UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Institut Català de la Salut

Clustering Patients with Tensor Decomposition

Matteo Ruffini [1] Ricard Gavaldà [1] Esther Limón [2]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain, [2]Institut Català de la Salut

Objective

Task: to segment patients in groups with similar clinical profiles.
Motivation:

1. Similar patients → Similar cares.
2. Find recurrent comorbidities.
3. Assigning and planning resources.

The data

All hospitalizations in Catalonia, Spain, in 2016. Each row is a patient's visit containing up to 10 ICD-9 diagnostics of the patient. Data can be converted into a binary matrix $\mathcal{X}$: $(\mathcal{X})_{i,j} = 1$ if patient i has disease j.

	Disease 1	Disease 2	Disease 3	...
Patient 1	1	1	0	...
Patient 2	0	1	1	...
...	...	...	...	...

Objective: cluster the rows of the matrix.
Challenges: sparse and high dimensional data.

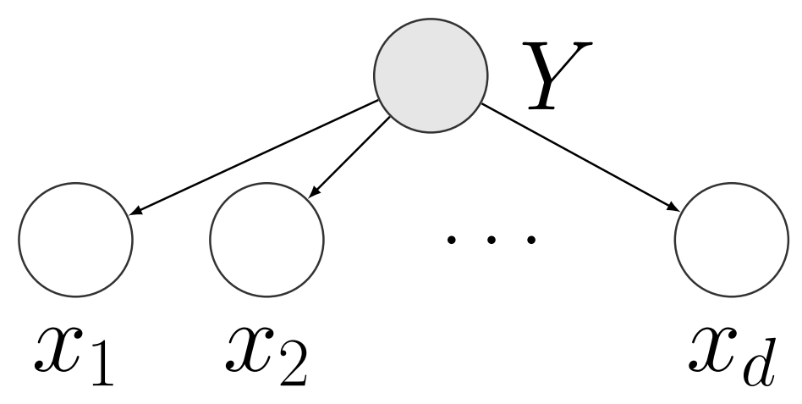
State of the art

Distance-based clustering methods [1](*k-means, k-medioids, single linkage*) : poor performances on high dimensional sparse data; manual definition of a distance function.
Tensor factorization for patients phenotyping[2, 3, 4] *(Limestone, Marble, Rubik...)*: require many sources of information; no generative model is assumed.

Proposed approach: mixture models

Data is modeled as a mixture of independent Bernoulli variables (Naïve Bayes model)

- Latent state → Medical status of a patient.
- Observed diseases depend on patient status.
- Once in a status, diagnostics are independent.



- $Y \in [k]$ is a latent variable.
- $X = (x_1, \dots, x_d)$ are observable, conditionally independent on Y.

Main advantages

- No distance is required.
- Generative model → clear interpretation.
- Flexible: works with single sources of data.

Mixture Model Clustering

Clustering: given a dataset, assign each sample to the most likely mixture component.

Given the **parameters of the mixture**:

$$\omega_j = \mathbb{P}(Y = j)$$
$$\mu_i = \mathbb{E}(X|Y = j), \quad M = [\mu_1|, ..., |\mu_k] \in \mathbb{R}^{d\times k}$$

and a sample $X = (x_1, ..., x_d)$,
the clustering rule is

$$Cluster(X) = argmax_{j=1,...,k}(\mathbb{P}(Y = j|X))$$

where, for mixture of independent Bernoulli

$$\mathbb{P}(Y = j|X) \propto \omega_j \prod_{i=1}^{d} (\mu_j)_i^{x_i}(1 - (\mu_j)_i)^{1-x_i}$$

We need the mixture parameters, (M, ω).

Learning mixture parameters: method of moments

Method of moments

A general approach from [5]:

- Estimate from data the *moments:*

$$M_1 := \textstyle\sum_{i=1}^{k} \omega_i\, \mu_i \in \mathbb{R}^d \quad (1)$$
$$M_2 := \textstyle\sum_{i=1}^{k} \omega_i\, \mu_i \otimes \mu_i \in \mathbb{R}^{d\times d} \quad (2)$$
$$M_3 := \textstyle\sum_{i=1}^{k} \omega_i\, \mu_i \otimes \mu_i \otimes \mu_i \in \mathbb{R}^{d\times d\times d} \quad (3)$$

- Obtain mixture's parameters with tensor decomposition on the moments:

$$\mathcal{TD}(M_1, M_2, M_3) \to (M, \omega)$$

  The decomposition of M_3 constrained to (1) and (2) is unique if M has full rank. There exist methods that, exploiting the structure of the moments, get (M, ω) efficiently.
- Improve the obtained parameters with EM.

Problem: No easy ways to estimate M_2 and M_3 for a mixture of independent Bernoulli.

Proposed approach: approximate estimates. Define

$$\tilde{M}_j^{(N)} := \sum_{i=1}^{N} \frac{X^{(i)\otimes j}}{N}, \quad \text{for} \;\; j = 1, 2, 3$$

where $X^{(i)}$ are our iid records. We can demonstrate that, for big enough samples, for $j = 1, 2, 3$

$$\Delta_j = ||\tilde{M}_j^{(N)} - \tilde{M}_j||_F$$

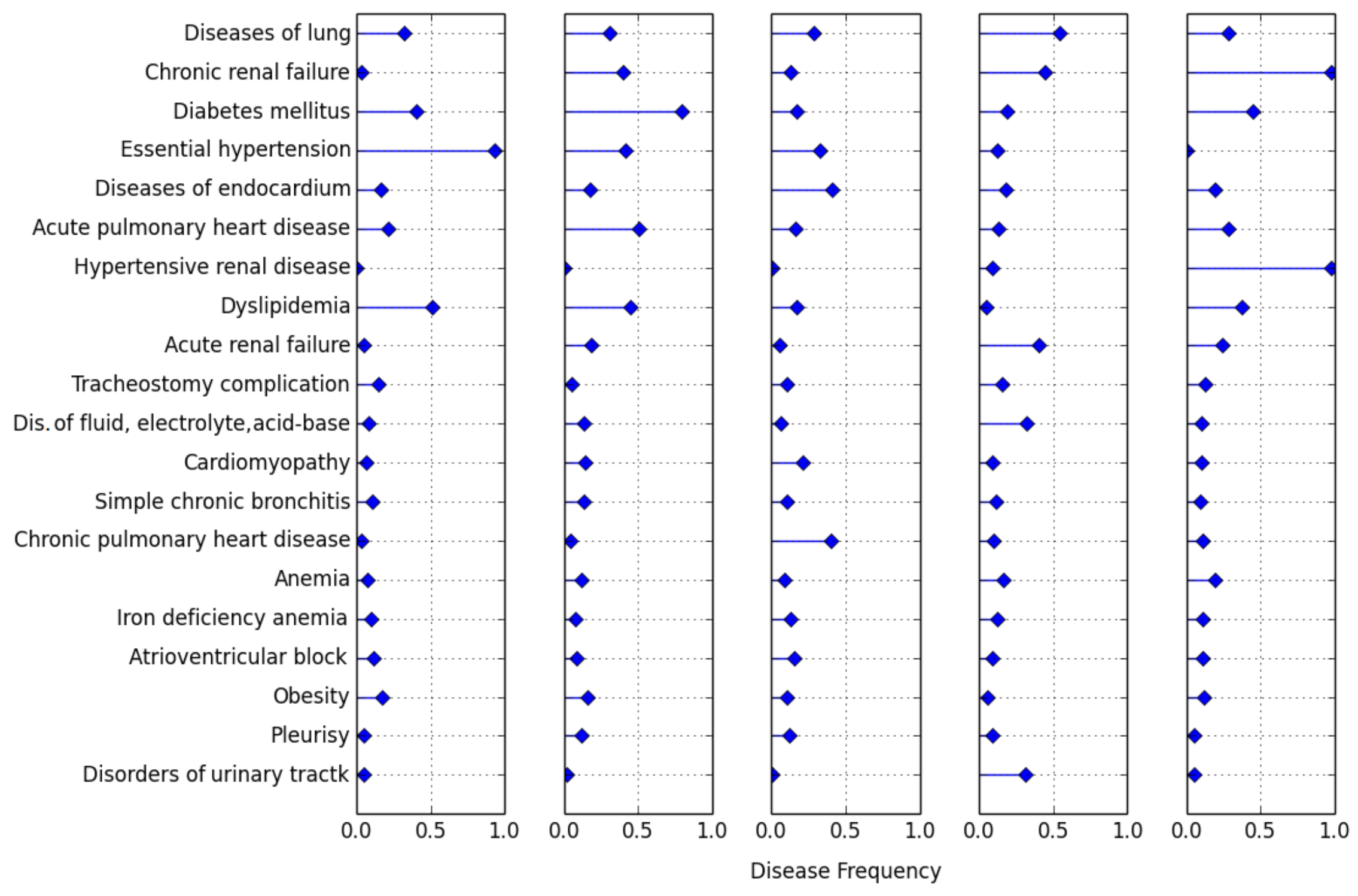
is small. This means that $\tilde{M}_j^{(N)}$ and $\tilde{M}_j$ are asymptotically close (but not equal, here is the approximation!).

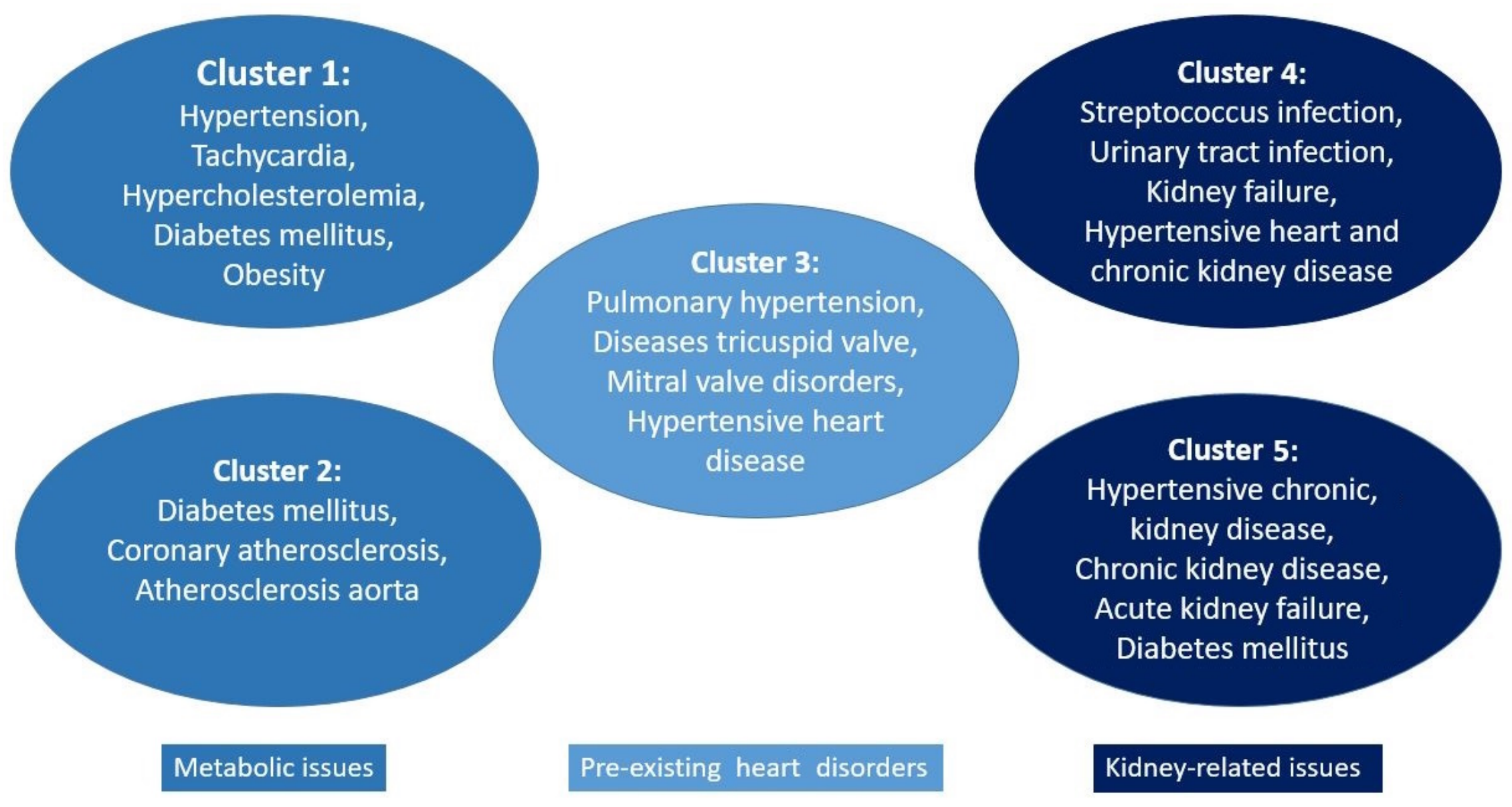
Key idea

- Estimate $\tilde{M}_2^{(N)}$ and $\tilde{M}_3^{(N)}$ (that are **biased**)
- Plug them into a tensor decomposition algorithm (we used SVTD [6]) to get $(\tilde{M}, \tilde{\omega})$ (biased as well)
- Remove the bias with EM.

Experiment: Heart Failure dataset

Patients affected by heart failure, having a diagnostic 428 (heart failure) in the ICD-9 code.




Cluster 1: Hypertension, Tachycardia, Hypercholesterolemia, Diabetes mellitus, Obesity

Cluster 2: Diabetes mellitus, Coronary atherosclerosis, Atherosclerosis aorta

Cluster 3: Pulmonary hypertension, Diseases tricuspid valve, Mitral valve disorders, Hypertensive heart disease

Cluster 4: Streptococcus infection, Urinary tract infection, Kidney failure, Hypertensive heart and chronic kidney disease

Cluster 5: Hypertensive chronic, kidney disease, Chronic kidney disease, Acute kidney failure, Diabetes mellitus

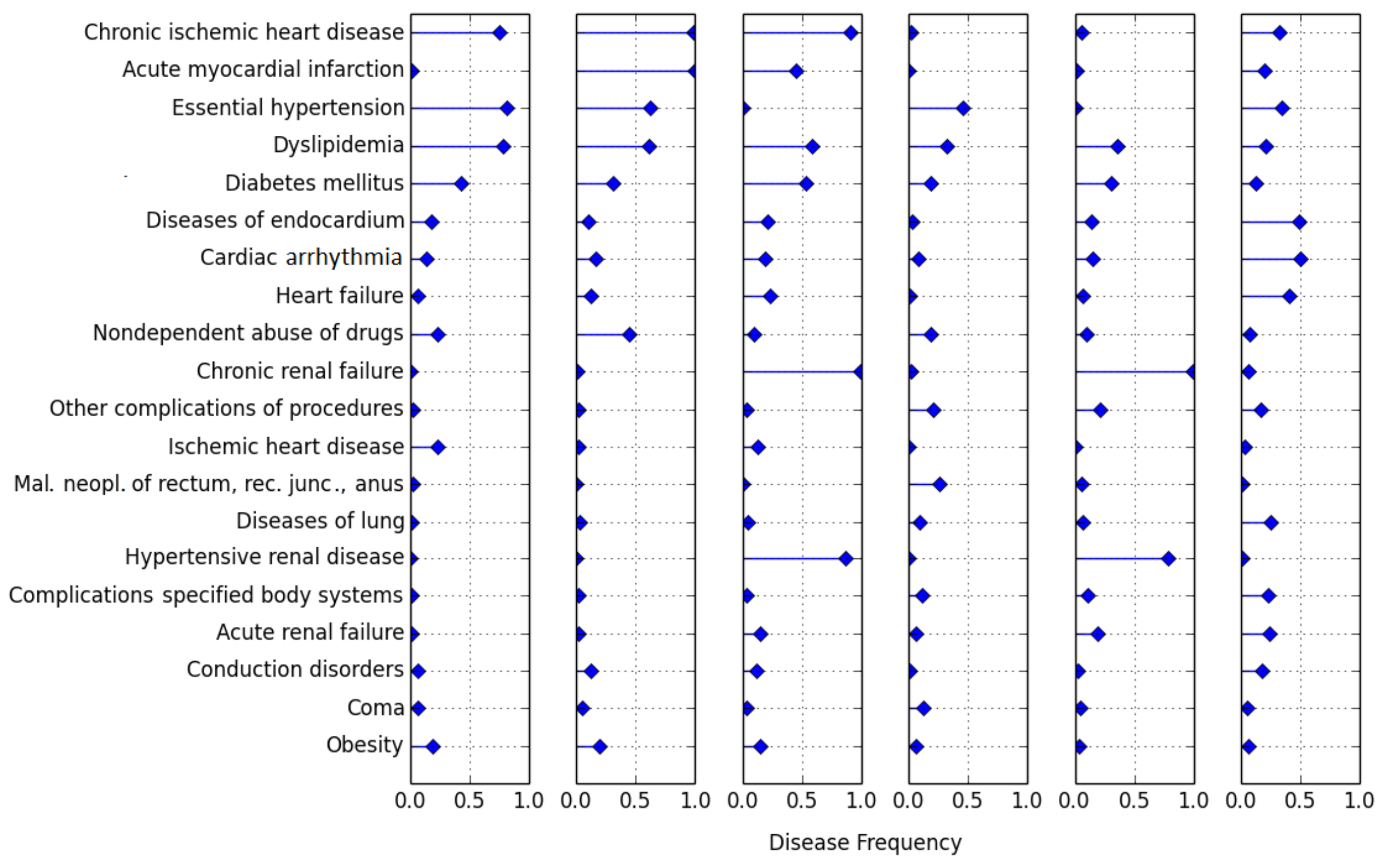
Metabolic issues

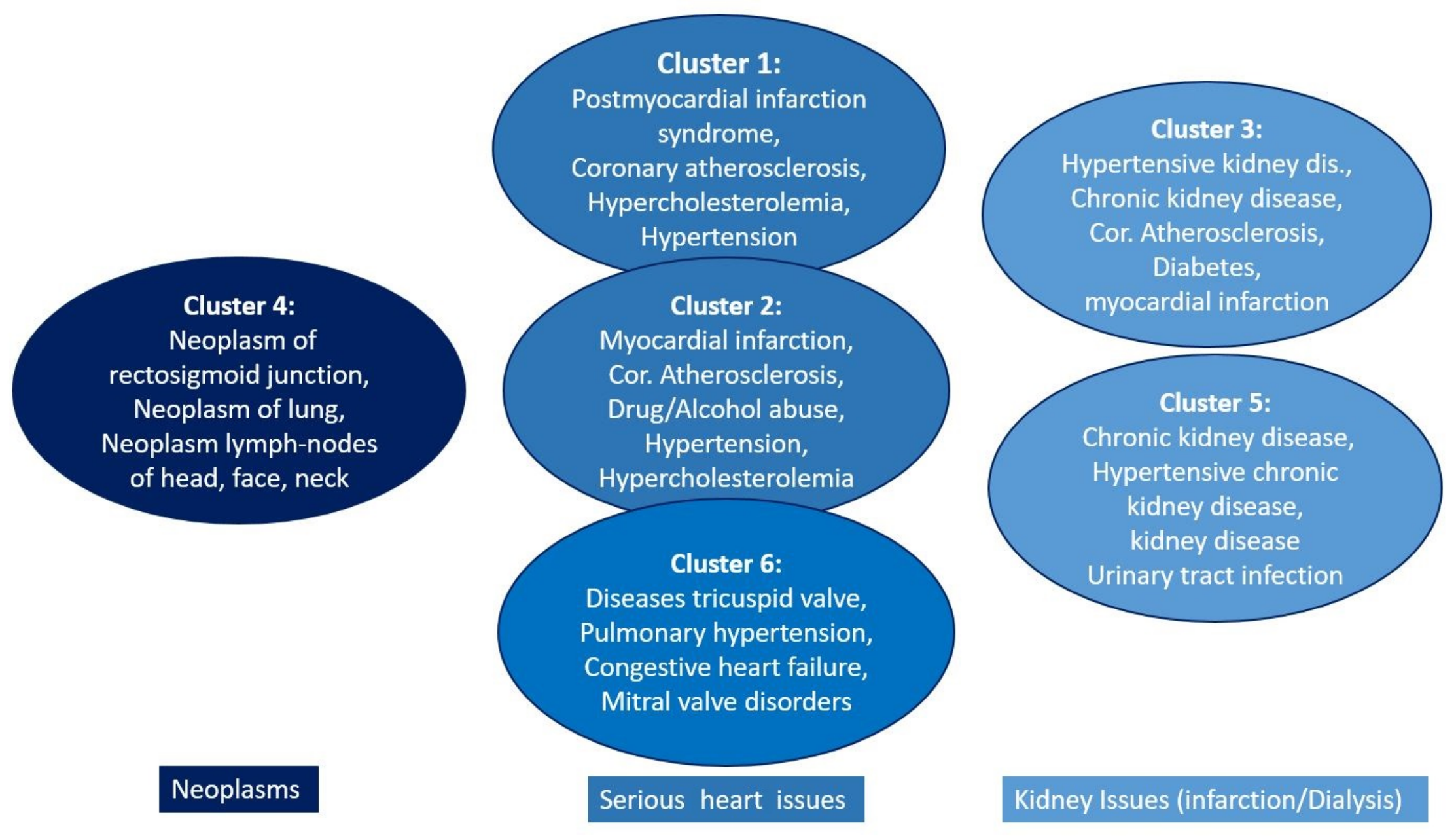
Pre-existing heart disorders

Kidney-related issues

Experiment: "Tertiary" Dataset

Patients with a serious disease, to be treated in reference hospitals in the area.




Cluster 1: Postmyocardial infarction syndrome, Coronary atherosclerosis, Hypercholesterolemia, Hypertension

Cluster 2: Myocardial infarction, Cor. Atherosclerosis, Drug/Alcohol abuse, Hypertension, Hypercholesterolemia

Cluster 3: Hypertensive kidney dis., Chronic kidney disease, Cor. Atherosclerosis, Diabetes, myocardial infarction

Cluster 4: Neoplasm of rectosigmoid junction, Neoplasm of lung, Neoplasm lymph-nodes of head, face, neck

Cluster 5: Chronic kidney disease, Hypertensive chronic kidney disease, kidney disease Urinary tract infection

Cluster 6: Diseases tricuspid valve, Pulmonary hypertension, Congestive heart failure, Mitral valve disorders

Neoplasms

Serious heart issues

Kidney Issues (infarction/Dialysis)

References

[1] K. Kshetri, (2011), Modelling patient states in intensive care patients.

[2] J. Ho et al, (2014), Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization.

[3] J. Ho et al. (2014), Limestone: High-throughput candidate phenotype generation via tensor factorization.

[4] Y. Wang et al (2015), Rubik: Knowledge guided tensor factorization and completion for health data analytics.

[5] A. Anandkumar et al, (2014), Tensor decompositions for learning latent variable models.

[6] M. Ruffini et al, (2017), A New Spectral Method for Latent Variable Models.