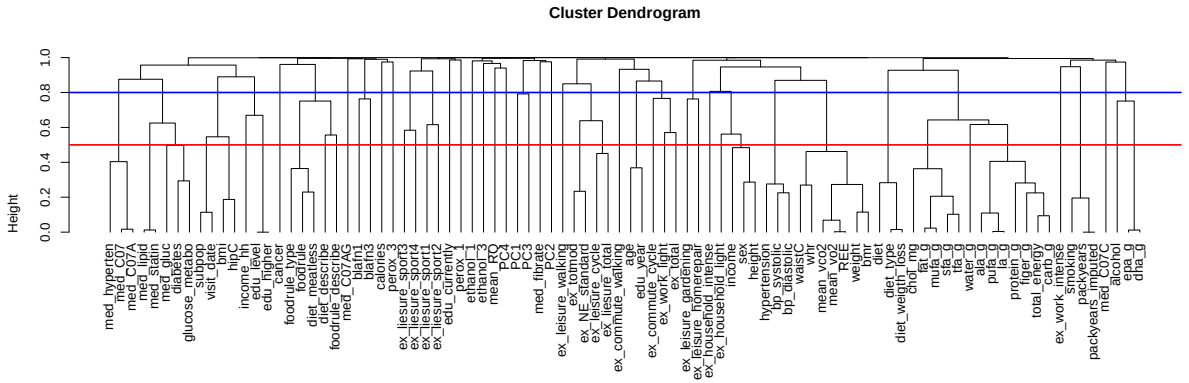
Leiden University
Medical Center

Hierarchical Clustering in Clinical Research

—— Using Venous Thrombosis as an example



EPI-OMICs-Meeting

Keyong Deng

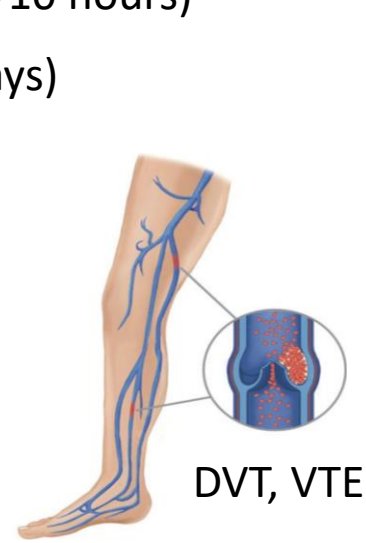
Department of Clinical Epidemiology

Background

- Most of the diseases are multifactorial, caused by a combination of various risk factors

Transient Provoking factors

- surgery
- long-distance travel (>10 hours)
- immobilization (> 3 days) within last 2 months

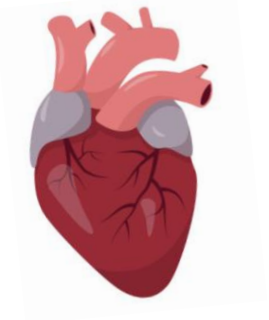


DVT, VTE

Unprovoking factors

Persistent Provoking factors

Active cancer / severe inflammatory diseases / …



<u>VTE Patients, but without provoking factors</u>

The issues of current risk stratification

- Not all patients benefit from the same treatment
  - DOACs / VKs

- Not all patients have the same clinical outcome
  - recurrent VTE
  - bleeding
  - prognosis
  - death
  - …

Is there any new phenotype indicating different risks of clinical complications/outcomes for the VTE patients?

Different Scenarios

Healthy Controls

Pre-disease → Follow-up → High risk group / incidence disease

Cases

Disease Onset → Follow-up → Recurrent events / Mortality / Prognosis

https://doi.org/10.1016/j.jtha.2023.01.025

jth

ORIGINAL ARTICLE

Exploring phenotypes of deep vein thrombosis in relation to clinical outcomes beyond recurrence

ELSEVIER

Contents lists available at ScienceDirect

Thrombosis Research

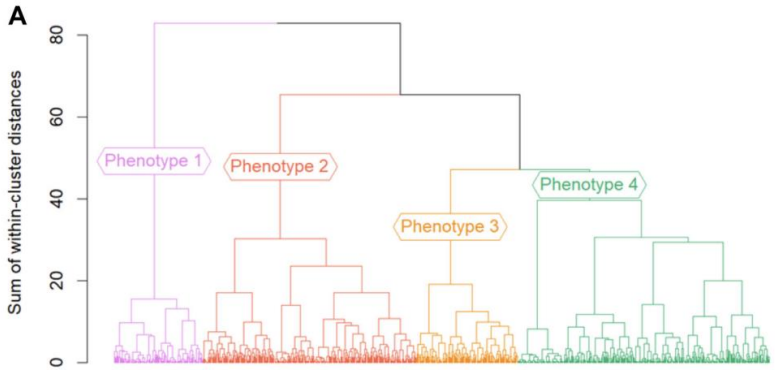
journal homepage: www.elsevier.com/locate/thromres

Full Length Article

Unsupervised clustering of venous thromboembolism patients by clinical features at presentation identifies novel endotypes that improve prognostic stratification

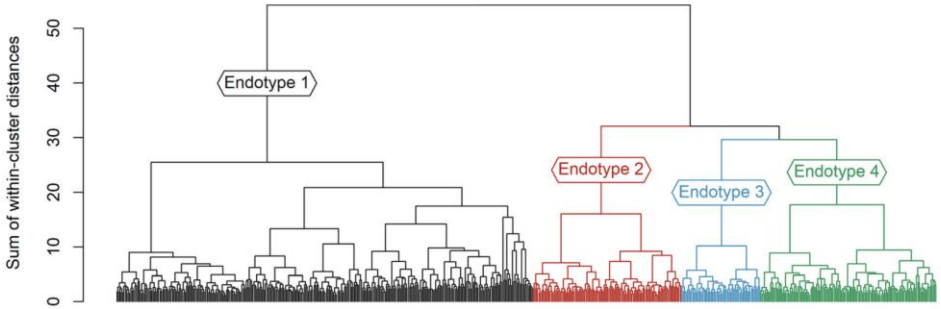
The main idea

825 Deep vein thrombosis patients



Hierarchical clustering applied to 23 variables (variables are mainly about risk factors)

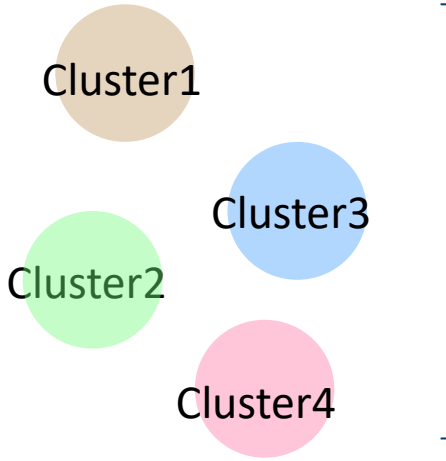
693 Acute venous thrombosis patients



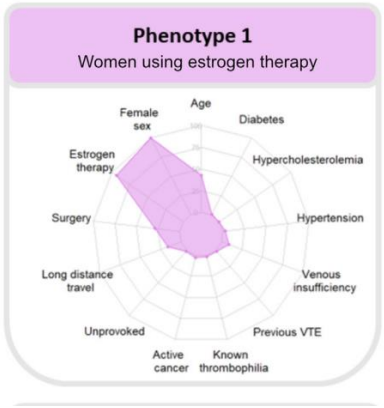
Hierarchical clustering applied to 58 variables (clinical and laboratory variables : continuous and dichotomous)

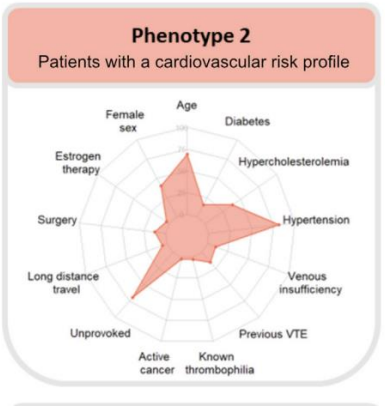
The main idea

After identifying the different clusters for individuals

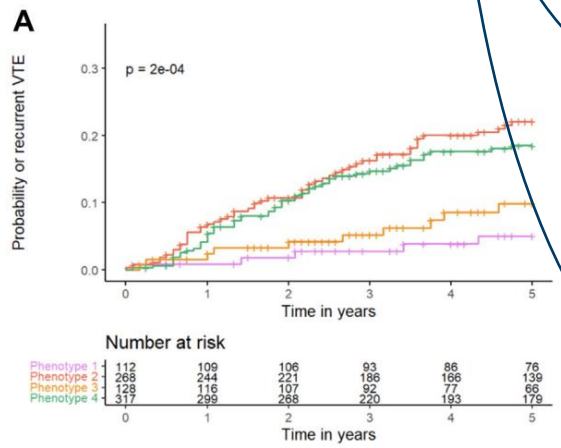


- Describe characteristic of different clusters





- Clinical Outcomes



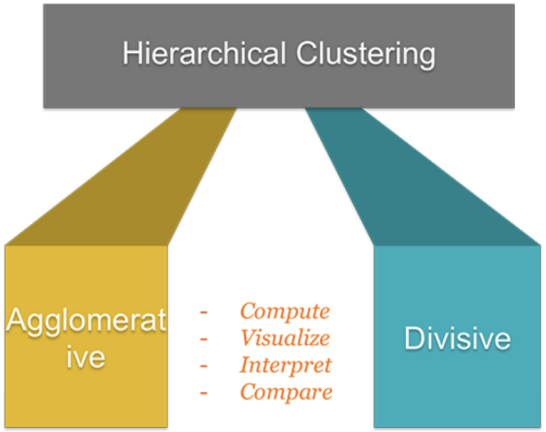
Hazard ratio for recurrent VTE (compared within different clusters)

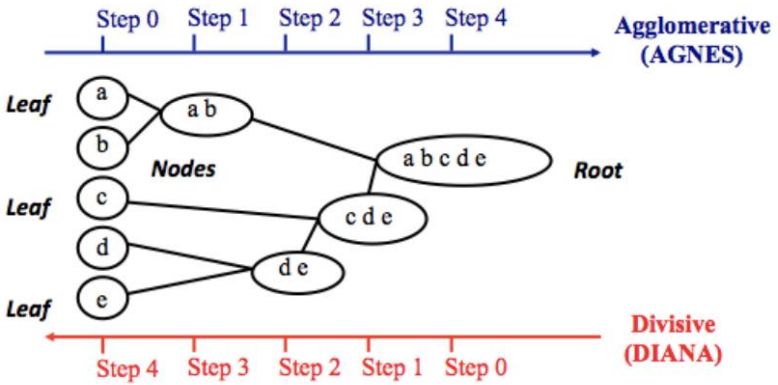
- Omics profiles

Differences in molecular traits among clusters

What's hierarchical clustering?

It is a method for grouping objects based on their similarity, in contrast to Kmeans clustering, hierarchical clustering doesn't require to pre-specify the number of clusters.





Two types:

- Agglomerative clustering

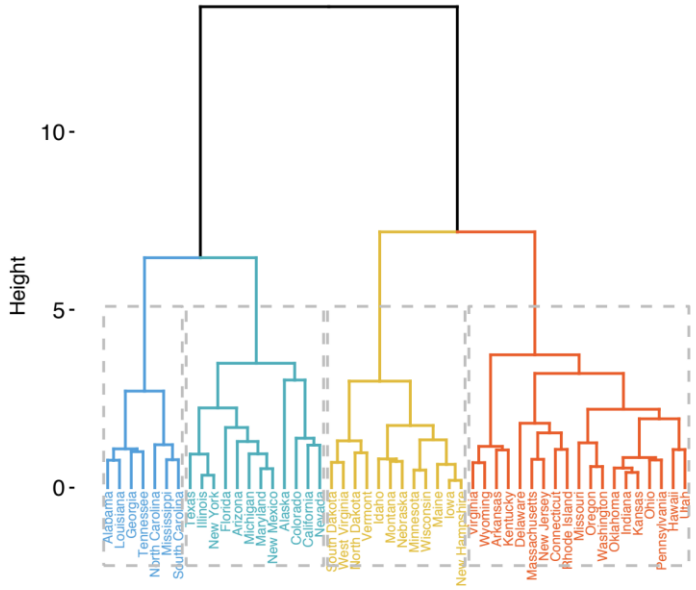
each observation is considered as a cluster, then the most similar clusters are merged

- Divisive clustering

all objects are included in one cluster, most different cluster are successively divided

Key Terms

1. Dendrogram: A representation of the records and the hierarchy of clusters



2. Distance: A measure of how close one record is to another

3. Dissimilarity: A measure of how close one cluster is to another

The leaves of the tree correspond to the records. The length of the branch in the tree indicates the degree of dissimilarity between corresponding clusters.

Steps to do hierarchical clustering

1. Preparing the data (computing distance between objects)
   - Euclidean distance

$$d(x,y) = \left(\sum_{j=1}^{d} (x_j - y_j)^2\right)^{\frac{1}{2}}$$

   - Manhattan distance

$$d(x,y) = \sum_{j=1}^{d} |x_j - y_j|$$

2. Computing dissimilarity matrix between every pair of objects

3. Using Linkage function to group objects into cluster tree, based on distance information
   Clusters that are in close proximity are linked together

4. Determining where to cut the tree in dendrogram into clusters

methods to compute distance between clusters
"complete", "single", "average" , "Ward"

Different methods to measure dissimilarity

1. *complete-linkage* method: It tends to produce more compact cluster

$$D(A, B) = \max d\left(a_i, b_j\right) \text{ for all pairs } i, j$$

2. *average-linkage* method:

The distance between two clusters is defined as the average distance between the elements in cluster 1 and the elements in cluster 2.

3. *ward-linkage* method:

It minimizes the total within-cluster variance.

4. *single-linkage* method:

opposite to complete-linkage method, it tends to produce long clusters

Ward-linkage and complete-linkage method are generally preferred

Other things to consider

Hierarchical clustering doesn't tell how many clusters there are, or where to cut the dendrogram to form clusters

The optimal number of cluster is subjective and depends on the method used for measuring similarities

simple and popular solution : inspect the dendrogram produced using hierarchical clustering to see if there is a suggested number of cluster

1. *elbow* method:
This method chooses a number of clusters when within-cluster sum of square (WSS) is minimized

2. *average silhouette* method
It shows how well each object lies within the cluster, the higher value indicates better clustering

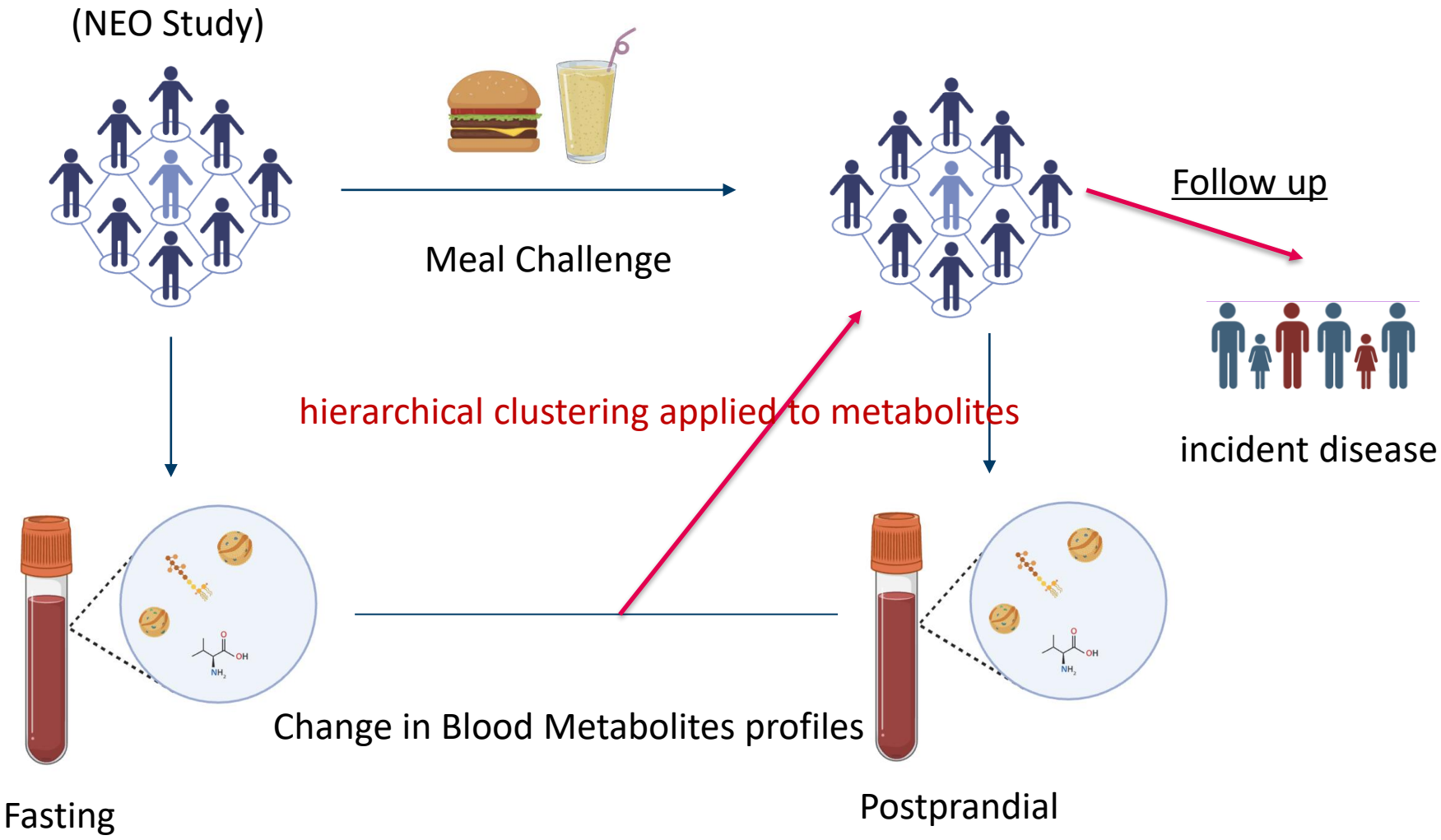
tion only. The optimal number of clusters for hierarchical clustering within a range of 2 to 10 was determined using the R package "NbClust", which uses the majority rule based on 30 different indices [21].

selected. To determine the optimal number of clusters, the NbClust R package was used. This package uses a majority rule based on 30 different indices to determine the optimal number of clusters in a given dataset [18].

Nbclust() function, can help determine the optimal number of clusters , based on 30 indices

index	the index to be calculated. This should be one of : "kl", "ch", "hartigan", "ccc", "scott", "marriot", "trcovw", "tracew", "friedman", "rubin", "cindex", "db", "silhouette", "duda", "pseudot2", "beale", "ratkowsky", "ball", "ptbiserial", "gap", "frey", "mcclain", "gamma", "gplus", "tau", "dunn", "hubert", "sdindex", "dindex", "sdbw", "all" (all indices except GAP, Gamma, Gplus and Tau), "alllong" (all indices with Gap, Gamma, Gplus and Tau included).

Combined with our "Own" Work

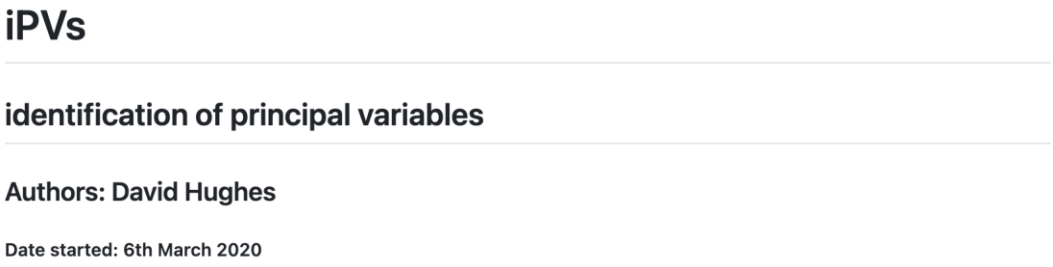


Metabolites Data

New problem ? (different from clinical characteristic)

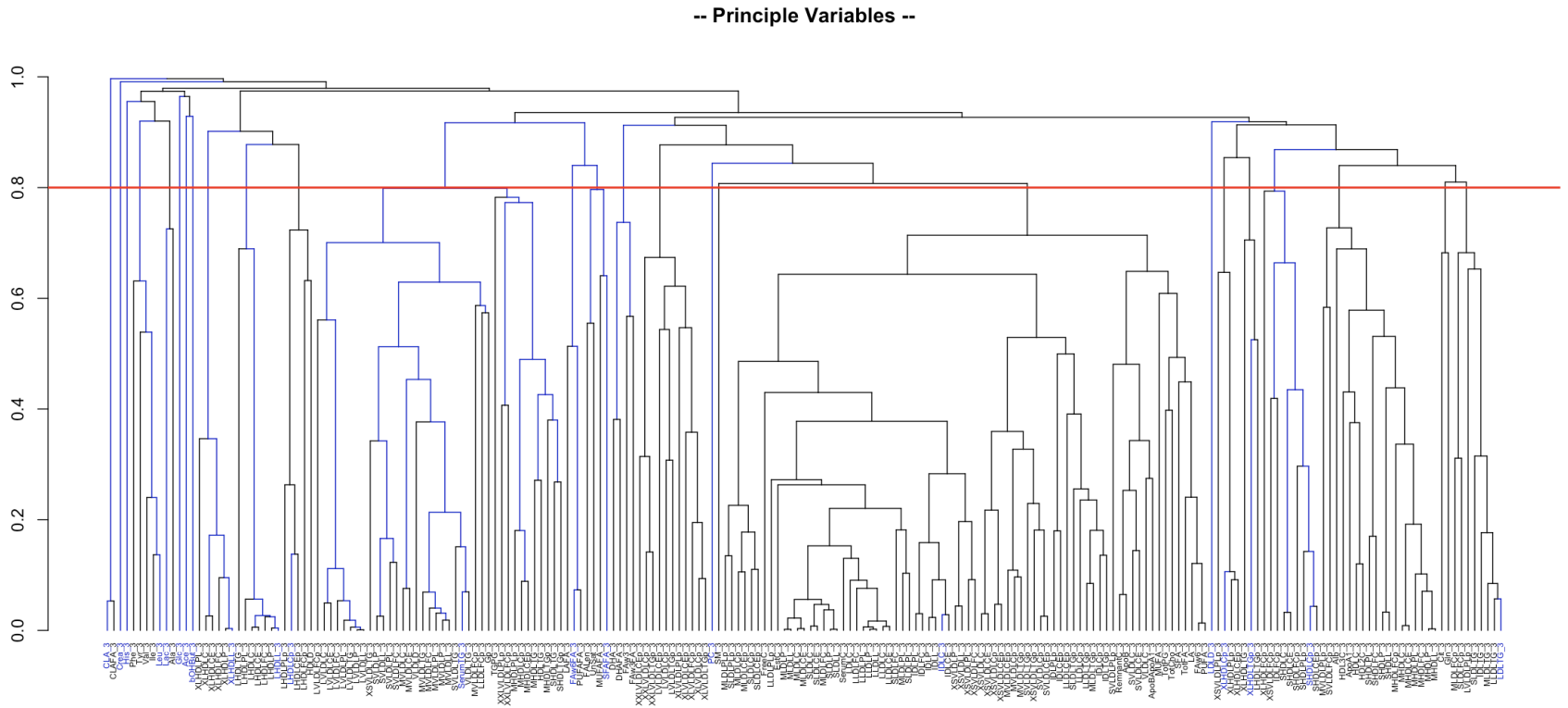
We have 200 metabolites level for each time point (fasting, postprandial), but most of these metabolites are highly correlated

High-dimensional data -> Select the independent variables / dimension reduction



https://github.com/hughesevoanth/iPVs

Results



Using hierarchical clustering method, the first step is to reduce the number of metabolites, keep the independent ones.

Based on these independent metabolites -> The number of clusters -> classify individuals

Leiden University
Medical Center

Thanks!

Keyong Deng

Department of Clinical Epidemiology