RUTGERS
Robert Wood Johnson Medical School
Cardiovascular Institute

Dept. of Statistics and Biostatistics

Personalized disease networks (PDN) for understanding and predicting cardiovascular diseases and other complex processes.

Javier Cabrera

Fei Wang

Nabil Adam, Jaideep Vaidya, Dhammika Amaratunga, William Kostis, John Kostis.

Research supported by a Chancellor seed grant from Rutgers Newark

Outline

- Big Data in medicine.
- MIDAS: cardiovascular diseases database.
- What are personalized disease networks?
- Analysis Plan.
- Enriched linear and non-linear PCA
- Proof of concept. Do PDN's help improve the prediction of cardiovascular outcomes.

Big Data in medicine

Electronic Medical Records → **Big Data**

- Databases of hospitalizations: ICD-9 or ICD-10 **diagnosis** and **procedural** codes with dates.
- Labs Data: Test results
- Health Insurance Providers: **Costs** of treatment, hospitalization, procedures and **drugs.**
- Medical Charts: Doctors handwriting, text on paper records. Patient data from devices: collected in real time by devices and uploaded into the cloud has many new applications.
- Objectives: Patient Summary Charts for doctor's uses.

Medical Research → Analysis of Big Data

The MIDAS dataset

MIDAS: All hospitalizations for cardiovascular diseases in New Jersey since 1980.
15 million hospitalizations
5 million patients with cardiovascular diseases.
MIDAS is housed at Rutgers Cardiovascular Institute

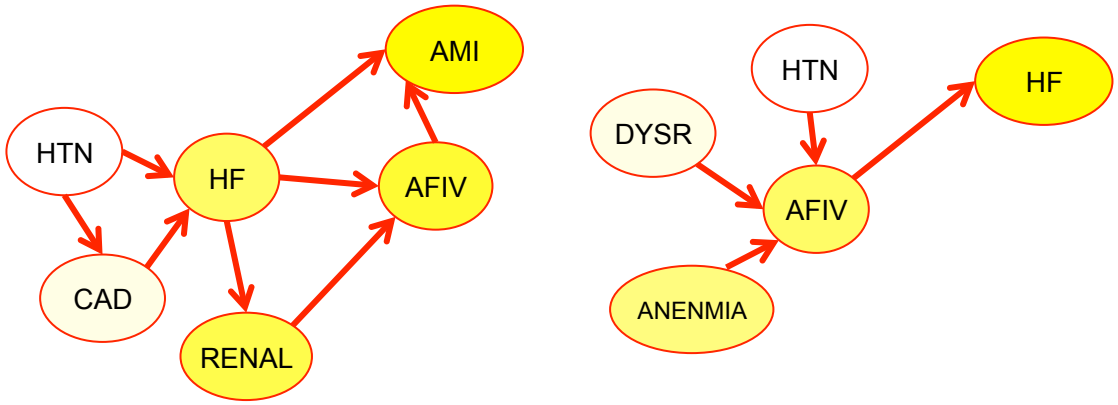
Plan: From MIDAS database PDN for cardiovascular diseases.
MIDAS database:

- Demographical information of the patients
- ICD9 codes for the diseases and the procedural codes and the corresponding date of the disease/procedure.
- Negotiating data sharing agreement with BlueCross/S

What are personalized disease networks?

- Personalized (patient) disease networks(PDN),
- Represent the evolution of a disease on a patient across subsequent diagnoses and medical interventions.
- Evolutionary steps of various cardiovascular diseases and procedures
- Network: represents a data based construction in a form of a connected graph to derive the disease evolution of each individual patient
- To build a PDN we use a list of events and their date
- Each event connected to another by a directed edge if the dates fulfill a certain criteria.

Example: PDN's of 2 patients that diagnosed with Atrium Fibrillation and Heart Failure. The arrows represent the order of events.



RUTGERS

Example of data from MIDAS

MR	CMTHY	CHD	HTN	DM	NEO	COPD	RENAL	STROKE	AMI
8/23/06	10/4/11	8/29/95	8/30/95	2/27/98	NA	3/20/03	9/9/10	NA	NA
5/22/95	NA	5/22/95	1/13/95	NA	NA	5/14/95	11/30/10	4/27/07	NA
NA	NA	9/25/07	9/18/99	2/12/07	NA	9/18/99	4/21/09	NA	NA
12/27/11	NA	10/14/97	10/19/01	8/3/07	3/8/03	1/22/02	10/14/97	10/19/01	4/30/03
NA	8/14/08	12/19/00	10/18/01	12/19/00	11/18/10	12/15/01	11/5/08	NA	NA
NA	NA	10/27/97	10/21/04	10/31/97	NA	NA	NA	NA	NA
NA	4/3/95	4/3/95	4/3/95	10/16/95	NA	3/31/02	9/3/09	NA	NA
NA	NA	11/22/96	11/22/96	NA	8/29/05	3/9/11	NA	NA	NA
NA	NA	5/2/95	2/25/98	NA	1/15/08	2/25/98	9/8/10	NA	NA
NA	11/18/07	6/2/06	3/26/02	3/26/02	NA	NA	11/18/07	NA	NA

Column labels correspond to ICD9 codes and dates .

Rules for building Personalized disease networks

Rule: Arrow connecting a Diagnosis/Procedure A to another Diagnosis/Procedure B where A and B are observed at times T_A and T_B

1. Naive Rules

If $T_A < T_B$ then A → B , If $T_A = T_B$ it will depend on A and B.

If $0 < T_B - T_A < K$ then A → B , If $T_A = T_B$ it will depend on A and B.

2. Flexible Rule:

If $0 < T_B - T_A < K(A,B)$ then A → B, If $T_A = T_B$ it will depend on A and B

To Calculate K(A,B) we may use several methods

(i) Expert Physicians : (It is been currently performed)

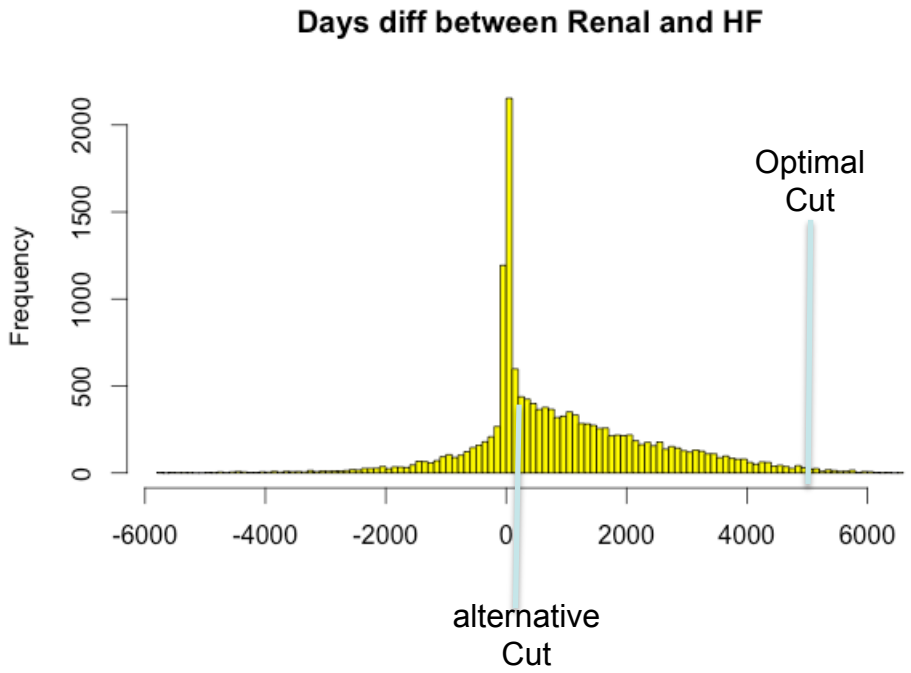
(ii) Find K(A,B) based on the prediction of some response.

3. Expert Physicians input. So far we got 3 physicians to give info.

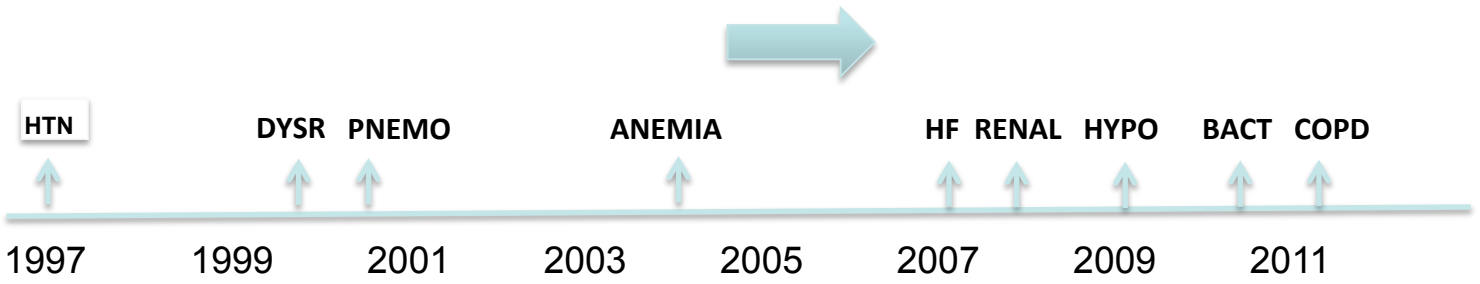
Expert Physicians input

So far we got 4 physicians: Drs J Kostis, W Kostis, Dr Moreira, Dr Pantazopoulos from CVI/Cardiology at Rutgers

Optimal cuts presented to for input and a decision is reached
Example: Renal dysfunction (kidney failure, dialysis) -> Heart Failure

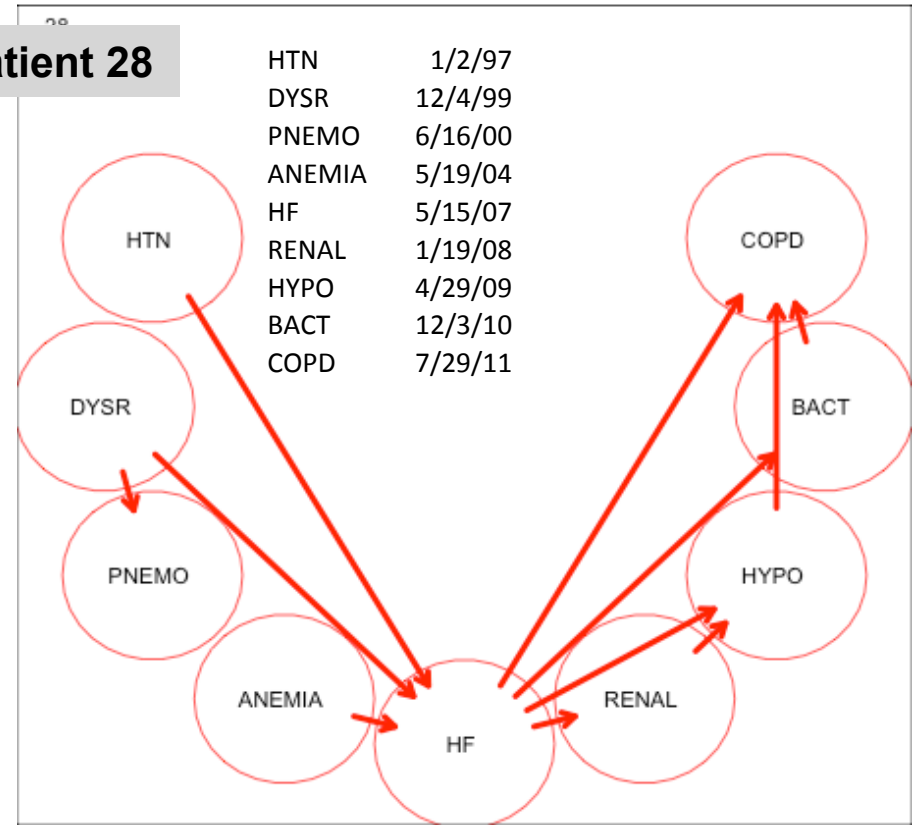


Personalized disease networks(Flexible rule . 2)

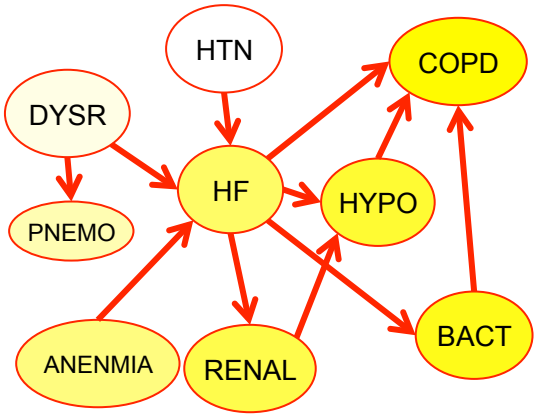


Patient 28

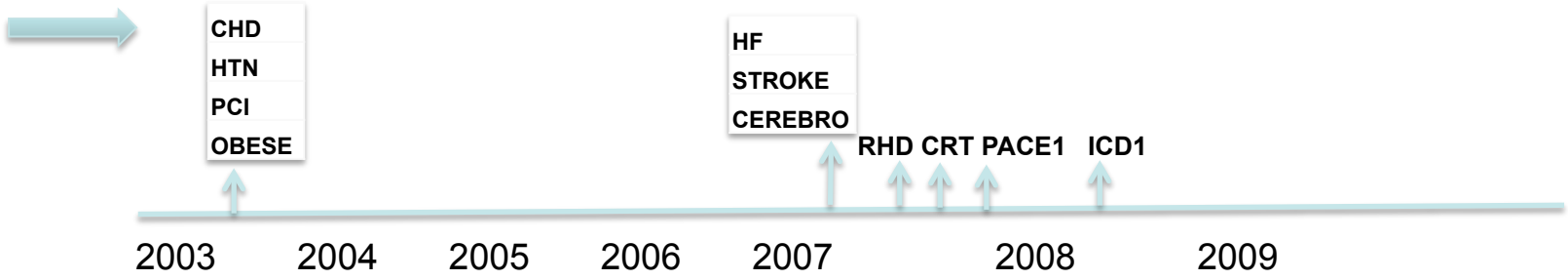
HTN	1/2/97
DYSR	12/4/99
PNEMO	6/16/00
ANEMIA	5/19/04
HF	5/15/07
RENAL	1/19/08
HYPO	4/29/09
BACT	12/3/10
COPD	7/29/11



Patient 28



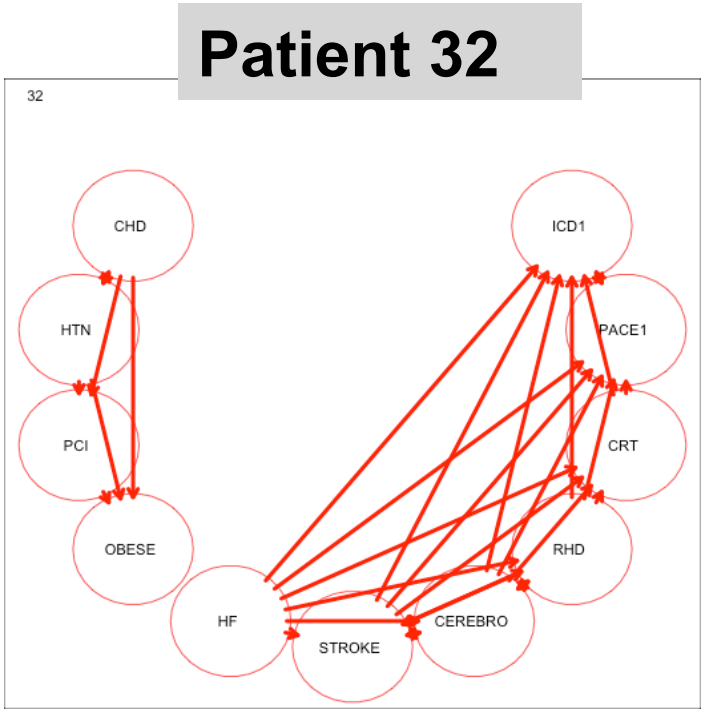
Naïve rule for building PDE's : K = 1000 days



Example of personalized (patient 32) disease networks(PDN)

Patient 32

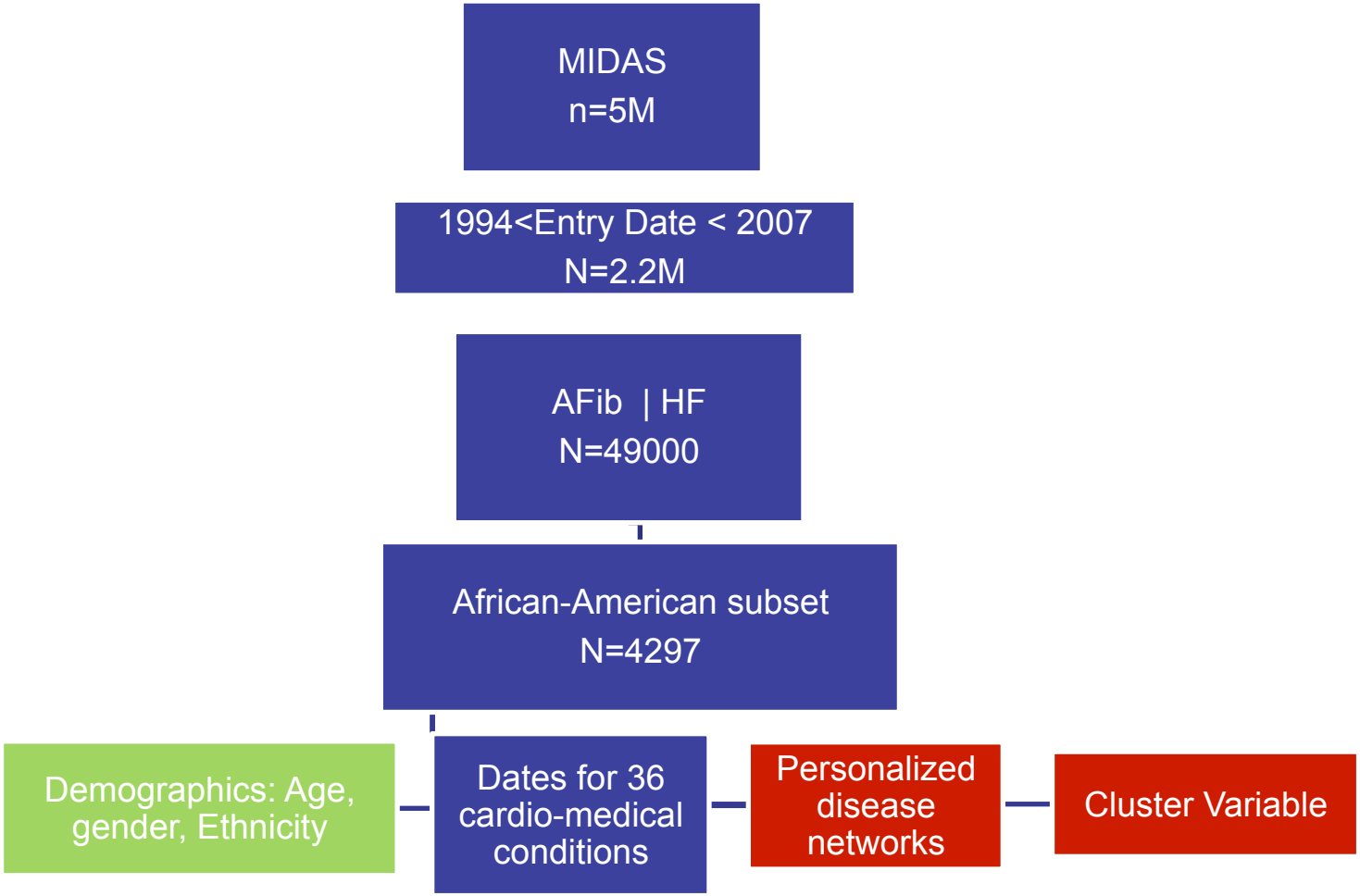
CODE	DATE
CHD	4/12/03
HTN	4/12/03
PCI	4/12/03
OBESE	4/12/03
HF	6/7/07
STROKE	6/7/07
CEREBRO	6/7/07
RHD	6/28/07
CRT	9/19/07
PACE1	9/27/07
ICD1	5/28/08



Data Matrix for personalized disease networks(PDN),

AFIB to HF	HF to AFIB	AFIB to MR	MR to AFIB	AFIB to CMTHY	CMTHY to AFIB	AFIB to CHD	CHD to AFIB	AFIB to HTN	HTN to AFIB
0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	1	0	1	0	1
0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	1	1	0
0	1	0	0	0	1	0	1	0	1
0	1	0	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	1	0	1	0	1	0	1	0	1
0	1	0	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Data Subsets



Analysis Plan

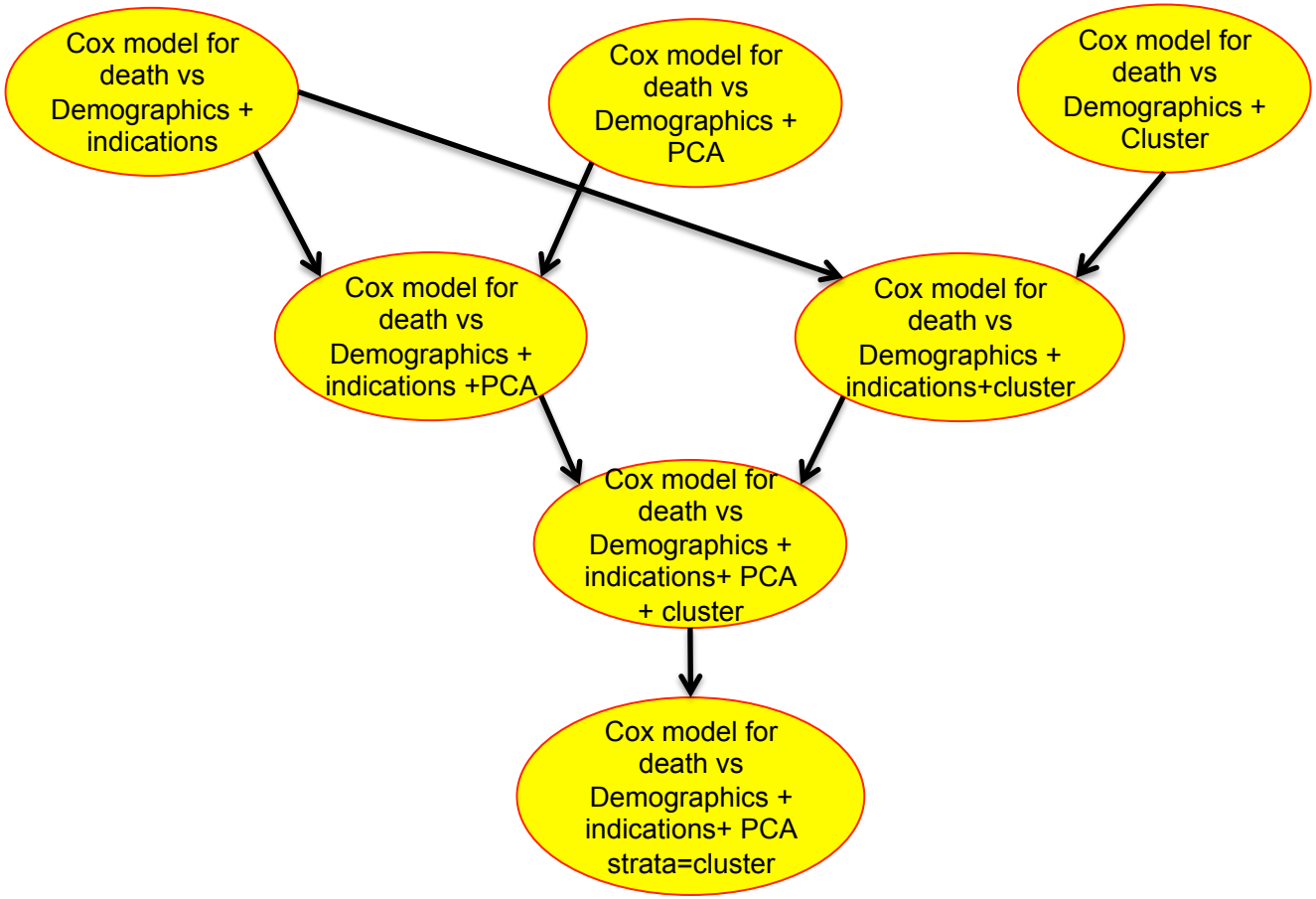
Demographics: age, gender, ethnicity

Dates for 36 cardio-medical indications

Personalized disease networks

Enriched PCA

Cluster Factor



Enriched PCA Scores

Enriched PCA scores. Amaratunga, Cabrera, Shkedy(2014)

Enriched PCA: $\underset{G\times G}{W} = \text{Diag}(W_i)$

n>=G

$$\underset{n\times G}{X^*} = \underset{n\times G}{X}\ \underset{G\times G}{W} = \underset{n\times G}{U^*}\ \underset{G\times G}{D^*}\ \underset{G\times G}{V^{*\prime}}$$

Covariance or Correlation $\underset{G\times G}{R^*} = \underset{G\times G}{V^*}\ \underset{G\times G}{D^{*2}}\ \underset{G\times G}{V^{*\prime}}$

n<G

$$\underset{G\times n}{X^*} = \underset{G\times G}{W}\ \underset{G\times n}{X} = \underset{G\times n}{U^*}\ \underset{n\times n}{D^*}\ \underset{n\times n}{V^{*\prime}}$$

Covariance or Correlation $\underset{G\times G}{R^*} = \underset{G\times n}{U^*}\ \underset{n\times n}{D^{*2}}\ \underset{n\times G}{U^{*\prime}}$

W matrix can be calculated as a measure of importance of a variable.

Enriched PCA scores: Amaratunga, Cabrera, Shkedy(2014)

How to determine the weights? $W = Diag(W_i)$

The idea is to reduce the effect of the spurious signal that is create by G large number of variables,

Use any model $Y = f(X_i)$ and obtain the model *p-value* p_i .

If G is large convert *p-values* into *q-values* q_i .

Define the weights:

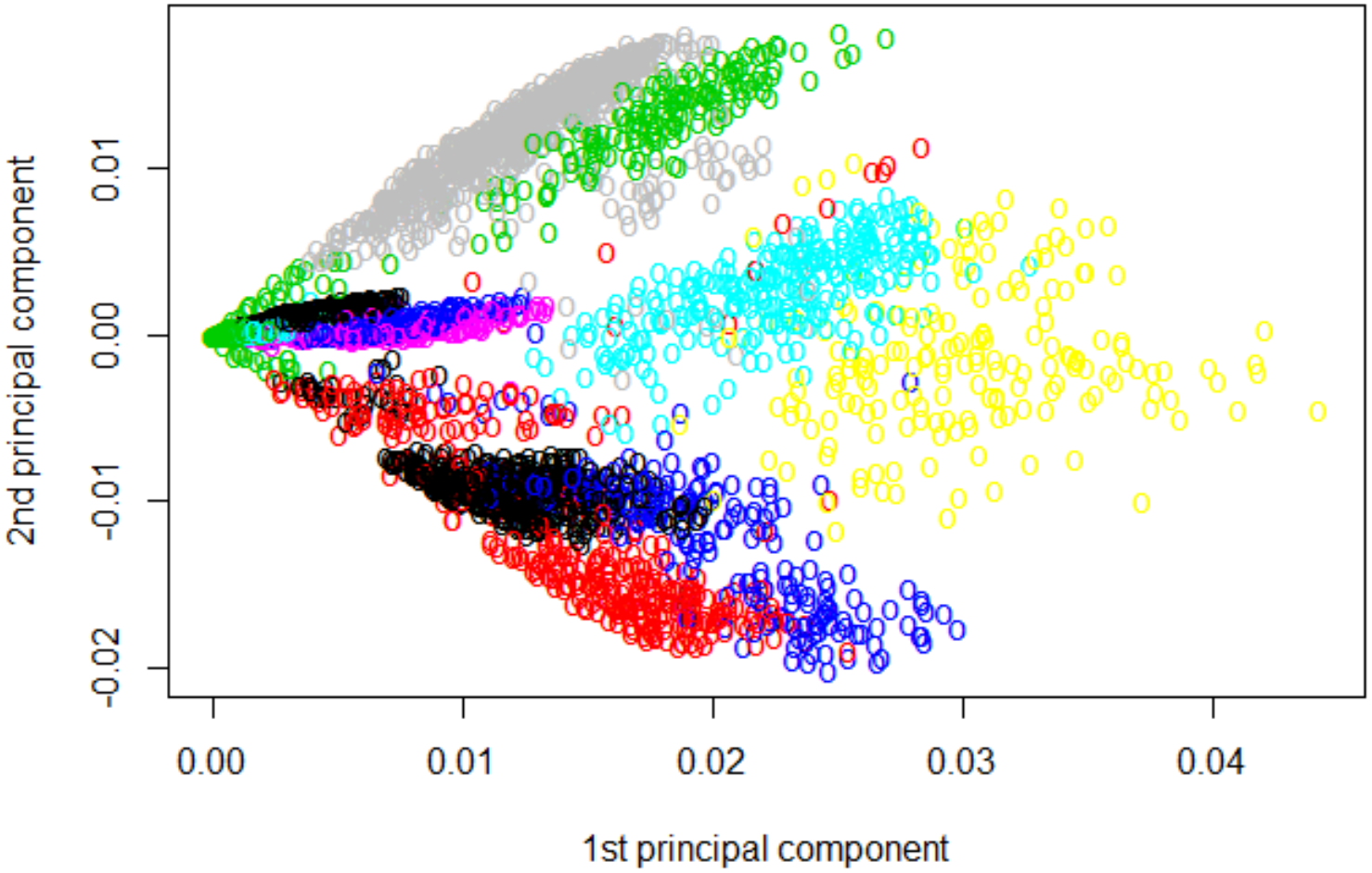
$$W_i = -log(q_i)$$

$$W_i = 1/q_i$$

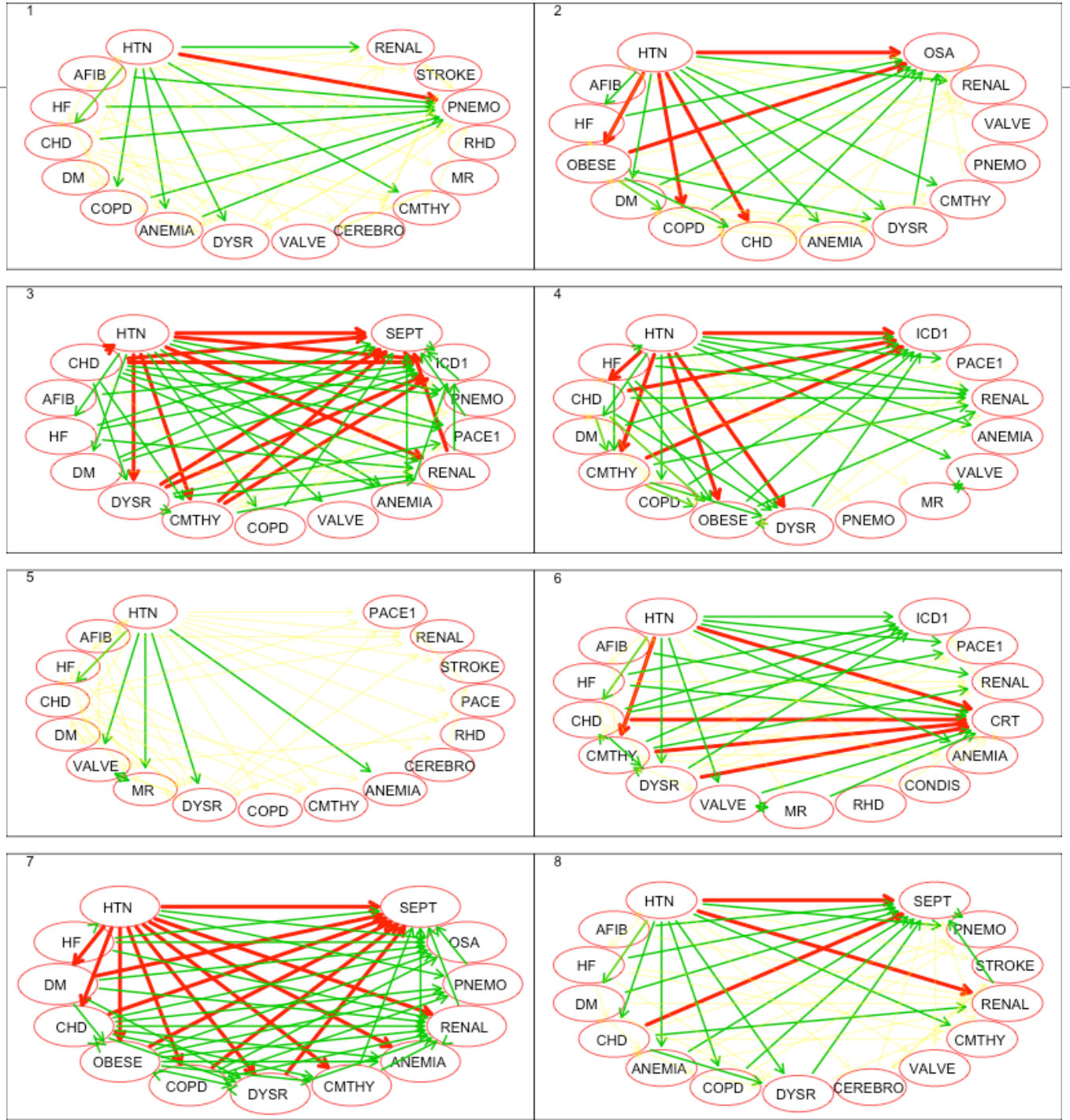
Enriched PCA and Clustering

Enriched PCA: 43 PCA

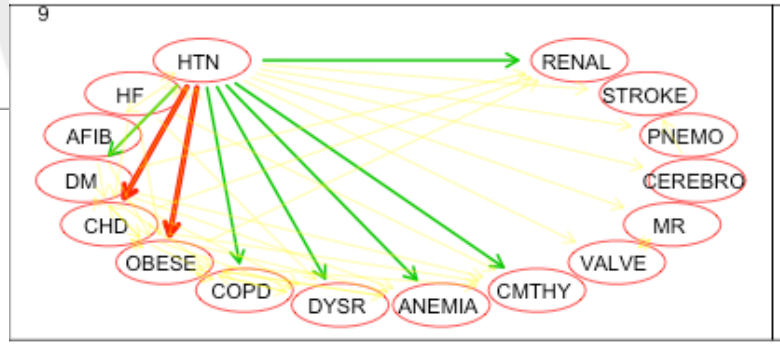
Hierarachical Clustering: 13 Clusters.

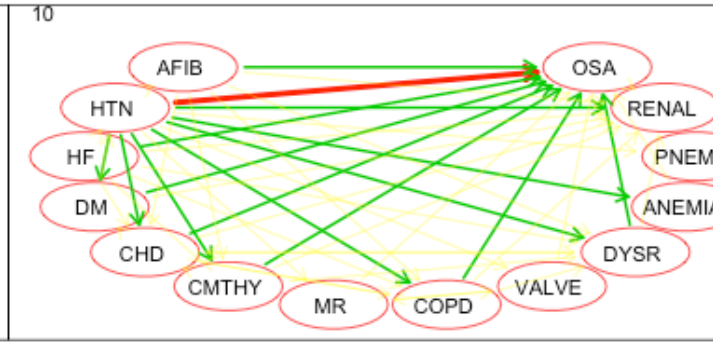


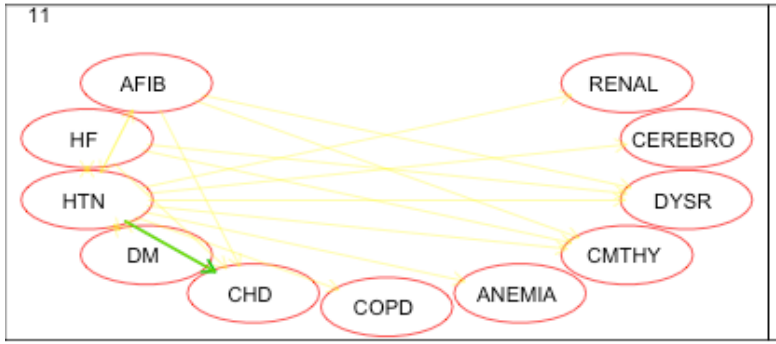
RUTGERS

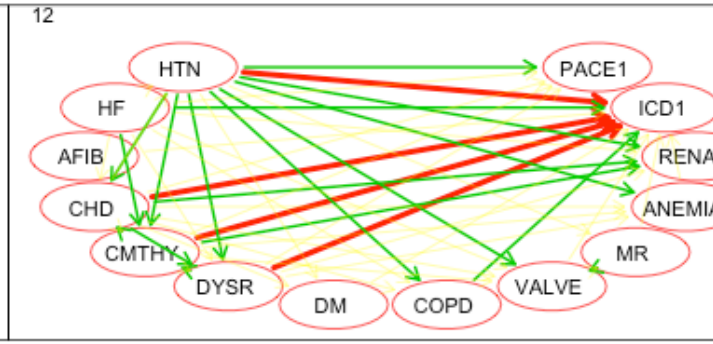

1
HTN
AFIB
HF
CHD
DM
COPD
ANEMIA
DYSR
VALVE
CEREBRO
CMTHY
MR
RHD
PNEMO
STROKE
RENAL
2
HTN
AFIB
HF
OBESE
DM
COPD
CHD
ANEMIA
DYSR
CMTHY
PNEMO
VALVE
RENAL
OSA
3
HTN
CHD
AFIB
HF
DM
DYSR
CMTHY
COPD
VALVE
ANEMIA
RENAL
PACE1
PNEMO
ICD1
SEPT
4
HTN
HF
CHD
DM
CMTHY
COPD
OBESE
DYSR
PNEMO
MR
VALVE
ANEMIA
RENAL
PACE1
ICD1
5
HTN
AFIB
HF
CHD
DM
VALVE
MR
DYSR
COPD
CMTHY
ANEMIA
CEREBRO
RHD
PACE
STROKE
RENAL
PACE1
6
HTN
AFIB
HF
CHD
CMTHY
DYSR
VALVE
MR
RHD
CONDIS
ANEMIA
CRT
RENAL
PACE1
ICD1
7
HTN
HF
DM
CHD
OBESE
COPD
DYSR
CMTHY
ANEMIA
RENAL
PNEMO
OSA
SEPT
8
HTN
AFIB
HF
DM
CHD
ANEMIA
COPD
DYSR
CEREBRO
VALVE
CMTHY
RENAL
STROKE
PNEMO
SEPT

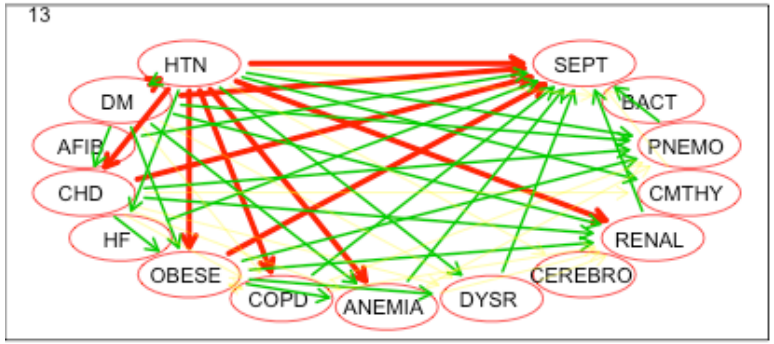
RUTGERS


9
HTN
HF
AFIB
DM
CHD
OBESE
COPD
DYSR
ANEMIA
CMTHY
VALVE
MR
CEREBRO
PNEMO
STROKE
RENAL

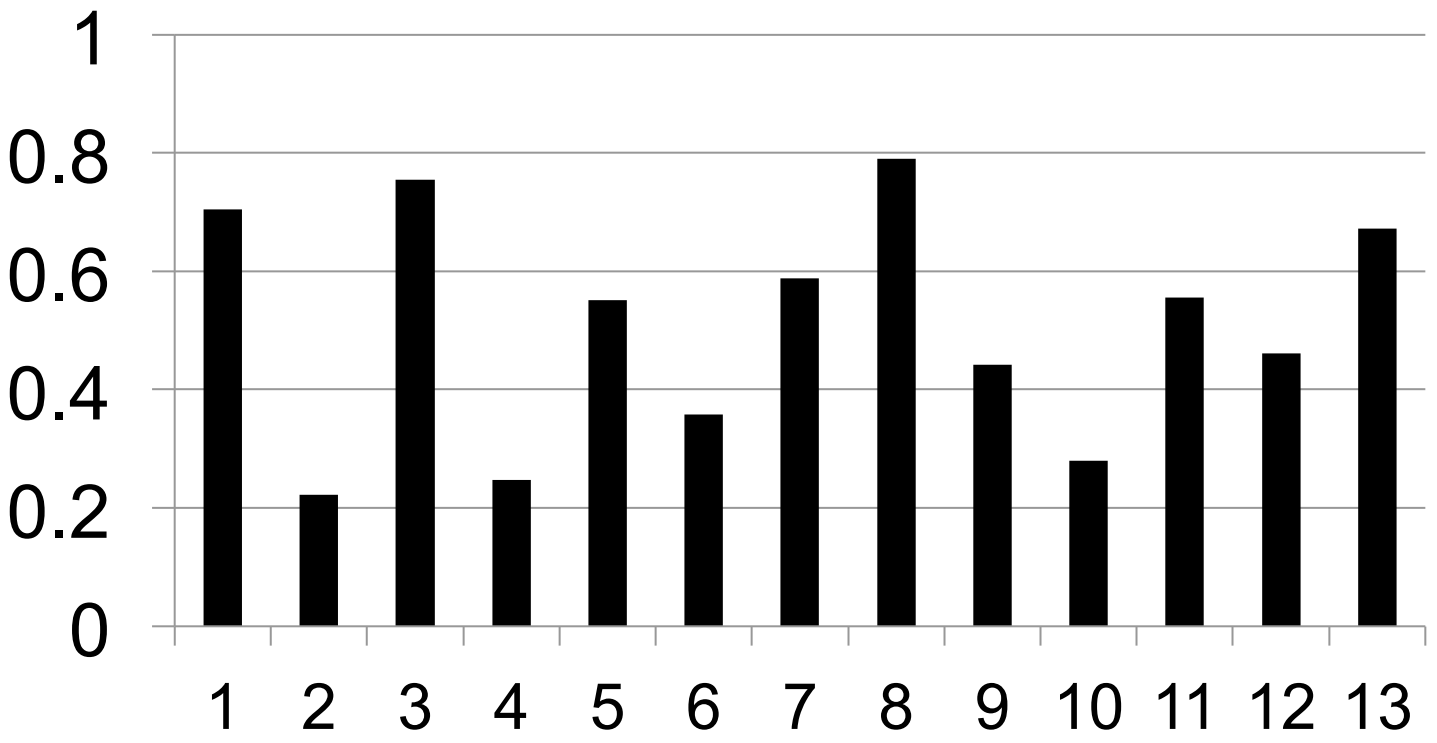

10
AFIB
HTN
HF
DM
CHD
CMTHY
MR
COPD
VALVE
DYSR
OSA
RENAL


11
AFIB
HF
HTN
DM
CHD
COPD
ANEMIA
CMTHY
DYSR
CEREBRO
RENAL


12
HTN
HF
AFIB
CHD
CMTHY
DYSR
DM
COPD
VALVE
MR
ICD1
PACE1


13
HTN
DM
AFIB
CHD
HF
OBESE
COPD
ANEMIA
DYSR
CEREBRO
RENAL
CMTHY
PNEMO
BACT
SEPT

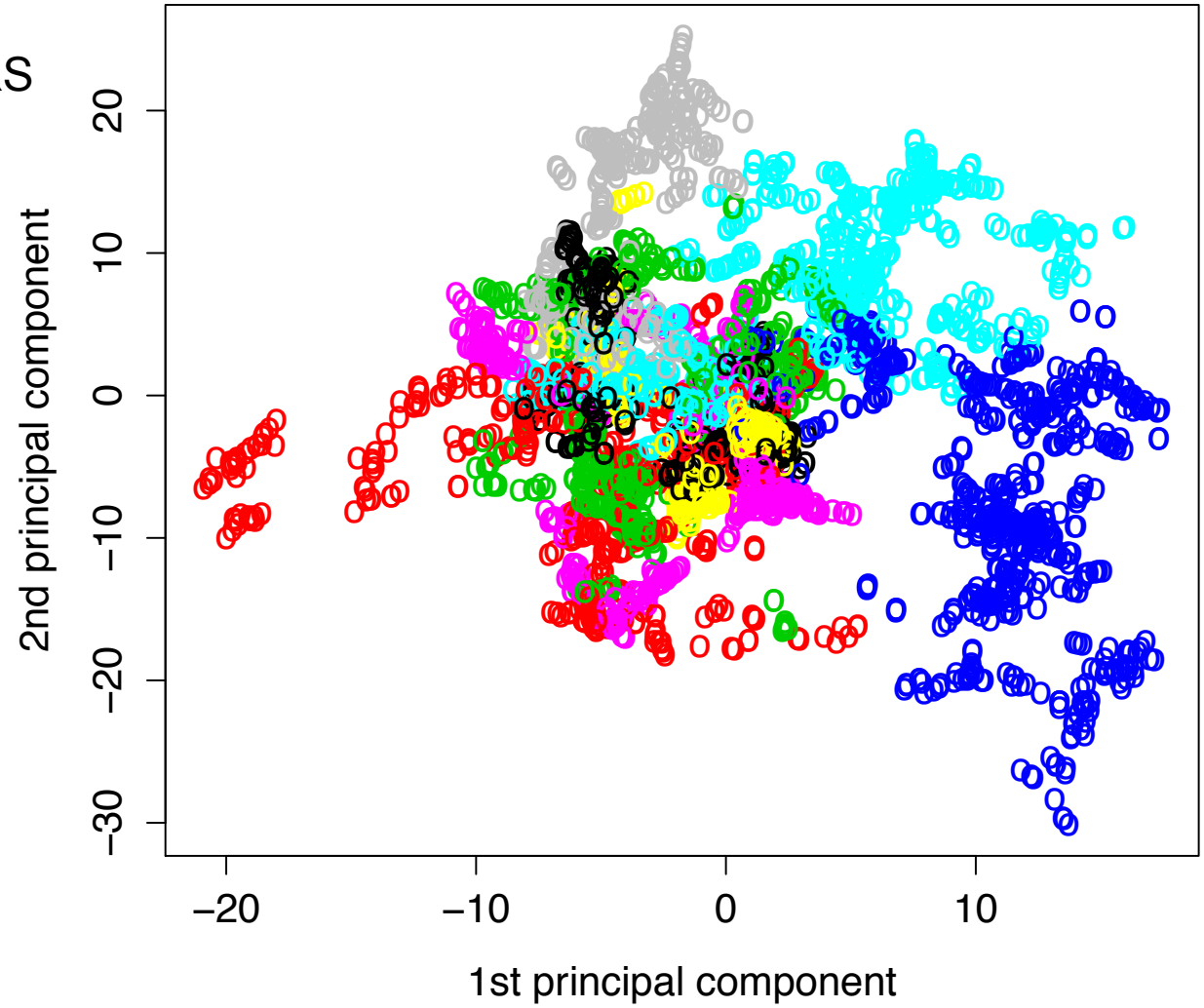
Mortality in each cluster

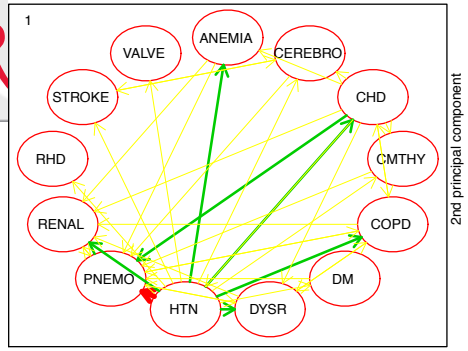


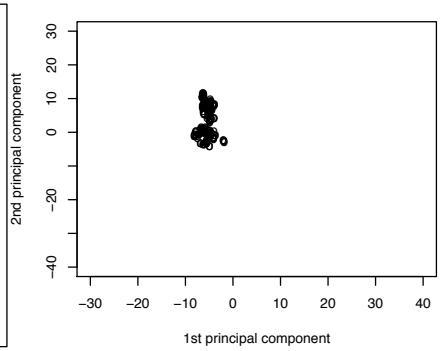
H Clustering with enriched nonlinear PCA.

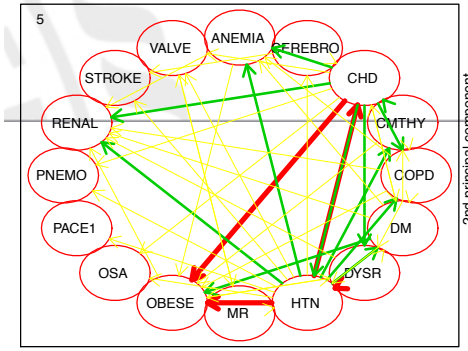
tsna: R package L.J.P. van der Maaten e.a.(2008)
Based on k-nearest neighbors.

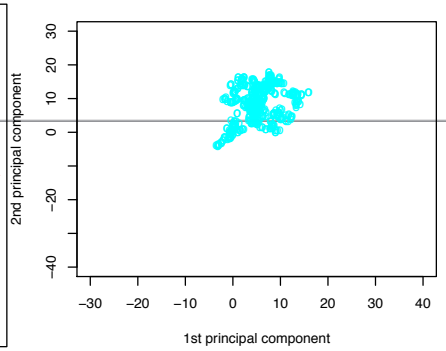
16 CLUSTERS

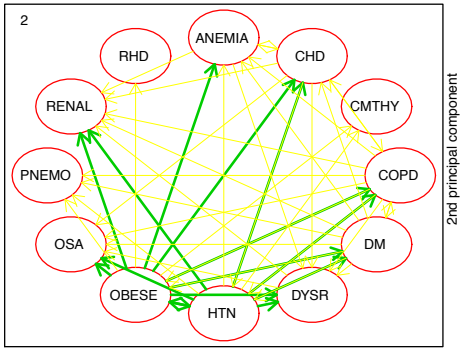


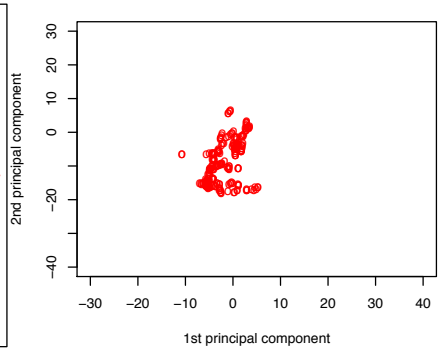

1
ANEMIA
VALVE
CEREBRO
STROKE
CHD
RHD
CMTHY
RENAL
COPD
PNEMO
DM
HTN
DYSR

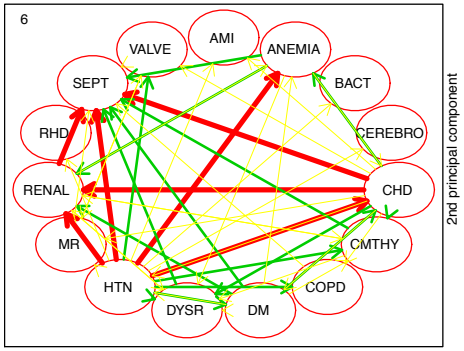

2nd principal component
1st principal component

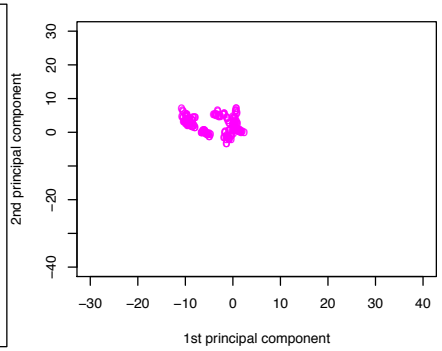

5
ANEMIA
VALVE
CEREBRO
STROKE
CHD
RENAL
CMTHY
PNEMO
COPD
PACE1
DM
OSA
DYSR
OBESE
MR
HTN

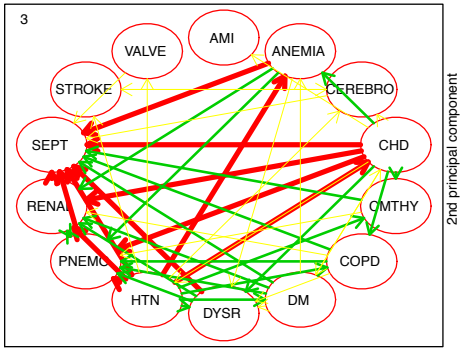

2nd principal component
1st principal component

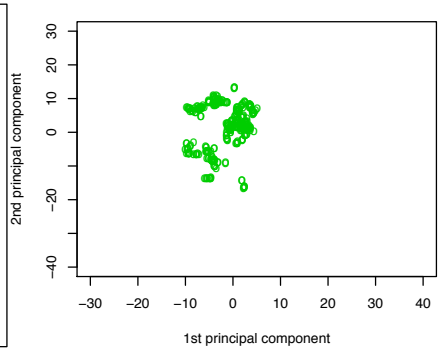

2
ANEMIA
RHD
CHD
RENAL
CMTHY
PNEMO
COPD
OSA
DM
OBESE
DYSR
HTN

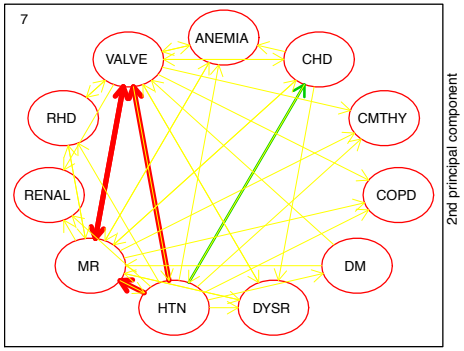

2nd principal component
1st principal component

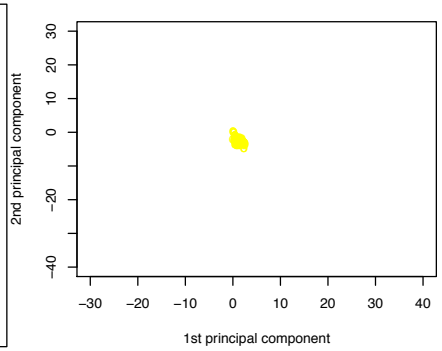

6
VALVE
AMI
ANEMIA
SEPT
BACT
RHD
CEREBRO
RENAL
CHD
MR
CMTHY
HTN
COPD
DYSR
DM

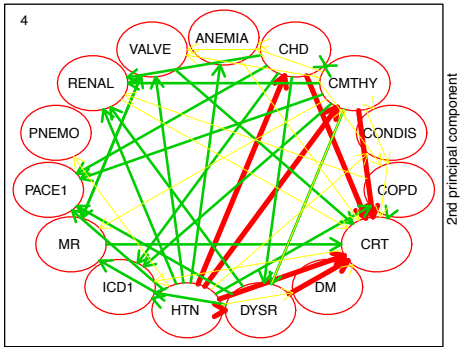

2nd principal component
1st principal component

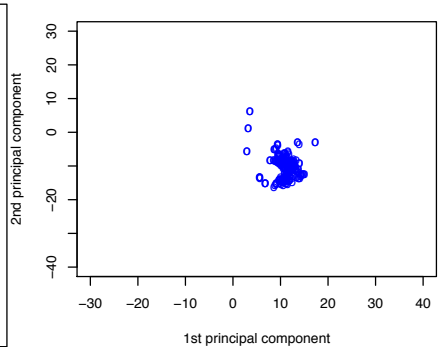

3
VALVE
AMI
ANEMIA
STROKE
CEREBRO
SEPT
CHD
RENAL
CMTHY
PNEMO
COPD
HTN
DM
DYSR

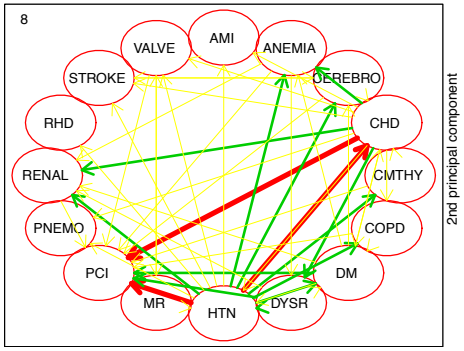

2nd principal component
1st principal component

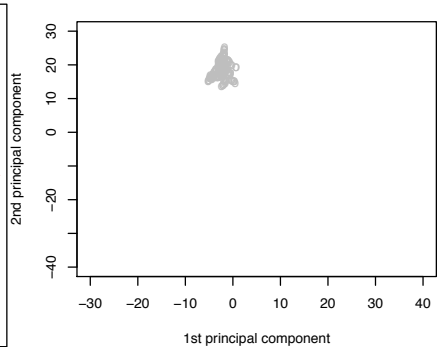

7
ANEMIA
VALVE
CHD
RHD
CMTHY
RENAL
COPD
MR
DM
HTN
DYSR


2nd principal component
1st principal component

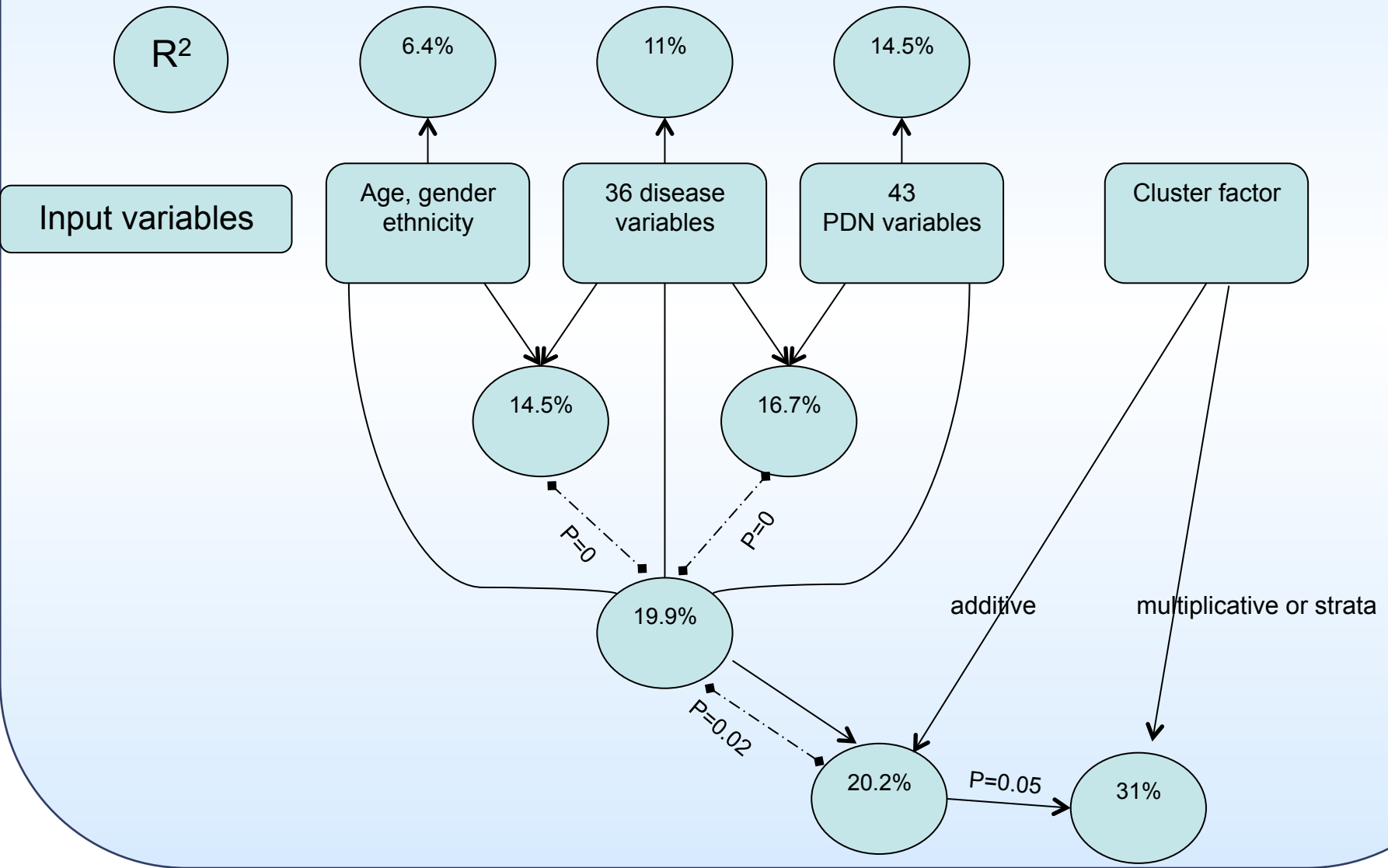

4
ANEMIA
VALVE
CHD
RENAL
CMTHY
PNEMO
CONDIS
PACE1
COPD
MR
CRT
ICD1
DM
HTN
DYSR


2nd principal component
1st principal component

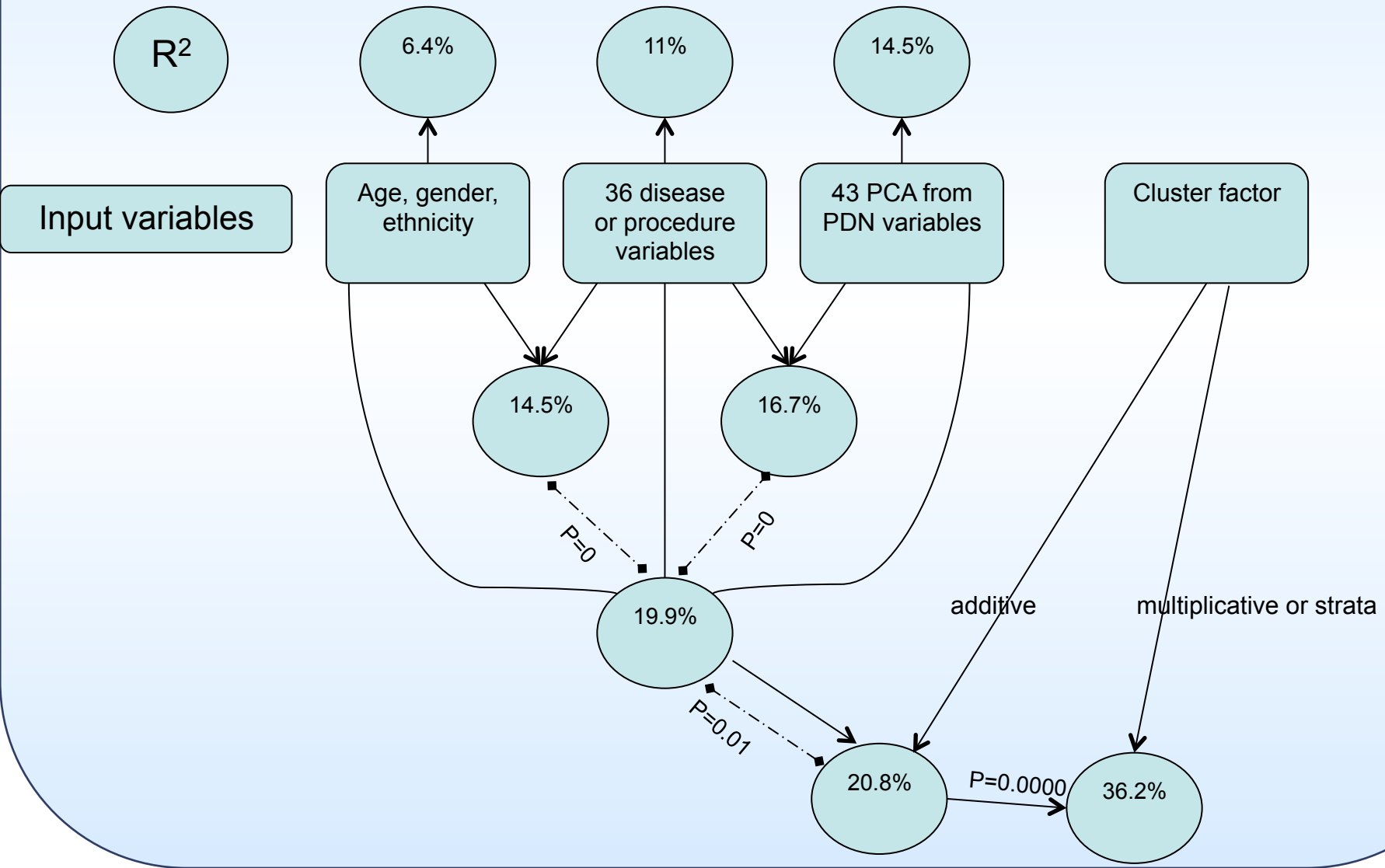

8
VALVE
AMI
ANEMIA
STROKE
CEREBRO
RHD
CHD
RENAL
CMTHY
PNEMO
COPD
PCI
DM
MR
HTN
DYSR


2nd principal component
1st principal component

Modeling all Mortality by using Cox PH models Enriched PCA



Modeling all Mortality by using Cox PH models Enriched nonlinear PCA



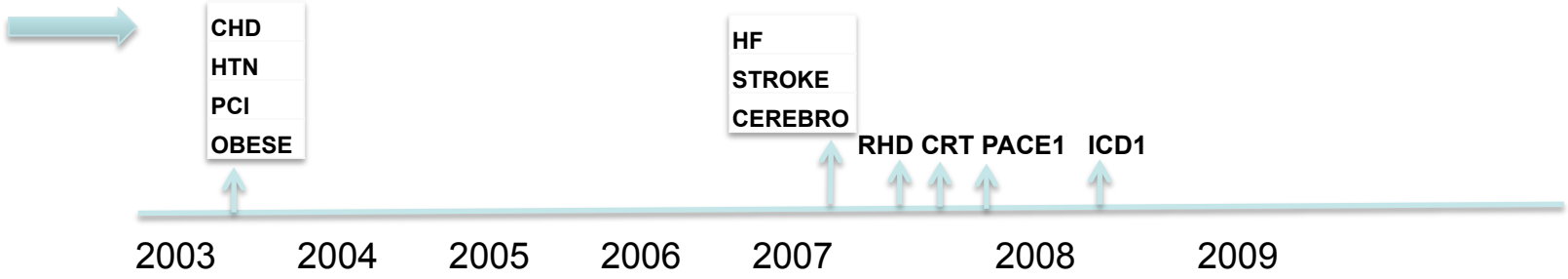
Conclusions

- We develop a novel algorithm for building patient level networks for prediction of medical outcomes
- We develop cluster methodologies (Hierarchical Cluster using enriched PCA scores) for individual network patient data and ways to summarize the clusters and study the within variability.
- In clinical outcome prediction, PDN's and cluster factors as input variables significantly improved goodness of fit of survival models in our dataset.

Future directions

- Optimize the network visualization. Time sequence, from horse shoes to more complex spatial network display .
- Network summaries. Median, mean, quantiles. Variability.
- Prediction of cardiovascular outcomes like cardiovascular death, Heart attack, Stroke, etc using the method. Training and testing sets. Cross validation.
- Apply the method for the larger subsets of MIDAS for other demographics and conditions.
- Simulations and
- Build an R package to implement our methodology.

Construction of Personalized disease networks



Example of personalized (patient 32) disease networks(PDN)

Patient 32

CODE	DATE
CHD	4/12/03
HTN	4/12/03
PCI	4/12/03
OBESE	4/12/03
HF	6/7/07
STROKE	6/7/07
CEREBRO	6/7/07
RHD	6/28/07
CRT	9/19/07
PACE1	9/27/07
ICD1	5/28/08