History-dependent Modeling of Patient Health Trajectories

Nils Haug[1,2], Stefan Thurner[1–4] and Peter Klimek[1,2]

[1]Section for Science of Complex Systems, Medical University of Vienna, Austria

[2]Complexity Science Hub Vienna, Austria

[3]Santa Fe Institute, USA

[4]IIASA, Austria

Complexity Science Hub Vienna
23 May 2019

Motivation

- Noncommunicable diseases such as dorsaglia, hypertension, respiratory diseases or diabetes affect a large fraction of the world's population and decrease the quality of life of people affected by them
- In many countries, healthcare expenditures amount to a considerable portion of the GDP, and are expected to rise in the future
- In order to make health care more efficient, we need to further our understanding on how the health state of a population proceeds based on its health history
- Data which is recorded routinely in the healthcare systems of many countries can be used for this purpose

Research question

Research question

How can we use this data to model the disease progression of a population depending on its entire observed health history?

Challenges

- There exists an astronomically high number of disease combinations, with many combinations occurring only in a single patient
- The time span covered by available data sets is limited
- Available data only provides a partial and distorted image of the real health state of a population

Past research

Past research has often focussed on binary relations between diseases, such as their *comorbidity*. Two diseases A and B are said to be comorbid if they often co-occur in the same patients, that is, if

$$\mathbb{P}(A \cap B) > \mathbb{P}(A) \cdot \mathbb{P}(B).$$

C A Hidalgo, N Blumm, A-L Barabási and N A Christakis
A Dynamic Network Approach for the Study of Human Phenotypes.
PLoS Comput. Biol., 5(4):e1000353, 2009.

A Chmiel, P Klimek and S Thurner
Spreading of diseases through comorbidity networks across life and gender.
New Journal of Physics, 16:115013, 2014.

The data

We analyse data on hospital stays in Austria from 1997 until 2014. The cohort of our study consists of the $M = 5,112,811$ individuals without a hospital diagnose with ICD-10 code from A00–N99 from 1997 until 2002.

Example: Data recorded for each patient

- Sex (male/female)
- Date of birth (5 year resolution)
- Data on hospital stays, e.g.:

Admission	Release	Main	Side
04.02.04	09.02.04	R35	E11, E28
19.11.06	20.12.06	U32	E11, E24, E28, I15, N14
04.07.11	14.11.11	I33	None

The model

At a given point in time t, the health state of the patient cohort is described by a binary $(M \times N)$ matrix $\mathbf{X}^{(t)}$, such that

$$(\mathbf{X}^{(t)})_{pj} = \begin{cases} 1 & \text{if patient } p \text{ has disease } j, \\ 0 & \text{otherwise.} \end{cases}$$

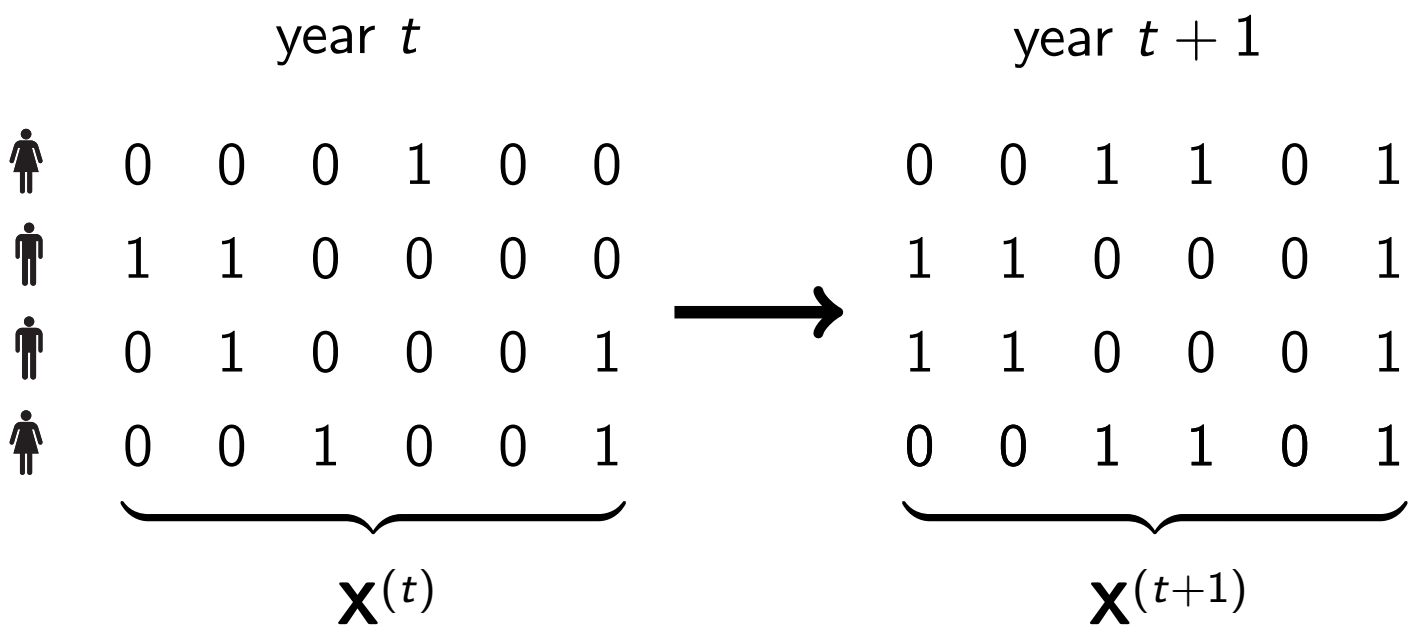
	Neoplasms	Retinal disorders	Diabetes	Obesity	Depression	Hypertension
	0	0	1	1	0	1

In our case, $M \simeq 5 \cdot 10^6$ and $N = 131$.

The model

We sample the health state of each patient of the cohort at the end of each year of the observation period from 2002 to 2014.

As time proceeds, patients acquire new diseases.



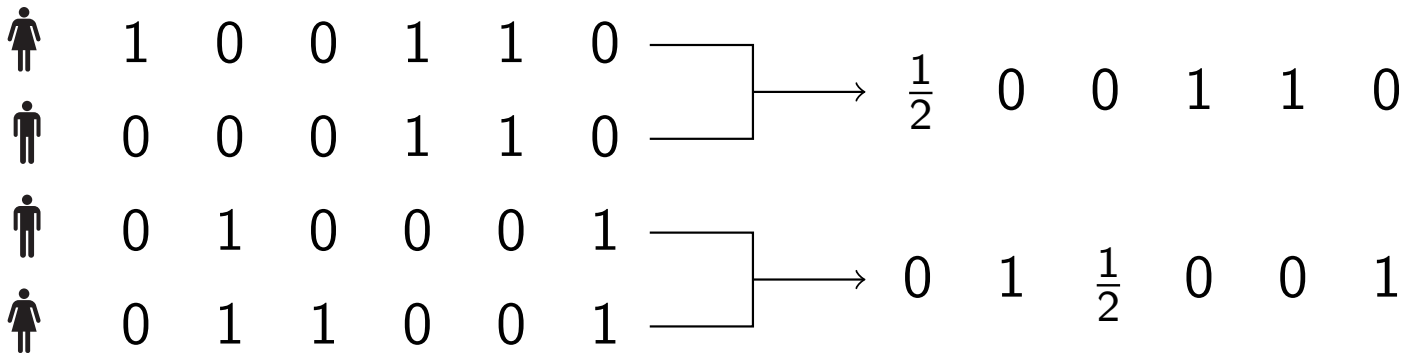
Patient clustering

We define the $((L \cdot M) \times N)$ matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{X}^{(0)} \\ \vdots \\ \mathbf{X}^{(L)} \end{pmatrix},$$

where $L = 13$ is the length of the observation period in years.

It is possible to partition the rows of $\mathbf{M}$ into K clusters.



Patient clustering

For $1 \leq k \leq K$, the centroid $\mathbf{C}^{(k)} = (C_1^{(k)}, \ldots, C_N^{(k)})$ of a cluster k is a vector with entries defined as

$$C_j^{(k)} = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} (\mathbf{M})_{ij},$$

where $|\mathcal{S}_k|$ is the set of rows contained in cluster k.

The homogeneity of a cluster k is measured in terms of the **inertia**

$$\mathcal{I}_k = \sum_{j=1}^{N} C_j^{(k)} (1 - C_j^{(k)}).$$

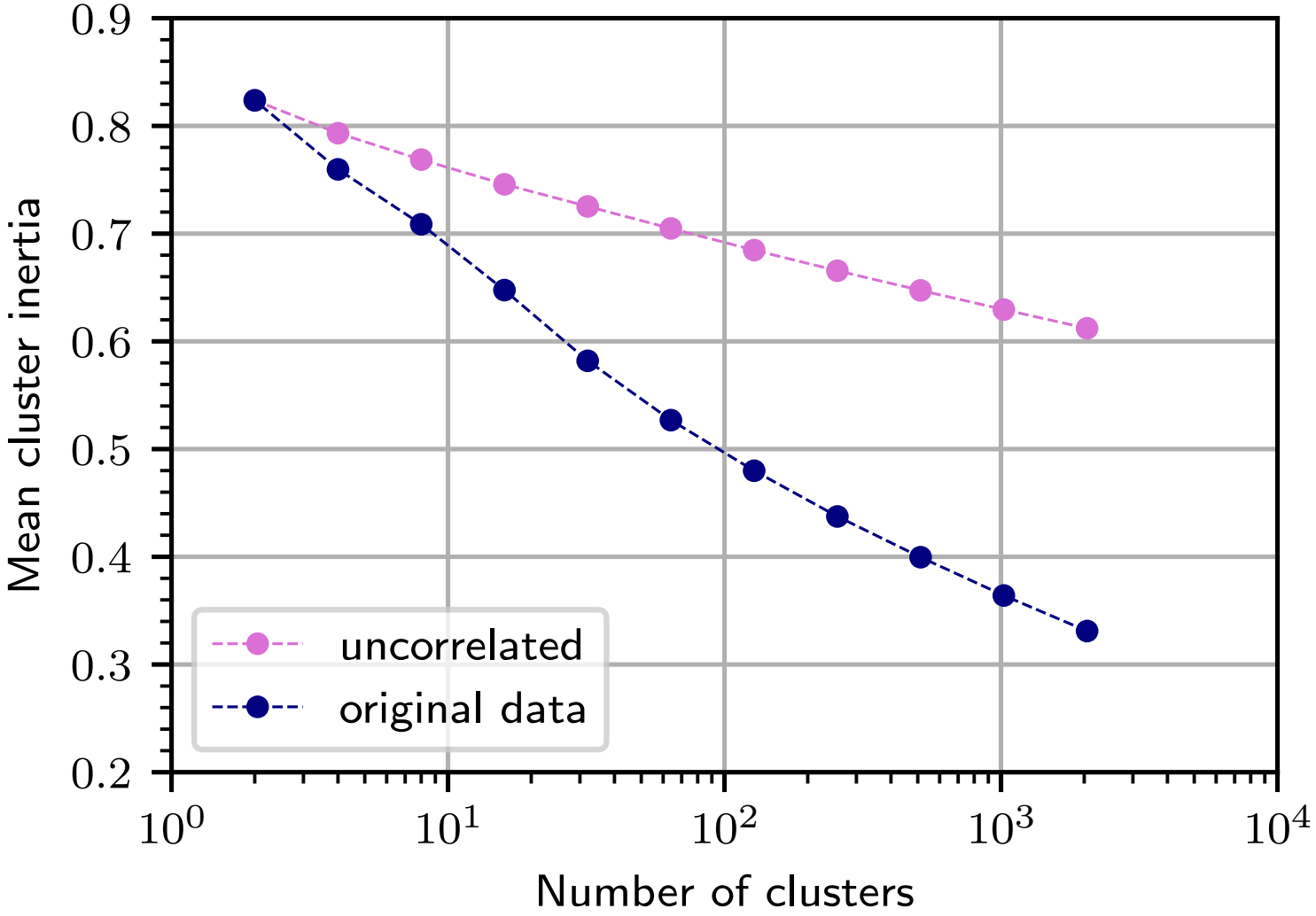
Patient clustering

We run a divisive clustering algorithm (DIVCLUS-T) on the patient cohort, such that at each splitting step, the decrease in the mean cluster inertia

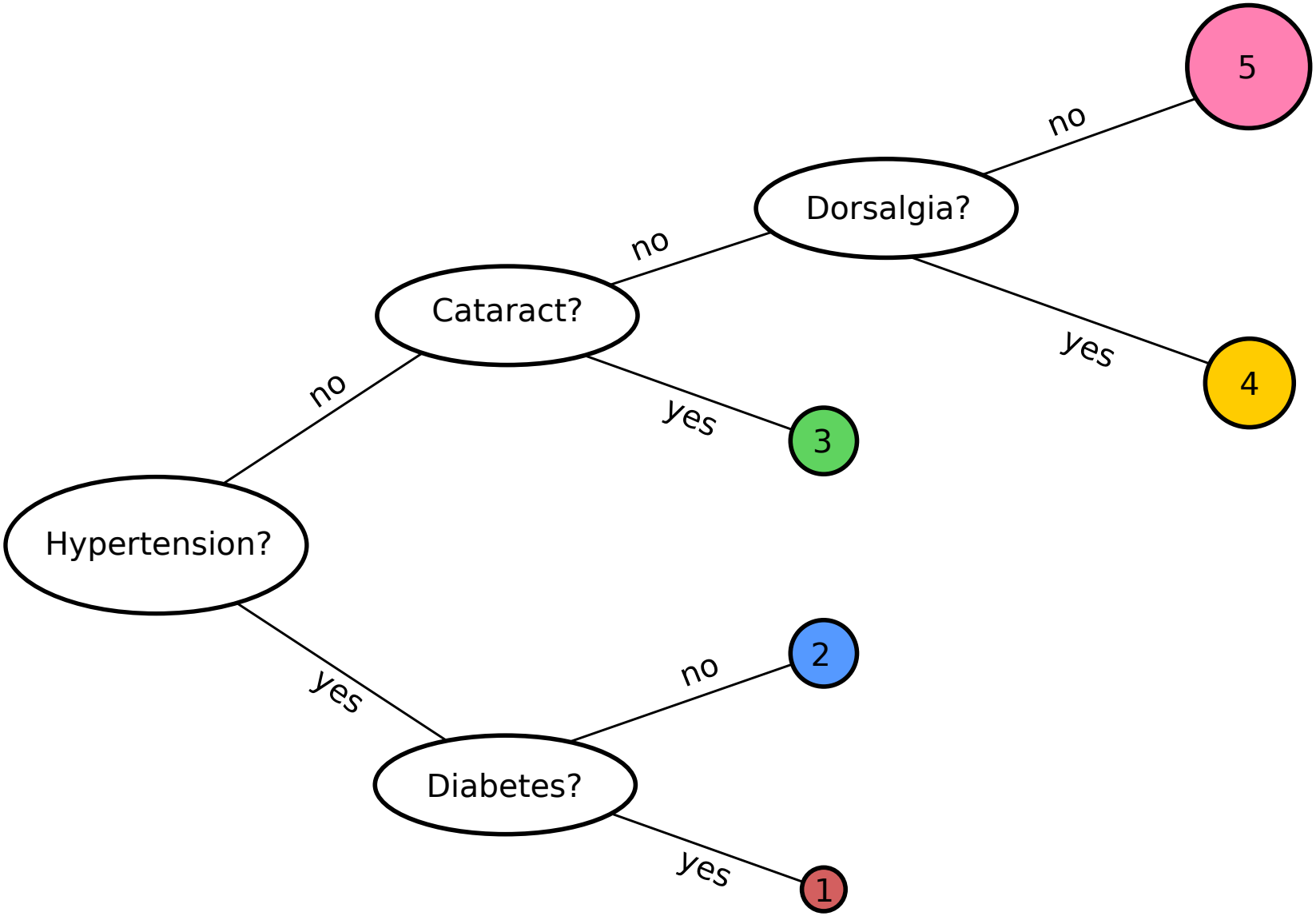
$$f = \frac{1}{\sum_k |\mathcal{S}_k|} \sum_{k=1}^{K} |\mathcal{S}_k| I_k$$

is maximal, where $|\mathcal{S}_k|$ is the number of observations belonging to cluster k and K is the number of clusters.

Selection of the number of clusters



Example clustering tree



Patient trajectories

Patients can change clusters by acquiring new diseases. The sequence of clusters a patient belongs to throughout the years describes his or her health trajectory.

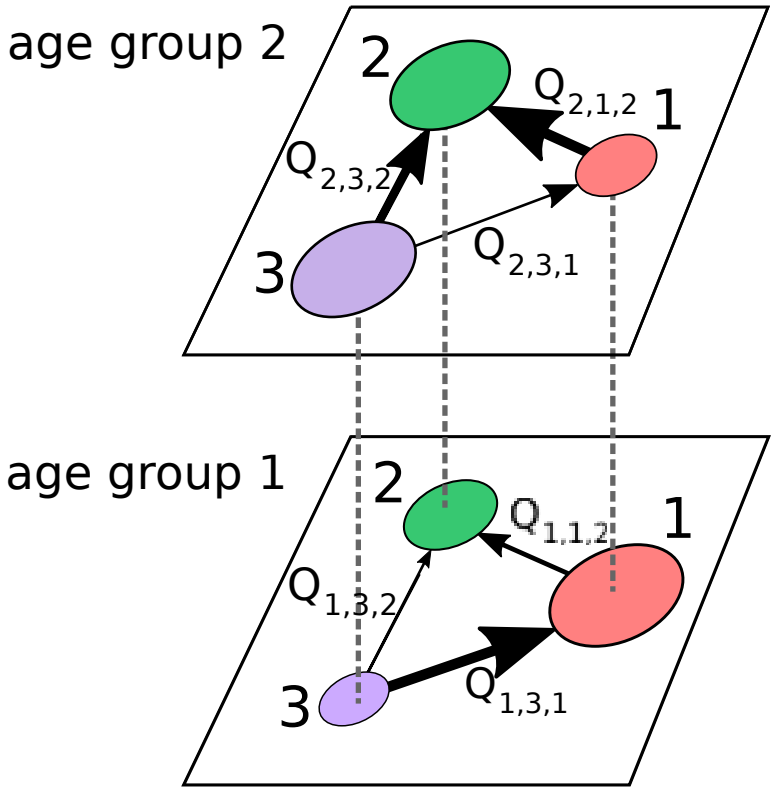
For each sex s and 10-year age group g, we calculate the $(K \times K)$ matrix $\mathbf{Q}^{(s,g)}$, where for $1 \leq j, k \leq K$,

$$(\mathbf{Q}^{(s,g)})_{j,k} = \frac{n_{j \to k}^{(g)}}{n_j^{(g)}},$$

where $n_j^{(s,g)}$ is the number of patients of sex s and age group g in cluster j and $n_{j \to k}^{(s,g)}$ is the number of patients of sex s and age group g stepping from cluster j to cluster k.

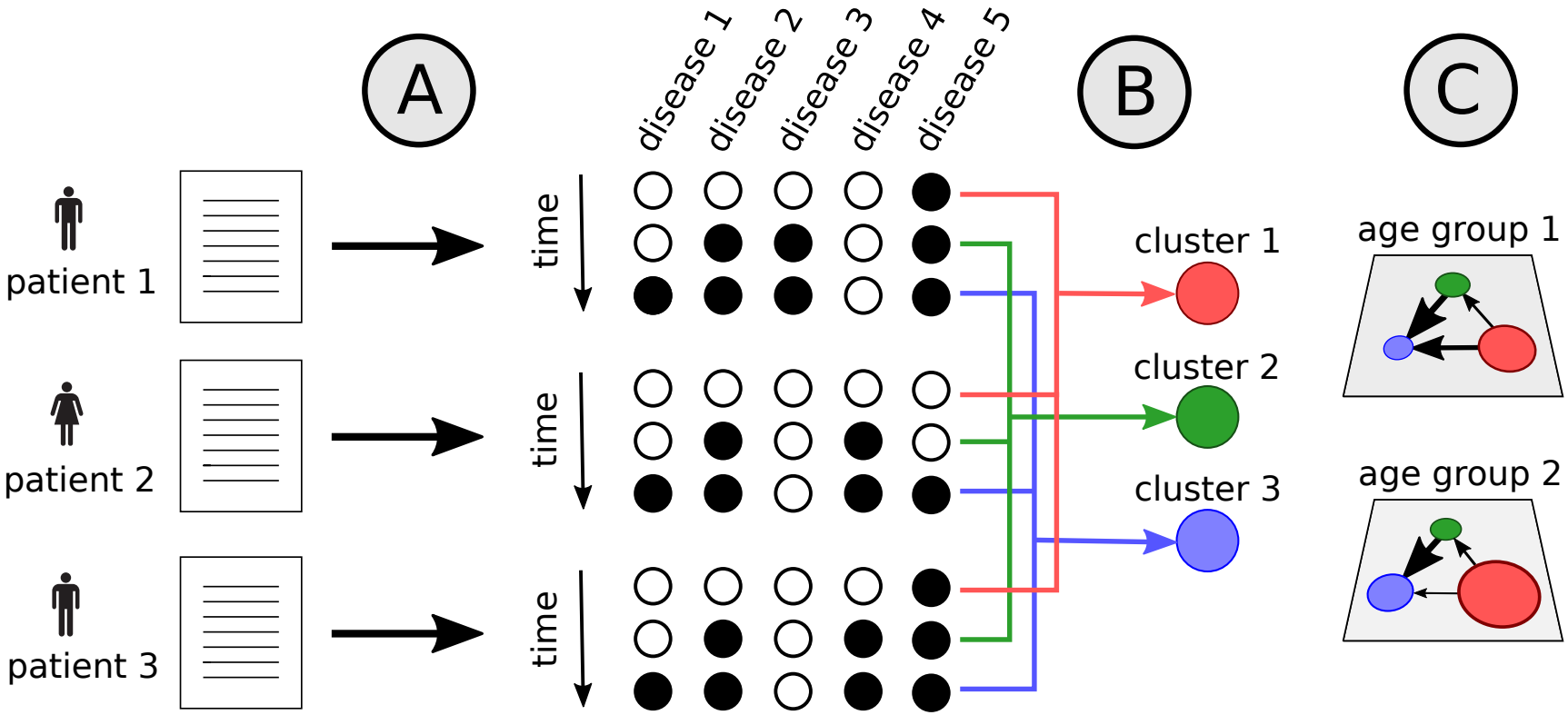
Modeling patient trajectories

We construct a multiplex network $\mathcal{M}$ of health state clusters, where layers correspond to sex and age groups.



We model the disease progression of patients as random walks on $\mathcal{M}$.

Our approach in a nutshell



Cluster conditions

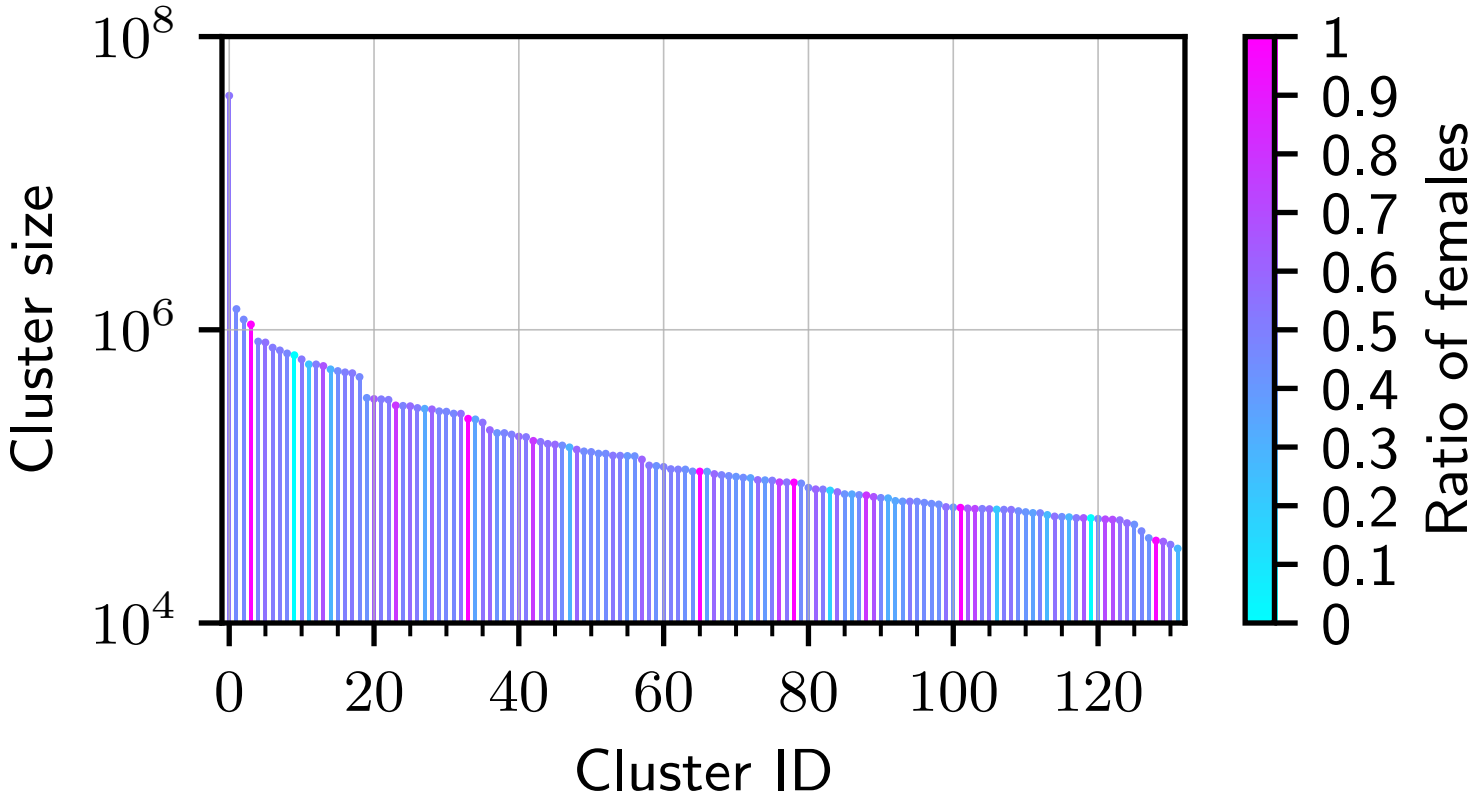
Each cluster is defined by a set of diseases which each patient in that cluster has been diagnosed with (inclusion criteria) and a set of diseases each patient in that cluster has *not* been diagnosed with (exclusion criteria).

Example

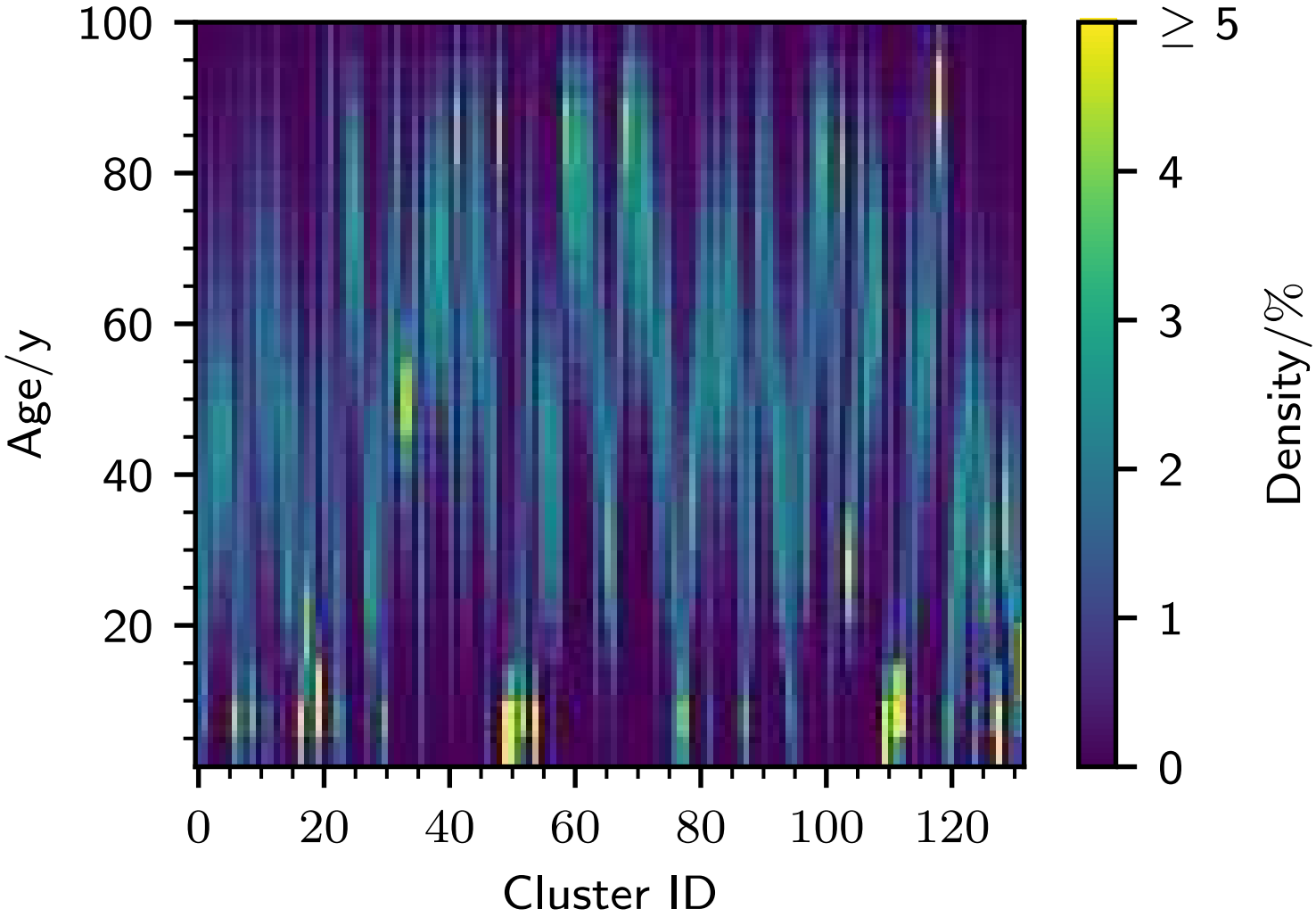
Diabetes mellitus (E10-E14)	✗
Metabolic disorders (E70-E90)	✓
Hypertensive diseases (I10-I15)	✓
Ischaemic heart diseases (I20-I25)	✓
Other forms of heart disease (I30-I52)	✗

Table: Inclusion and exclusion criteria for cluster 34.

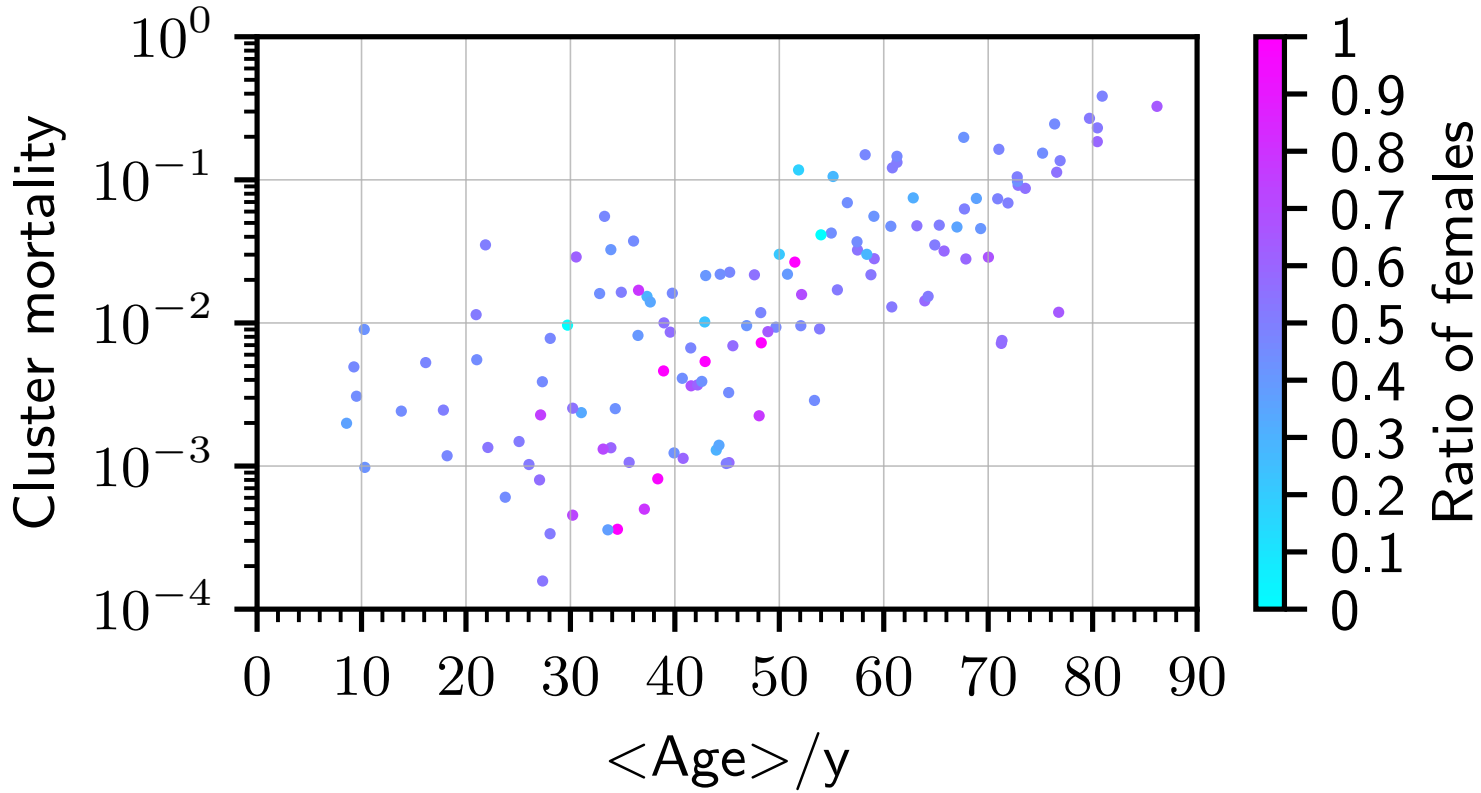
Distribution of cluster sizes



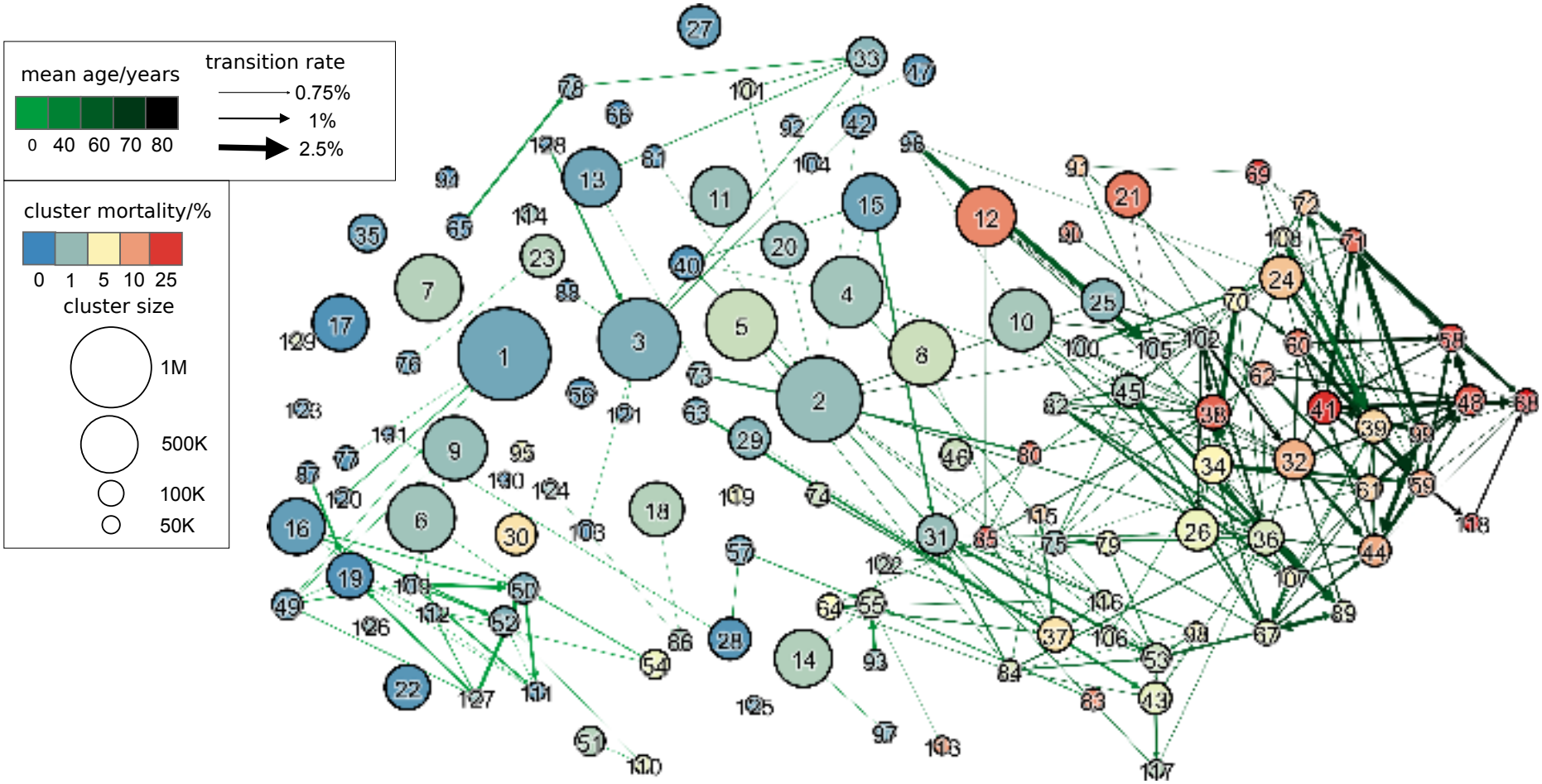
Age distribution in the different clusters



In-hospital mortality in the different clusters



Collapsed multiplex network of clusters



The highest mortality cluster (the sink state)

The in-hospital mortality of patients in cluster 68 is 38%.

Hypertensive diseases (I10-I15)	✓
Other forms of heart disease (I30-I52)	✓
Diseases of arteries, arterioles and capillaries (I70-I79)	✓
Renal failure (N17-N19)	✓

Table: Inclusion criteria for cluster 68.

The most frequent trajectory of length 3

Cluster 0	
all	✗

$\xrightarrow{129{,}961}$

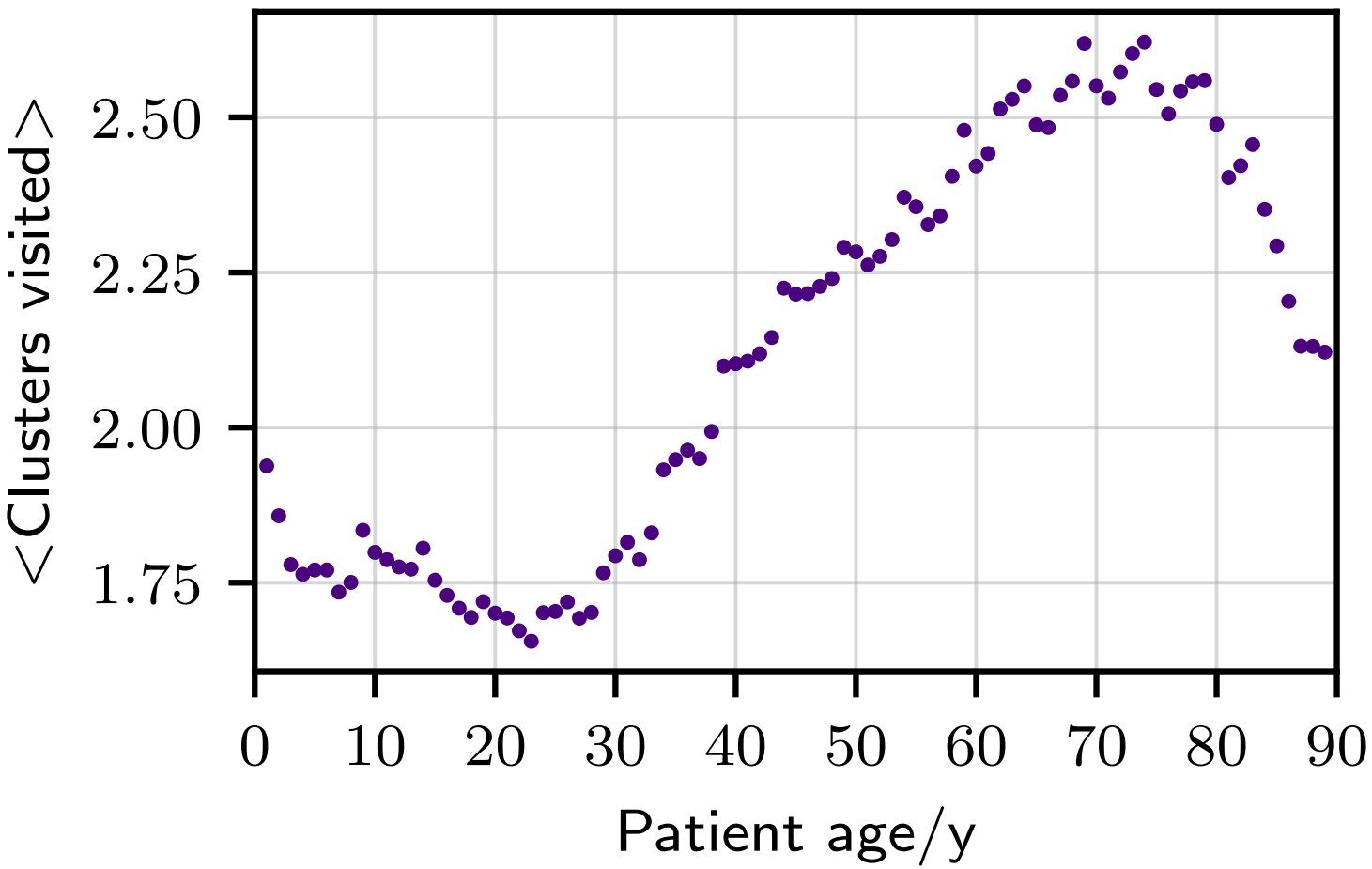
Cluster 2	
E70-E90	✗
I10-I15	✗
I80-I89	✗
J30-J39	✗
M00-M25	✓
M40-M54	✗
M60-M79	✗
M80-M94	✗

$\xrightarrow{5{,}983}$

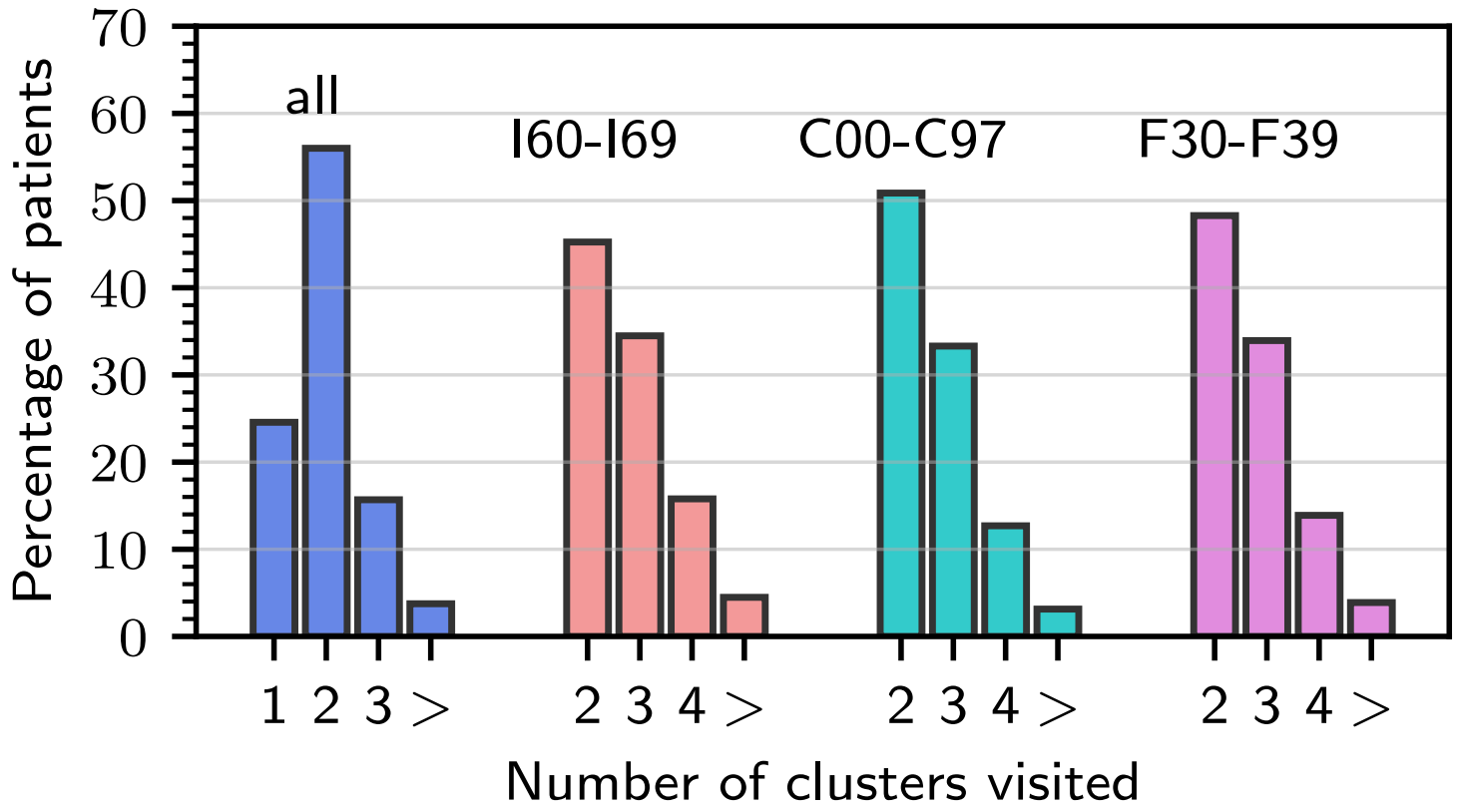
Cluster 31	
I10-I15	✗
J30-J39	✗
M00-M25	✓
M60-M79	✓
M80-M94	✗

E70-E90	Metabolic disorders
I10-I15	Hypertensive diseases
I80-I89	Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
J30-J39	Other diseases of upper respiratory tract
M00-M25	Arthropathies
M40-M54	Dorsopathies
M60-M79	Soft tissue disorders
M80-M94	Osteopathies and chrondropathies

Mean age-dependent trajectory length



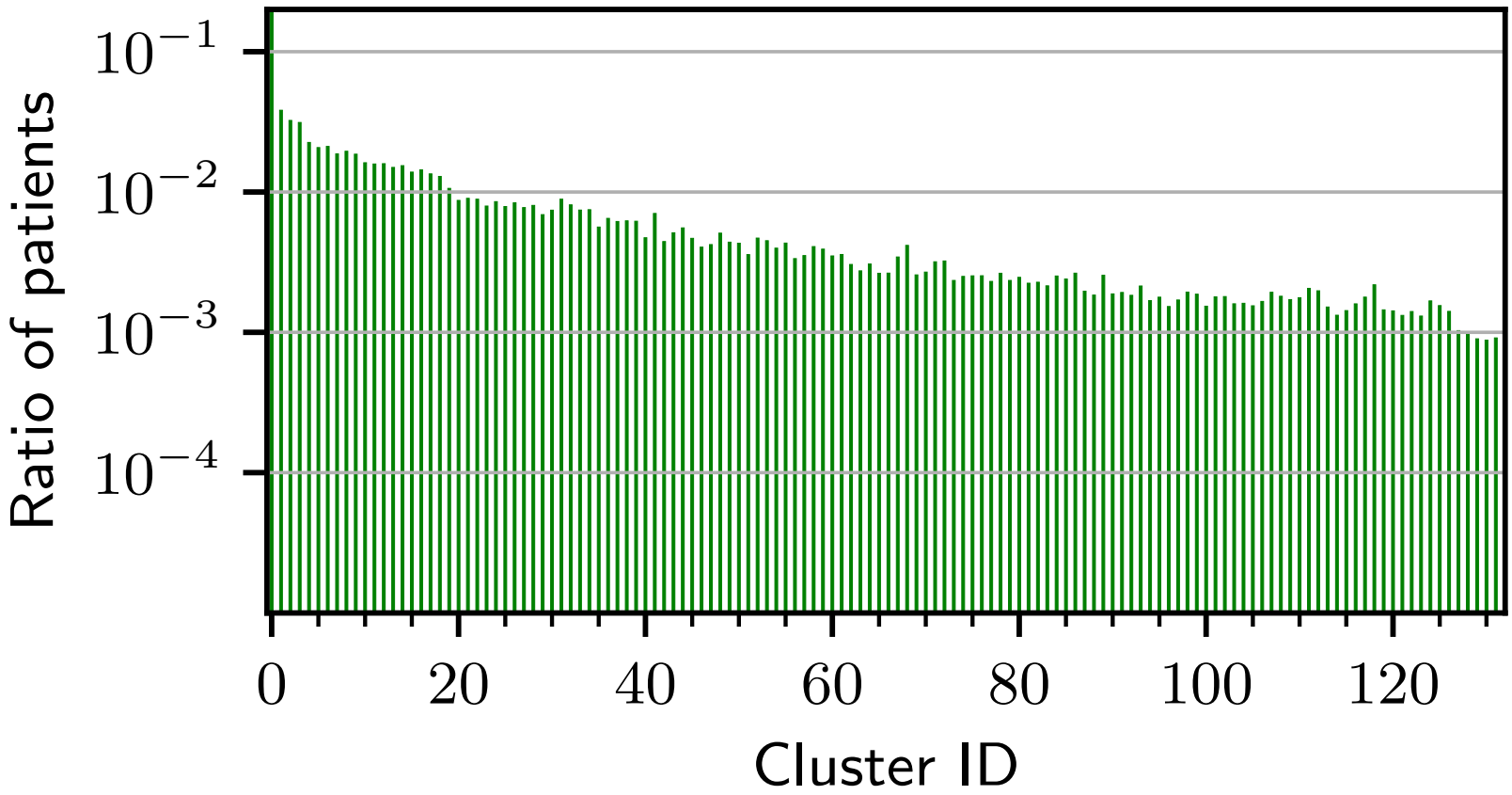
Distribution of trajectory lengths



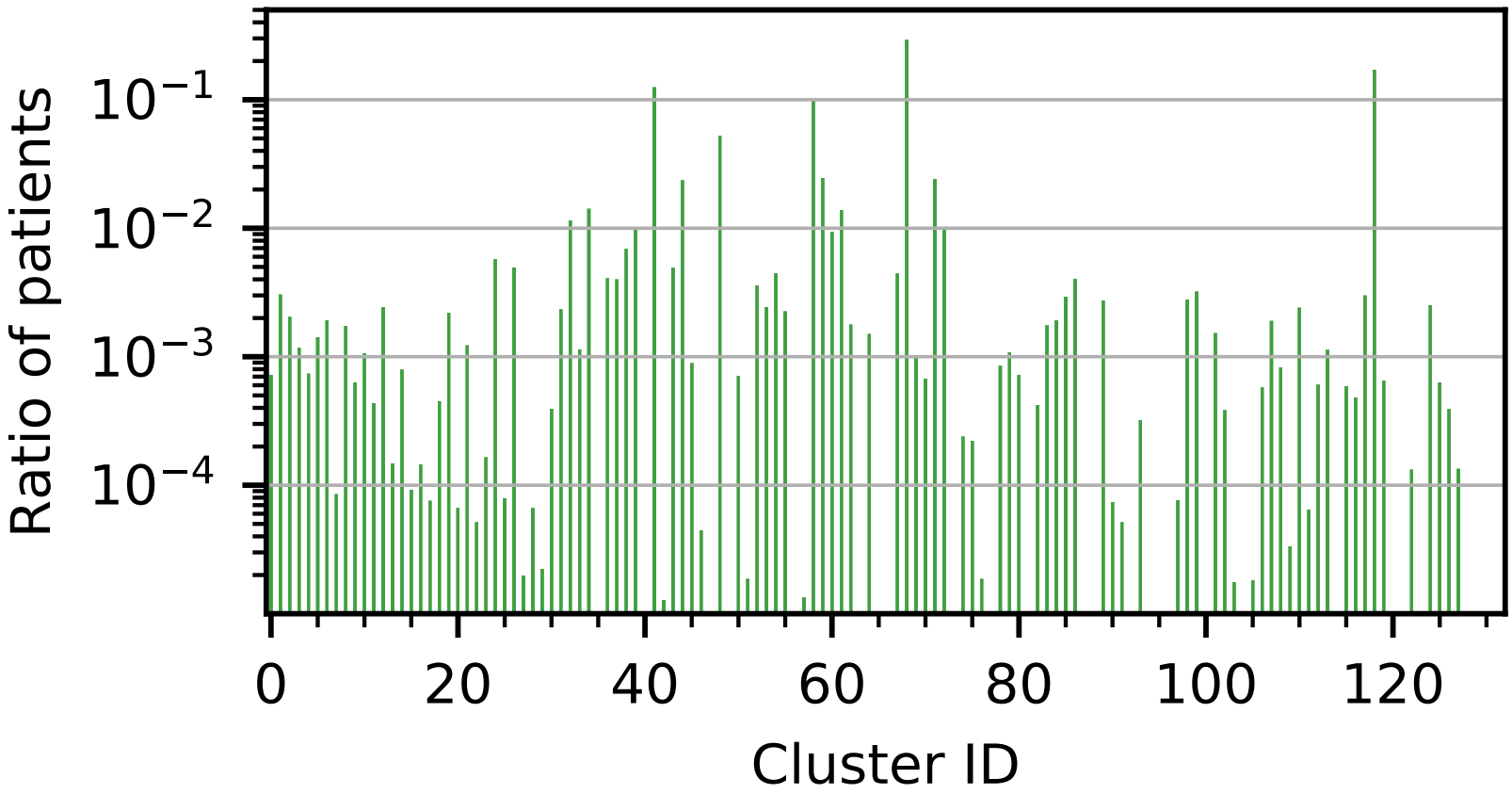
Example trajectories involving common diseases

- A total of 210,589 patients receive a diagnosis of mood or affective disorders. The most frequent trajectory of length 3 is followed by 1,936 patients. These patients first acquire a diagnosis of mental and behavioral disorders due to psychoactive substance abuse and subsequently one of mood disorders.
- A total of 312,787 patients get diagnosed with malignant neoplasms, among which 3,093 patients receive a diagnosis of hypertensive diseases after the cancer diagnosis.
- A total of 199,681 patients get a diagnosis of cerebrovascular diseases. The most frequent trajectory of length 3 is followed by 1,447 patients. These patients get diagnosed with hypertension and heart diseases beforehand.

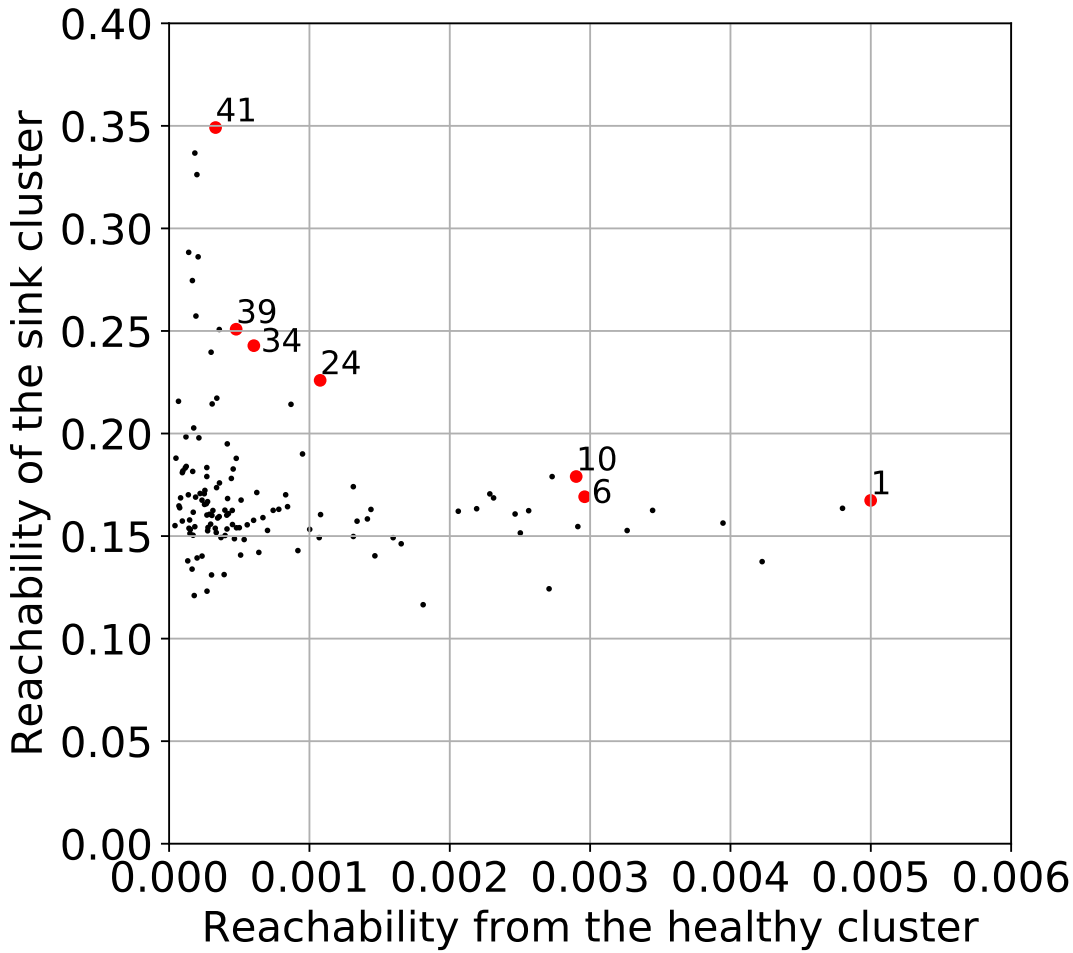
Cluster populations after 13 years



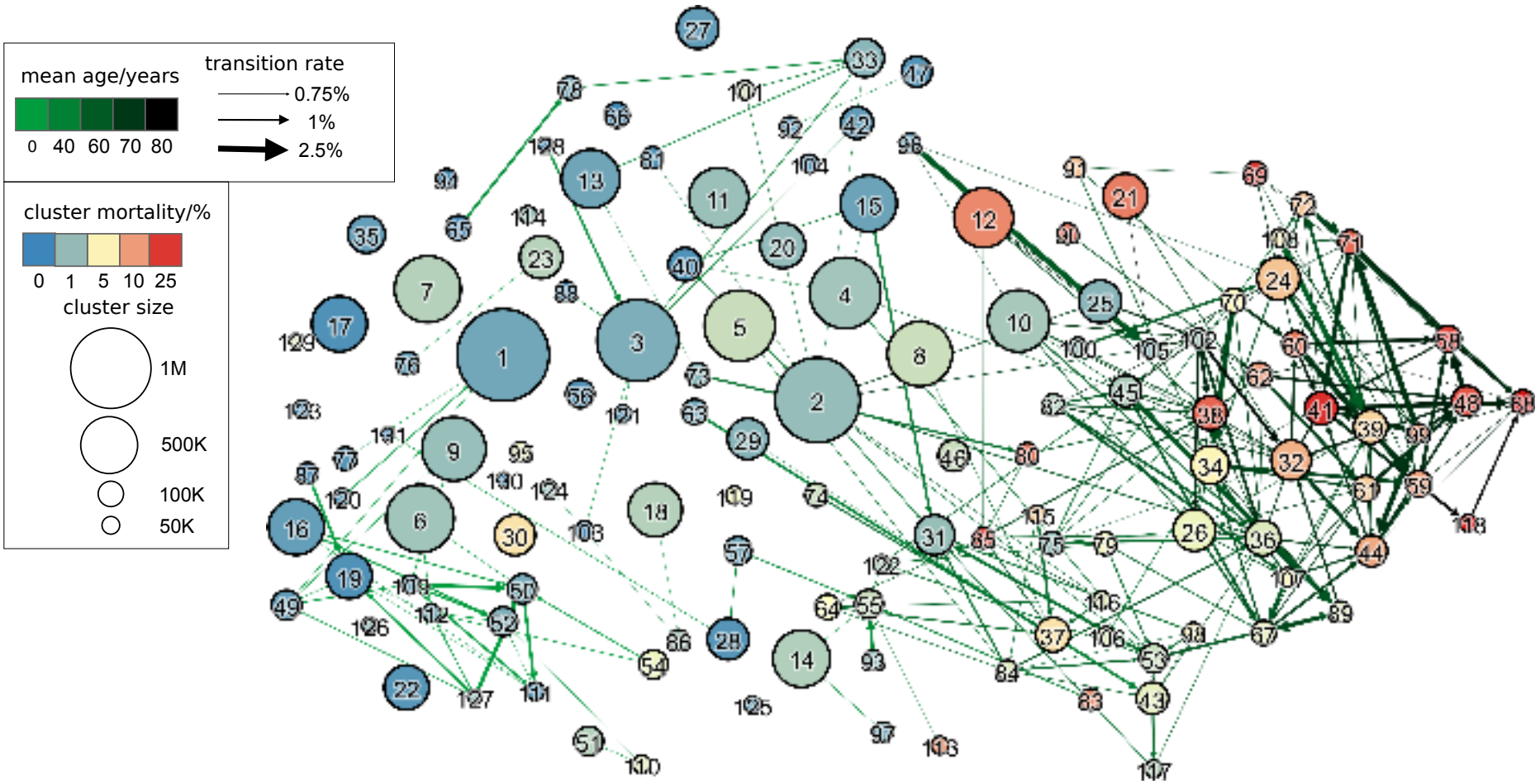
Projected cluster populations after 100 years



Gatekeepers to the sink state



Gatekeepers to the sink state



Gatekeepers to the sink state

Diabetes mellitus (E10-E14)	✓
Obesity and other hyperalimentation (E65-E68)	✗
Metabolic disorders (E70-E90)	✗
Hypertensive diseases (I10-I15)	✓
Other forms of heart disease (I30-I52)	✗

Table: Inclusion criteria and exclusion criteria for cluster 24.

Comparison with benchmarks

The vectors $\hat{\mathbf{p}} = (\hat{p}^{(1)}, \ldots, \hat{p}^{(N)})$ and $\mathbf{p} = (p^{(1)}, \ldots, p^{(N)})$ give the marginal disease probabilities at the end of the observation period as predicted by the model and as observed from the data, respectively.

	$\|\hat{\mathbf{p}}\|_1 / \|\mathbf{p}\|_1$	$\|\Delta\mathbf{p}\|_1$	$\|\Delta\mathbf{p}\|_2$	f
DIVCLUS-T	98%	0.14	0.021	0.48
Benchmark 1	88%	0.30	0.039	0.55
Benchmark 2	94%	0.17	0.029	0.48

Summary and Limitations

Summary

- We presented a novel method to describe the disease progression of a population depending on its disease history. The approach is based on a coarse grained description of the health states of patients in terms of a small number of easily interpretable clusters
- From a statistical analysis of the transition rates of patients between the different clusters, we created a model which can be used to generate synthetic patient health trajectories

Limitations

- The time span covered by the data set is limited
- The data only provides a partial and distorted image of the health state of the population
- The waiting times for the transitions are assumed to be exponentially distributed, which is not always the case

The End.