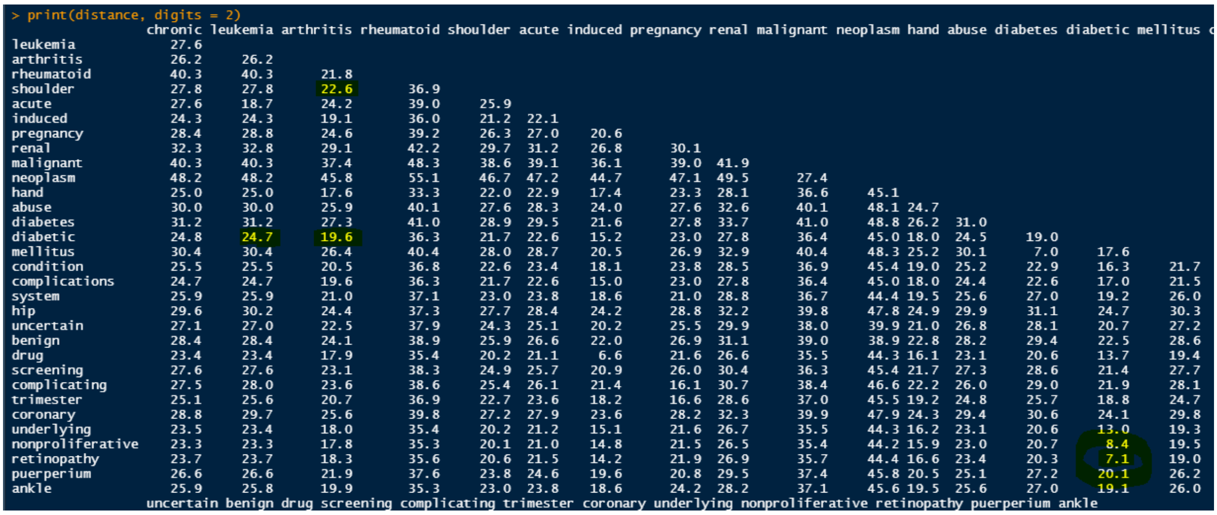
After trying to cluster the patients we would like to take action on our second intentional task, trying to prepare an approach for potential comorbidity detection. Therefore, we will use a standard Text-Mining strategy.

- Load the data
  - For this task, our current dataframe isn't much of a help as we'd like to have descriptions of the patients' diagnoses, so we can compute the similarity of the terms. Let us load the desired data set into a new dataframe called "*newdata*".

- Process the data by building corpus
  - Create a corpus for the new dataframe so we can load it into a term matrix.
  - Also make sure to try to clean the data from potential items like *numbers*, *punctuations*, *stopwords*, or *predefined misleading terms*.

- Calculate the distance
  - Now it helps us to calculate the distance of the terms within our matrix. It helps us to recognize which terms might show up in the same cluster.

  - See for that the yellow marked values as examples:
    - The lower the value between two terms is → the higher the possibility these two terms might show up in the same cluster. There might be some relationship between suffering on "*retinopathy*" and being "*diabetic*"
    - The higher the value → the lower the possibility that these terms are assigned in the same cluster.



```
> print(distance, digits = 2)
                 chronic leukemia arthritis rheumatoid shoulder acute induced pregnancy renal malignant neoplasm hand abuse diabetes diabetic mellitus c
leukemia            27.6
arthritis           26.2     26.2
rheumatoid          40.3     40.3      21.8
shoulder            27.8     27.8      22.6       36.9
acute               27.6     18.7      24.2       39.0     25.9
induced             24.3     24.3      19.1       36.0     21.2  22.1
pregnancy           28.4     28.8      24.6       39.2     26.3  27.0    20.6
renal               32.3     32.8      29.1       42.2     29.7  31.2    26.8      30.1
malignant           40.3     40.3      37.4       48.3     38.6  39.1    36.1      39.0  41.9
neoplasm            48.2     48.2      45.8       55.1     46.7  47.2    44.7      47.1  49.5      27.4
hand                25.0     25.0      17.6       33.3     22.0  22.9    17.4      23.3  28.1      36.6     45.1
abuse               30.0     30.0      25.9       40.1     27.6  28.3    24.0      27.6  32.6      40.1     48.1 24.7
diabetes            31.2     31.2      27.3       41.0     28.9  29.5    21.6      27.8  33.7      41.0     48.8 26.2  31.0
diabetic            24.8     24.7      19.6       36.3     21.7  22.6    15.2      23.0  27.8      36.4     45.0 18.0  24.5     19.0
mellitus            30.4     30.4      26.4       40.4     28.0  28.7    20.5      26.9  32.9      40.4     48.3 25.2  30.1      7.0     17.6
condition           25.5     25.5      20.5       36.8     22.6  23.4    18.1      23.8  28.5      36.9     45.4 19.0  25.2     22.9     16.3     21.7
complications       24.7     24.7      19.6       36.3     21.7  22.6    15.0      23.0  27.8      36.4     45.0 18.0  24.4     22.6     17.0     21.5
system              25.9     25.9      21.0       37.1     23.0  23.8    18.6      21.0  28.8      36.7     44.4 19.5  25.6     27.0     19.2     26.0
hip                 29.6     30.2      24.4       37.3     27.7  28.4    24.2      28.8  32.2      39.8     47.8 24.9  29.9     31.1     24.7     30.3
uncertain           27.1     27.0      22.5       37.9     24.3  25.1    20.2      25.5  29.9      38.0     39.9 21.0  26.8     28.1     20.7     27.2
benign              28.4     28.4      24.1       38.9     25.9  26.6    22.0      26.9  31.1      39.0     38.9 22.8  28.2     29.4     22.5     28.6
drug                23.4     23.4      17.9       35.4     20.2  21.1     6.6      21.6  26.6      35.5     44.3 16.1  23.1     20.6     13.7     19.4
screening           27.6     27.6      23.1       38.3     24.9  25.7    20.9      26.0  30.4      36.3     45.4 21.7  27.3     28.6     21.4     27.7
complicating        27.5     28.0      23.6       38.6     25.4  26.1    21.4      16.1  30.7      38.4     46.6 22.2  26.0     29.0     21.9     28.1
trimester           25.1     25.6      20.7       36.9     22.7  23.6    18.2      16.6  28.6      37.0     45.5 19.2  24.8     25.7     18.8     24.7
coronary            28.8     29.7      25.6       39.8     27.2  27.9    23.6      28.2  32.3      39.9     47.9 24.3  29.4     30.6     24.1     29.8
underlying          23.5     23.4      18.0       35.4     20.2  21.2    15.1      21.6  26.7      35.5     44.3 16.2  23.1     20.6     13.0     19.3
nonproliferative    23.3     23.3      17.8       35.3     20.1  21.0    14.8      21.5  26.5      35.4     44.2 15.9  23.0     20.7      8.4     19.5
retinopathy         23.7     23.7      18.3       35.6     20.6  21.5    14.2      21.9  26.9      35.7     44.4 16.6  23.4     20.3      7.1     19.0
puerperium          26.6     26.6      21.9       37.6     23.8  24.6    19.6      20.8  29.5      37.4     45.8 20.5  25.1     27.2     20.1     26.2
ankle               25.9     25.8      19.9       35.3     23.0  23.8    18.6      24.2  28.2      37.1     45.6 19.5  25.6     27.0     19.1     26.0
                 uncertain benign drug screening complicating trimester coronary underlying nonproliferative retinopathy puerperium ankle
```