Tracking clusters of patients over time enables extracting information from medico-administrative databases

JDB 2022 – 17.11.2022

Judith LAMBERT
Anne-Louise LEUTENEGGER
Anne-Sophie JANNOT
Anaïs BAUDOT

Marseille Medical Genetics

NEURODIDEROT
Unité Mixte de Recherche UMR 1141
Inserm-Université Paris Diderot

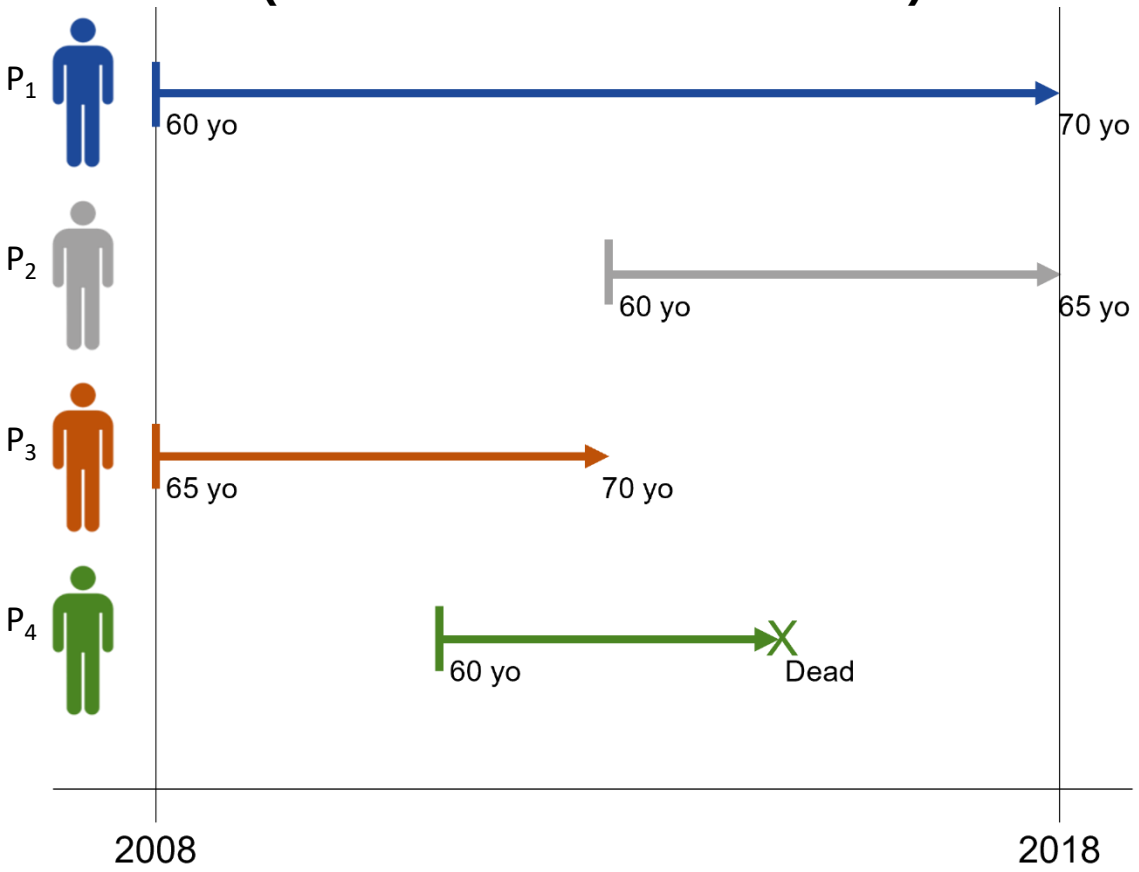
CRC
CENTRE DE RECHERCHE DES CORDELIERS

HeKA

Université Paris Cité

Context

Medico-administrative databases (Healthcare reimbursements)

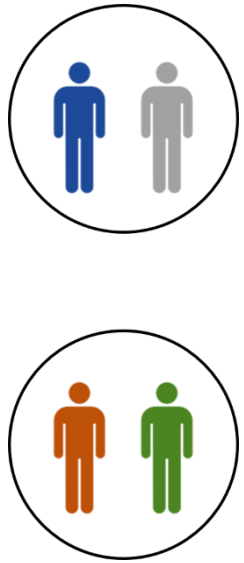


Patients are characterized by different types of longitudinal variables which are measured over different follow-up periods (→), generating truncated data.

Clusters of patients (subgroups)

Goal: better understand disease heterogeneity [1]

Clustering of patients



[1] Sula Windgassen et al. (2018)

Classical approaches to cluster patients with longitudinal data

Longitudinal approaches

	Raw-data-based approaches[1]	Feature-based approaches[1]	Model-based approaches[1]
Principe	Approaches directly applied on raw longitudinal data	Features extracted from raw longitudinal data + non-longitudinal clustering approaches	Model estimation
Drawback	Patients with truncated data must be removed from analysis or their data must be imputed		

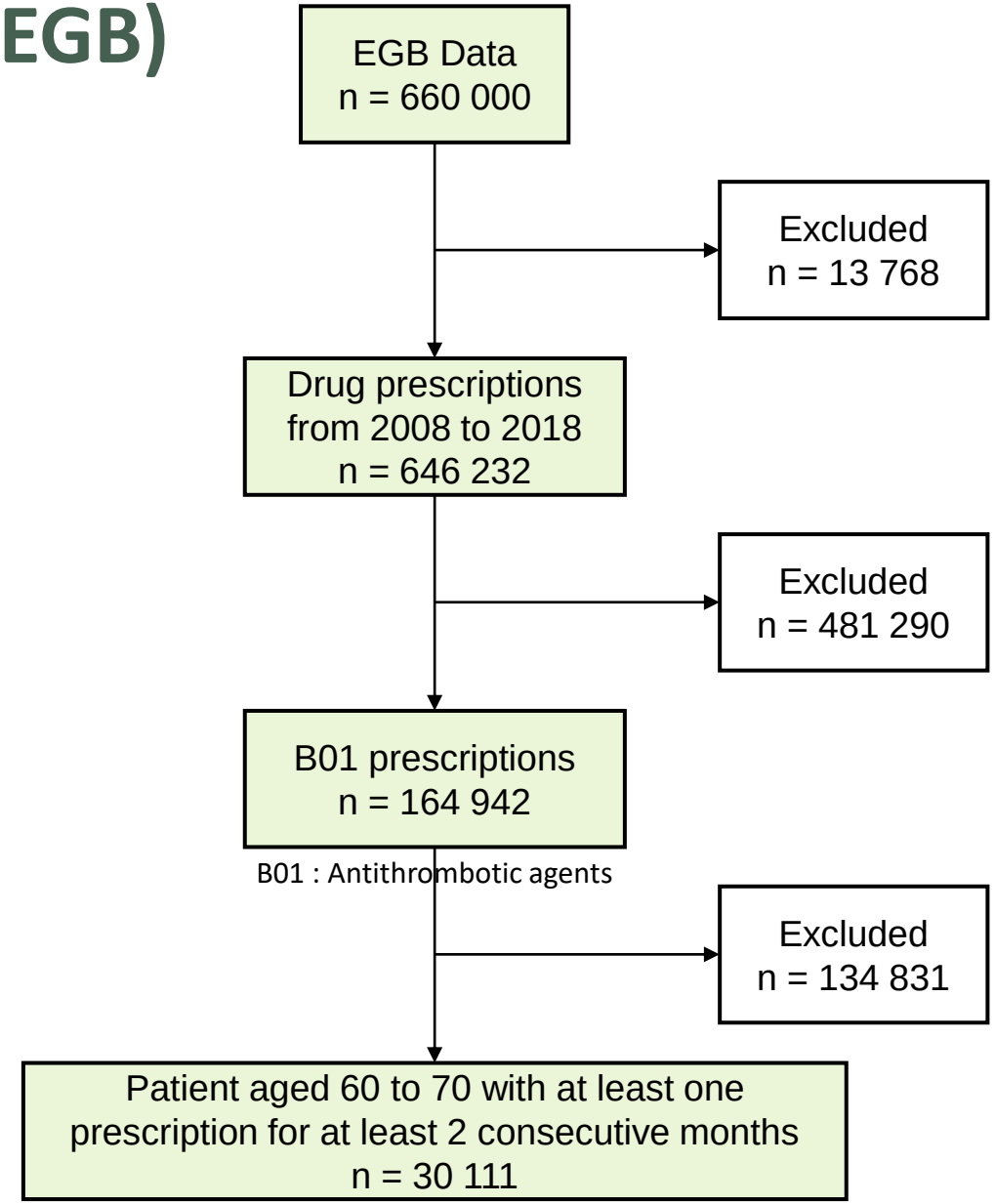
Adapted approaches required

↓

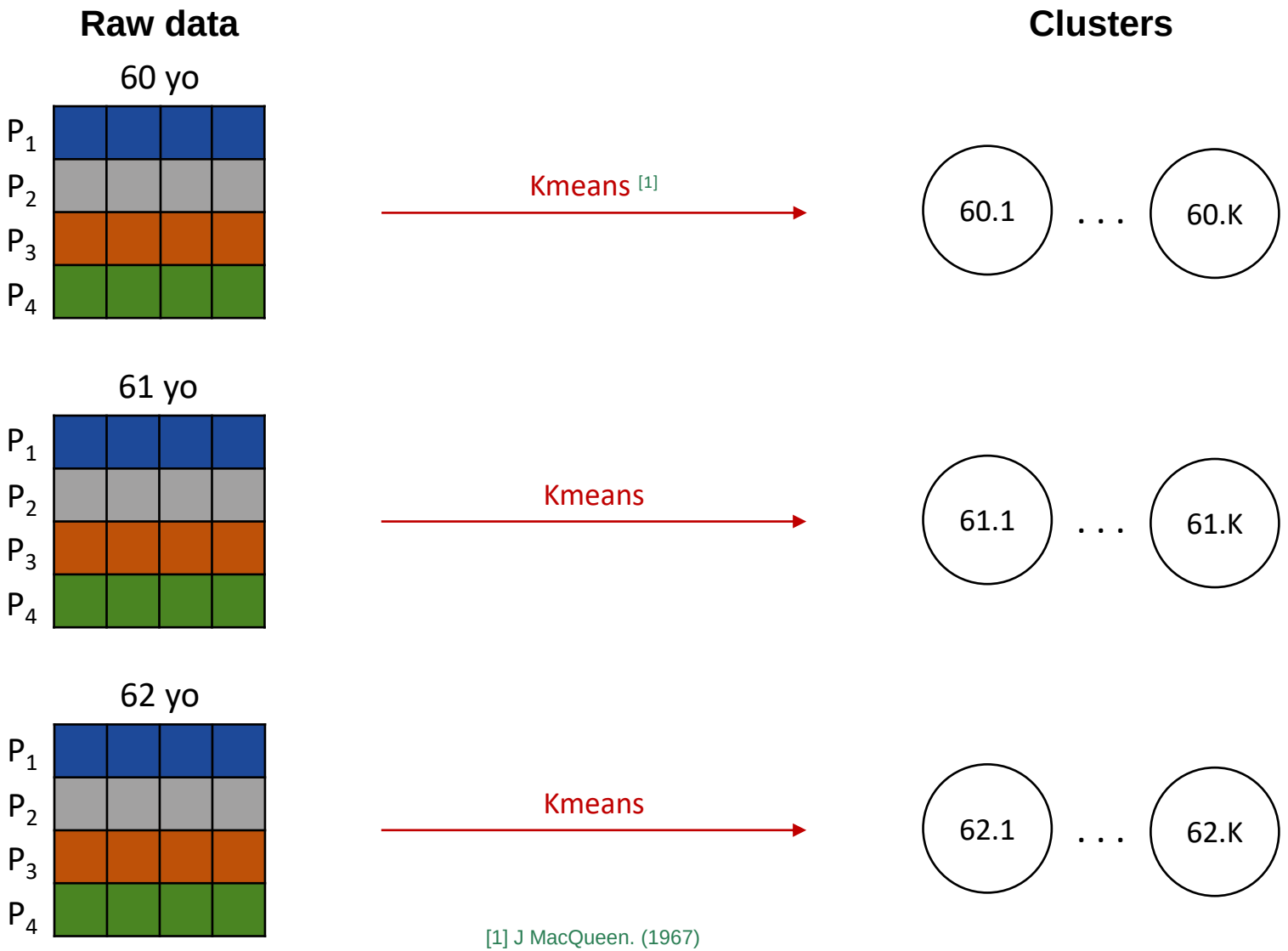
Cluster-tracking approaches

[1] Liao, T. Warren. "Clustering of time series data—a survey." (2005)

Use case : Echantillon Généraliste des Bénéficiaires (EGB)



Cluster-tracking: raw-data-based approach



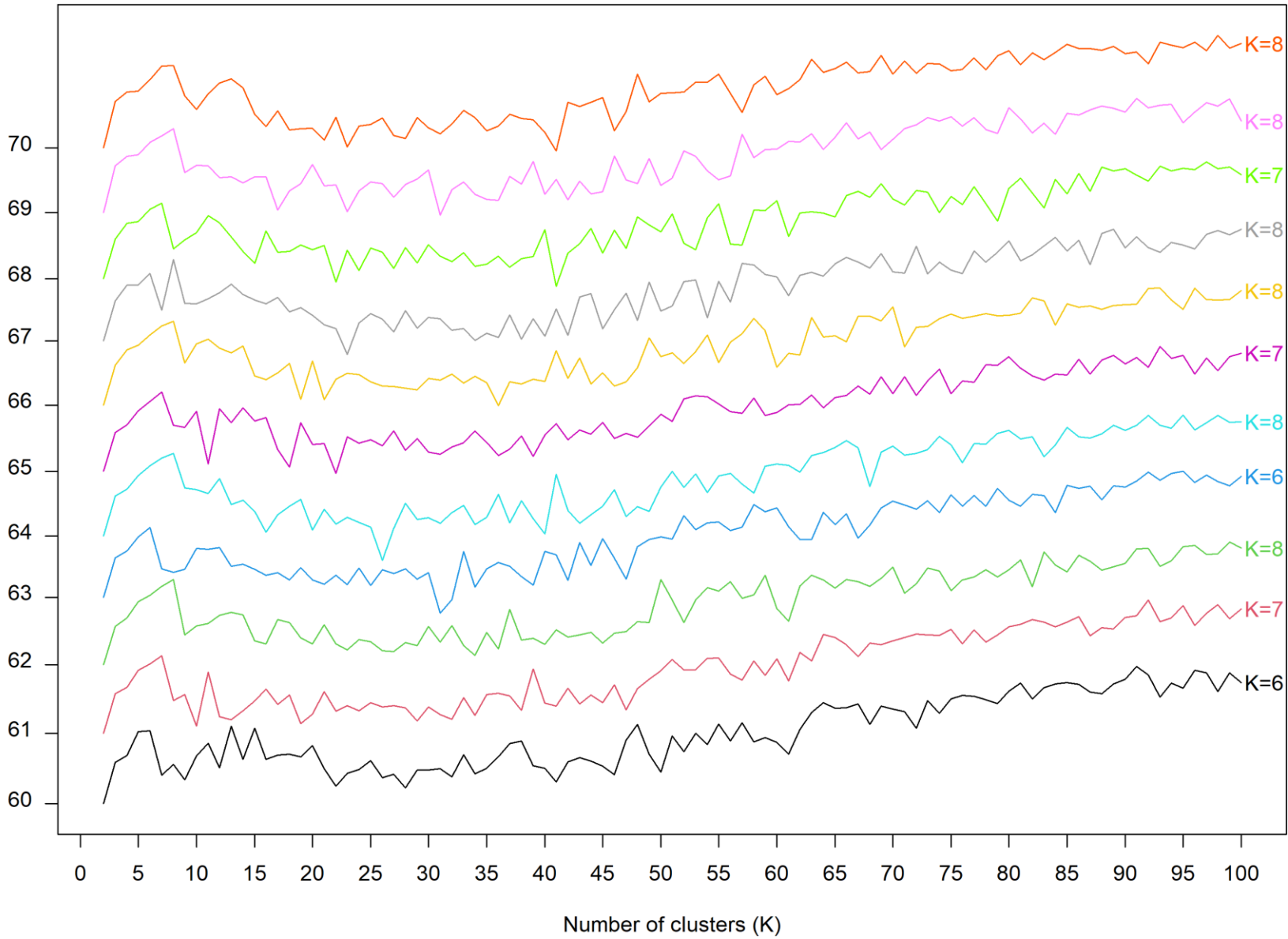
[1] J MacQueen. (1967)

Cluster-tracking: raw-data-based approach

Silhouette score: assess the clustering quality

$$S^i = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{p\in P^k} s_p^i$$

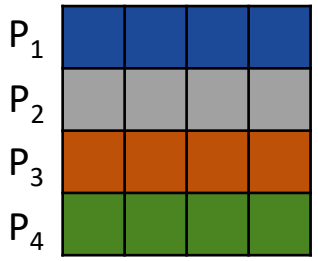
n_k: number of patients in the cluster k
K: number of clusters
i: age
s_p^i: silhouette score of patient p



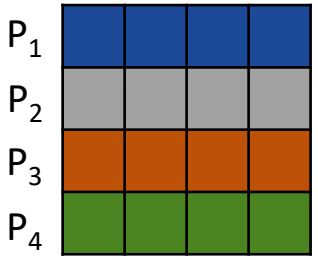
Cluster-tracking: network-based approach

Raw data

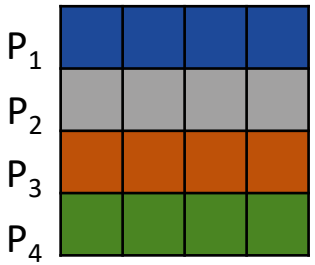
60 yo



61 yo



62 yo



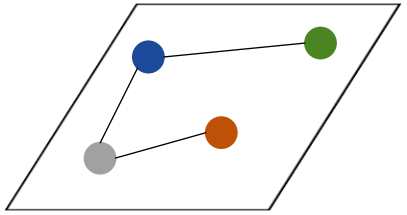
Cosine similarity [1]

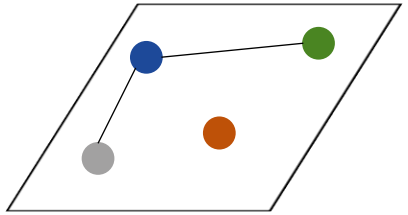
$$\cos\theta = \frac{A \cdot B}{||A|| \cdot ||B||} \geq 0.8$$

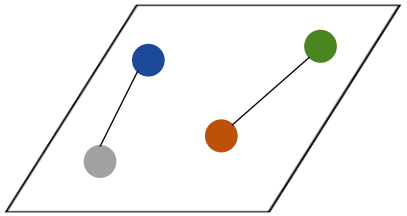
Cosine similarity

Cosine similarity

Patient networks







MCL [2]

MCL

MCL

Clusters

60.1 . . . 60.K

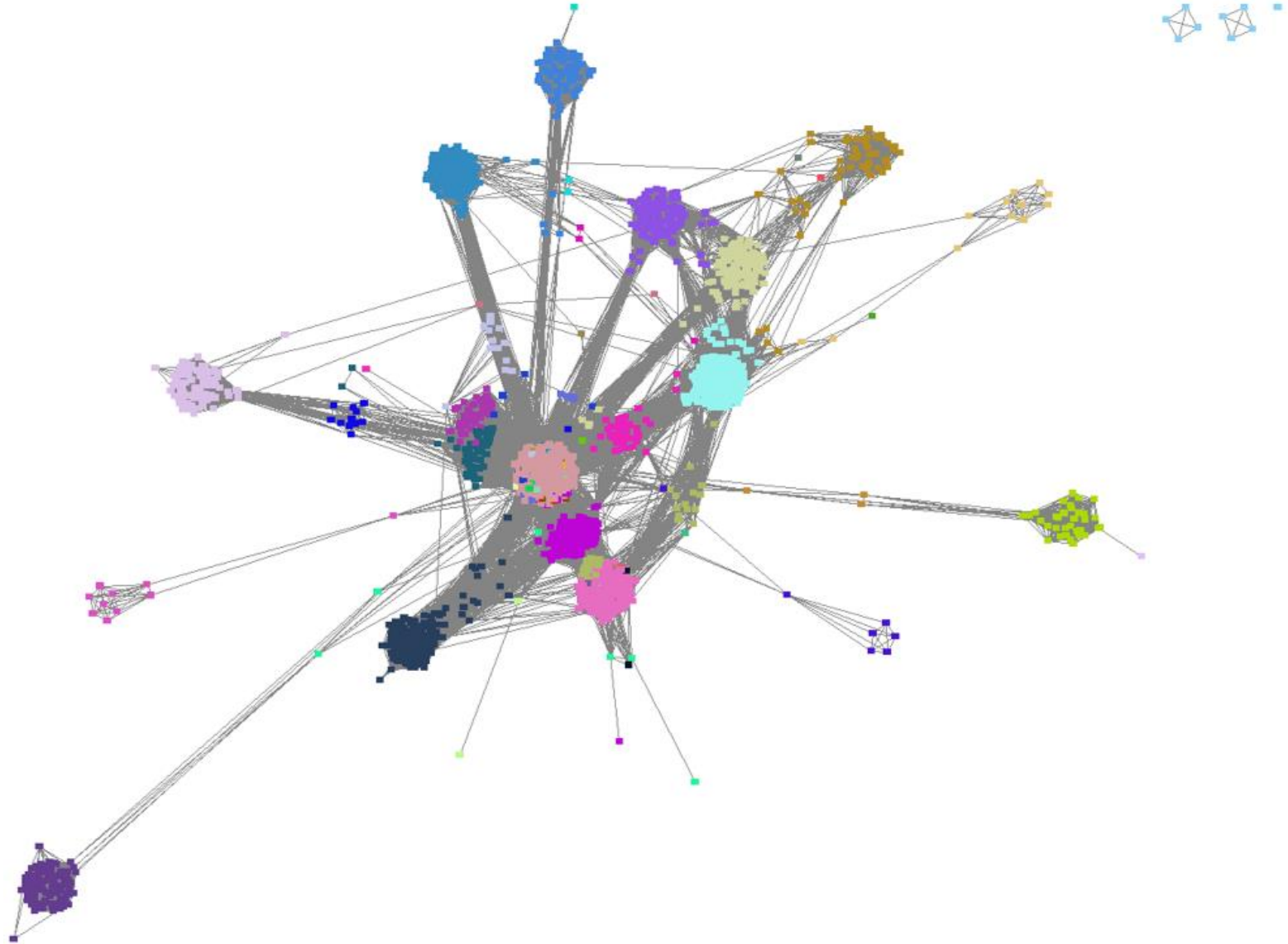
61.1 . . . 61.K

62.1 . . . 62.K

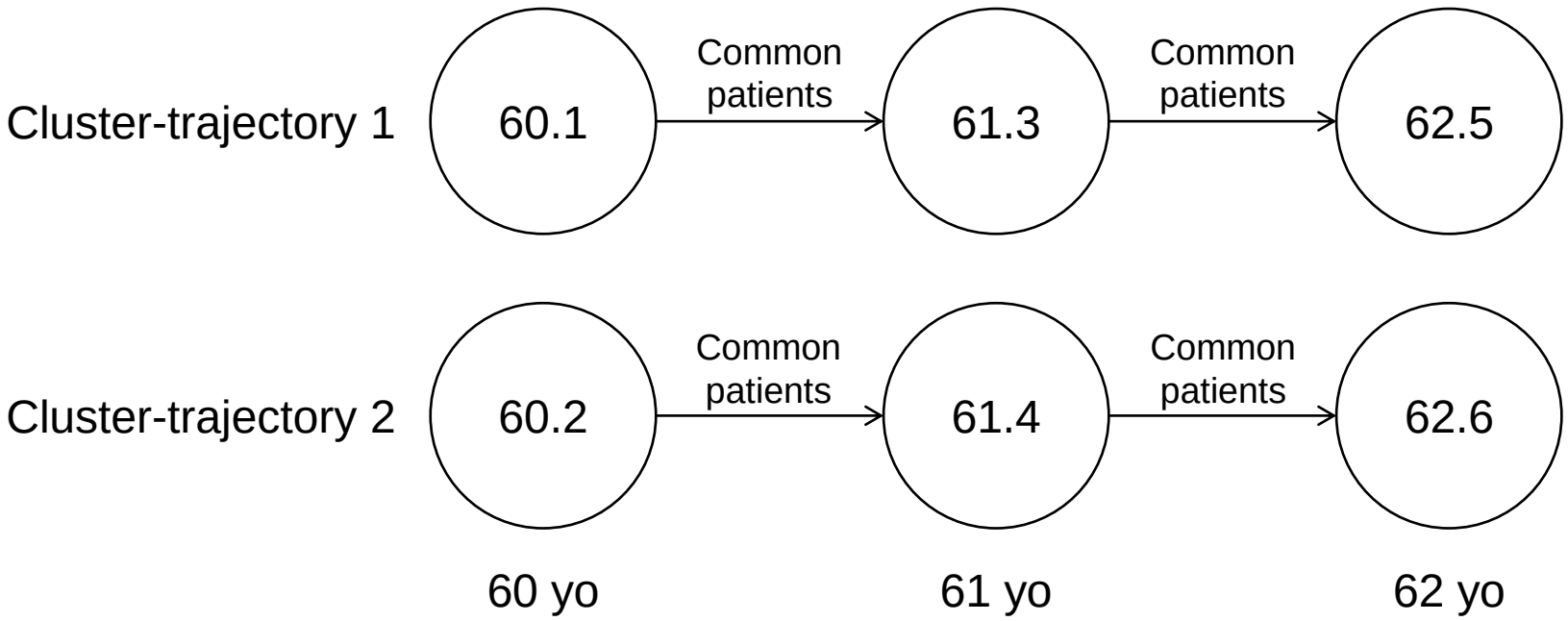
[1] Amit Singhal et al. (2001)

[2] vanDongen, Stijn. (2000)

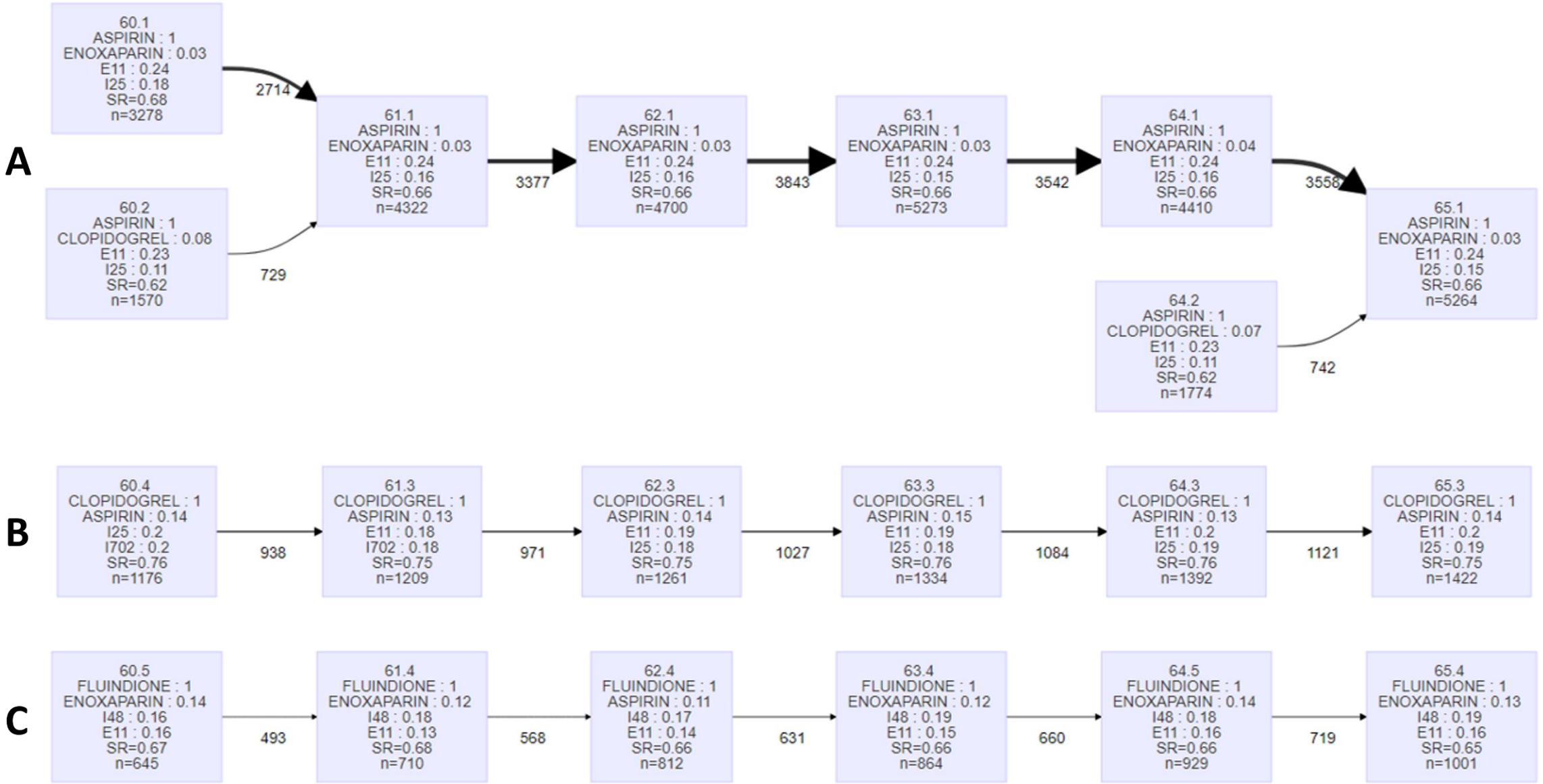
Cluster-tracking: network-based approach



Cluster-tracking: raw-data and network-based approaches

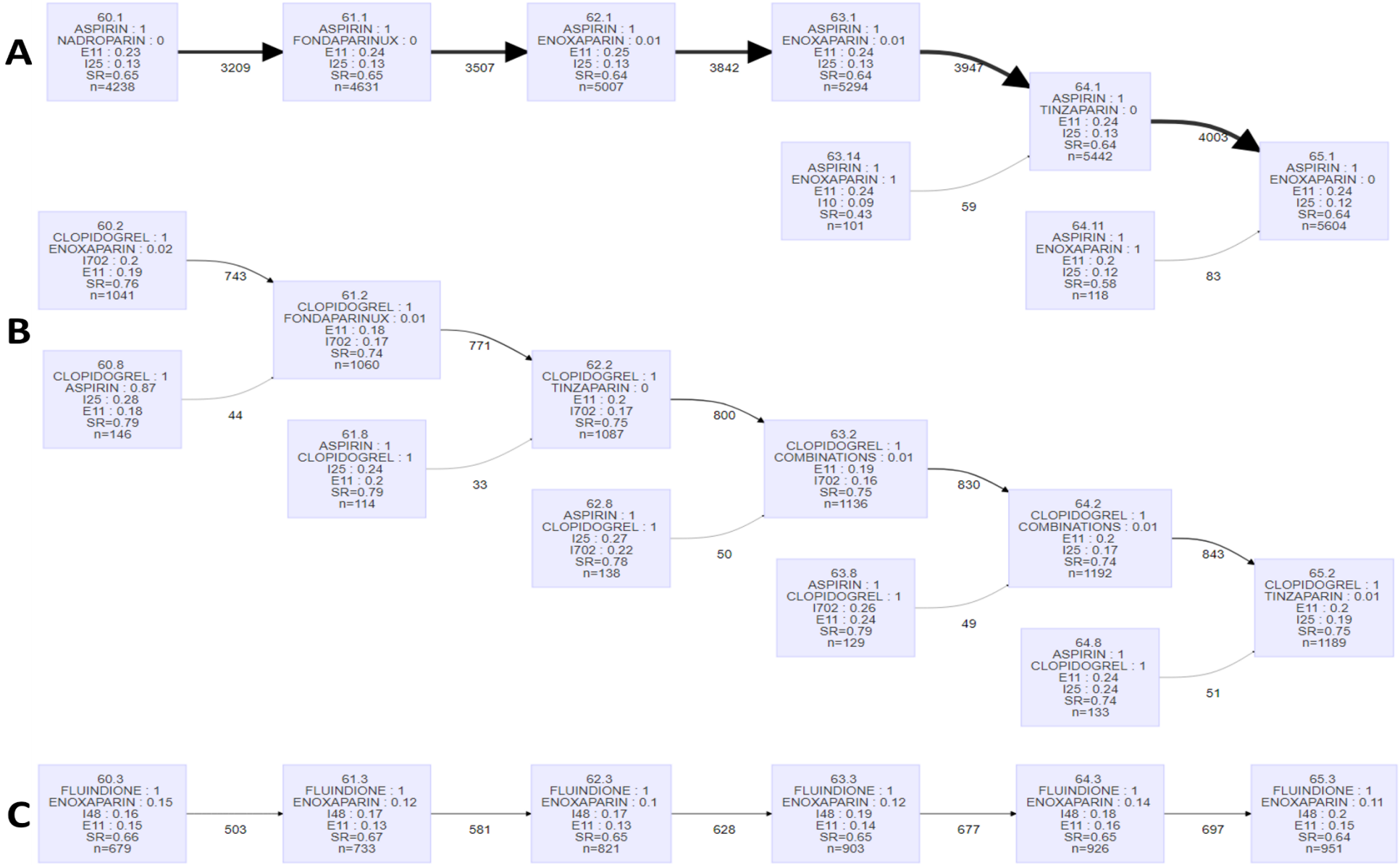


Trajectories identified with the raw-data-based approach



E11: type 2 diabetes mellitus, I25: chronic ischemic heart disease, I702: atherosclerosis of extremities, I48: atrial fibrillation

Trajectories identified with the network-based approach



E11: type 2 diabetes mellitus, I25: chronic ischemic heart disease, I10: essential primary hypertension, I702: atherosclerosis of extremities, I48: atrial fibrillation

Shiny app

On R:

```
library(shiny)

runGitHub("Cluster-tracking", "JudithLamb")
```

To visualize the tracking of clusters and the cluster-trajectories from a simulated dataset of 5594 patients with their drug prescriptions

Cluster-tracking approach from 60 to 70 years old

Clustering strategy

Network-based

Cluster limit size

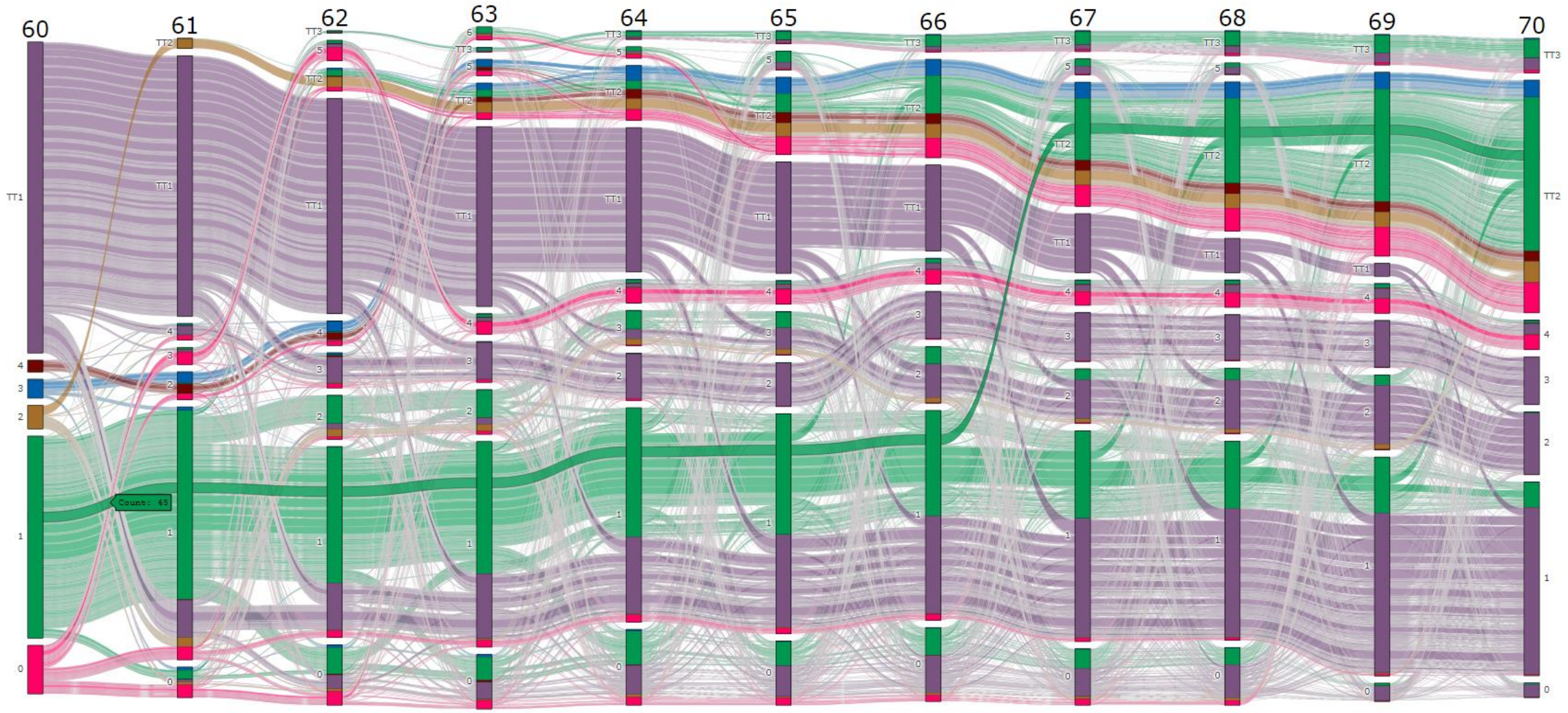
0 50 100

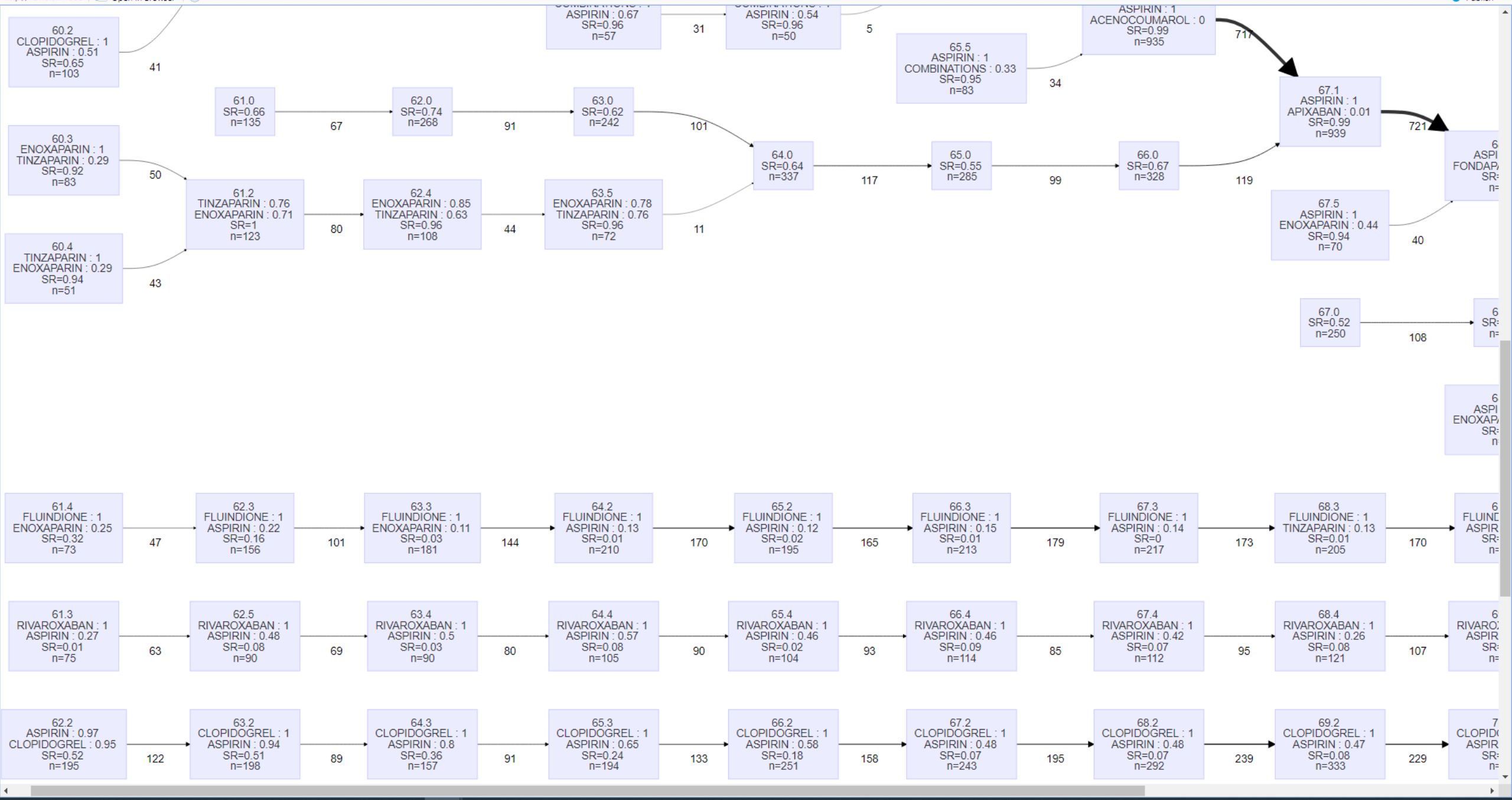
0 10 20 30 40 50 60 70 80 90 100

Display Clear

Alluvial Plot | Cluster Trajectories

Display Clear





Comparison with three classical longitudinal clustering approaches

Silhouette score: assess the clustering quality

$$S = \frac{1}{|I|} \sum_{i \in I} S^i$$

$S \approx 1$: well-separated clusters
$S \approx 0$: overlapping clusters
$S < 0$: misassigned clusters

$$S^i = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{p \in P^k} s_p^i$$

n_k: number of patients in the cluster k
K: number of clusters
i: age
s_p^i: silhouette score of patient p

Raw-data-based cluster-tracking	Network-based cluster-tracking	Raw-data-based longitudinal-clustering[1]	Feature-based longitudinal-clustering[1]	Model-based longitudinal-clustering[1]
0.57	0.50	0.27	0.20	-0.33

[1] Liao, T. Warren. "Clustering of time series data—a survey." (2005)

Conclusion

- We identified several cluster-trajectories corresponding to pathologies
- We identified more homogeneous groups of patients with our cluster-tracking approaches as compared to three classical longitudinal approaches
- Our cluster-tracking approaches do not need any imputation of data or exclusion of patients

Thank you for your attention !!!

Judith et al. (2022): https://www.medrxiv.org/content/10.1101/2022.08.05.22278468v1
Code Availability: https://github.com/JudithLamb/Cluster-tracking

judith.lambert@inserm.fr