Séminaire interne du SESSTIM

Segmentation de cohortes de patients par clustering de parcours de soins : une approche séquentielle univariée-multivariée

Nicolas Grevet*, Ndiaga Dieng, Nathalie Grafféo, Roch Giorgi

*nicolas.grevet@univ-amu.fr

SESSTIM, Faculté des Sciences Médicales et Paramédicales, Aix-Marseille Université, Marseille, France
https://sesstim.univ-amu.fr/

Introduction

- Analyser les parcours de soins pour
  1. décrire
  2. comprendre
  3. améliorer

  les prises en charge des patients et/ou l'organisation des soins.

- Données longitudinales issues des bases de données hospitalières, SNDS ou autres sources similaires

- Challenges
  - La nature séquentielle des données
  - La représentation des trajectoires de soins : états vs événements, **univariée vs multivariée**

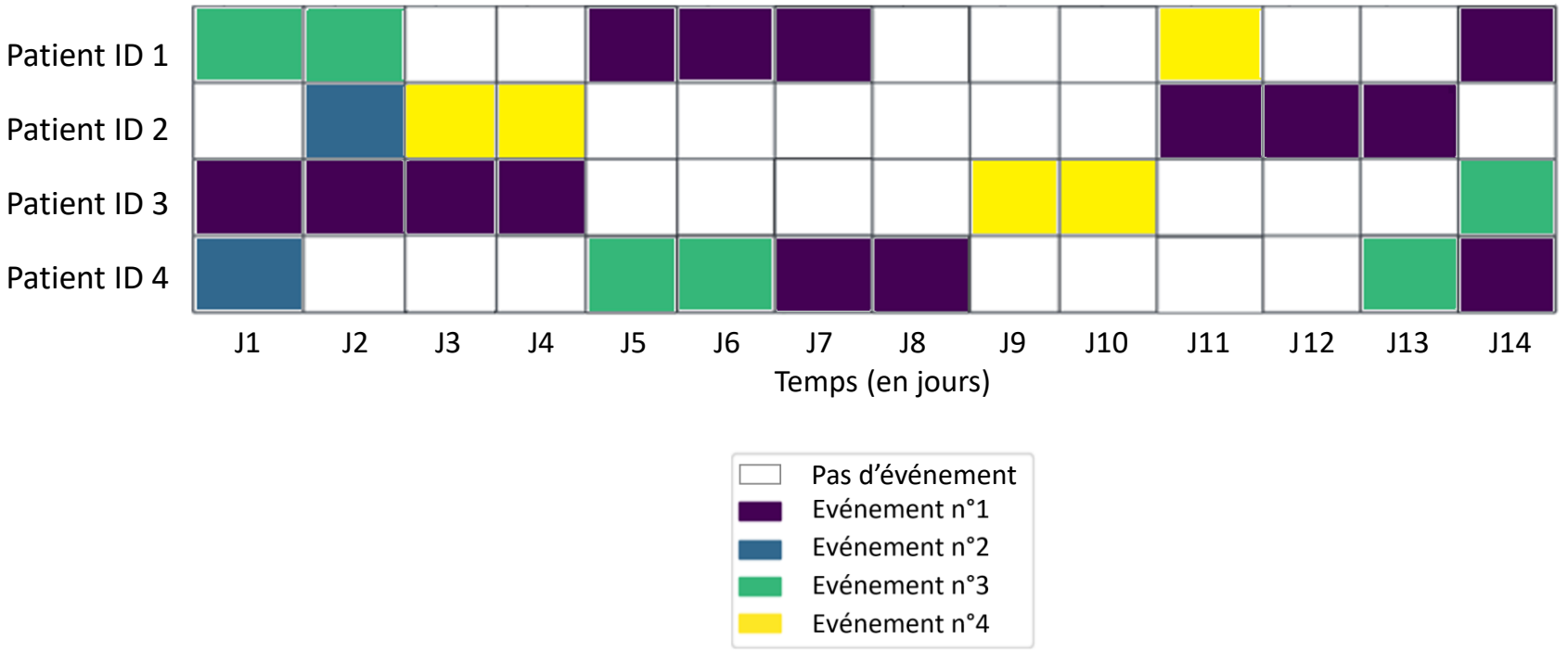
Objectifs

- Développer une approche combinant des techniques d'analyses univariées et multivariées pour regrouper efficacement les patients en fonction de leurs parcours de soins

- Implémenter l'approche dans un package Python : **Trajectory Clustering Analysis** (TCA)

Représentation des trajectoires de soins

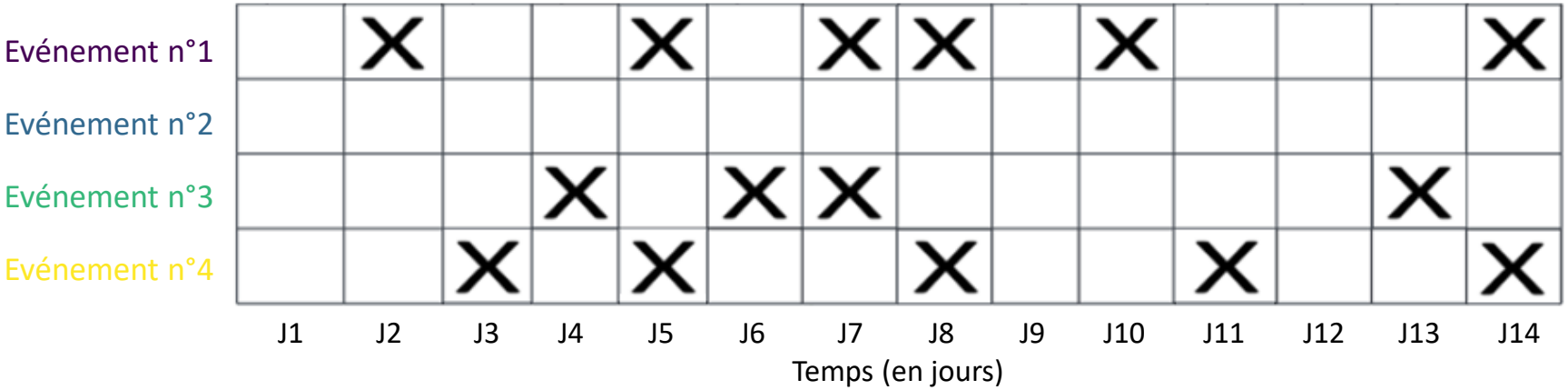
- Trajectoires **univariées**
  - Structure de données en **deux** dimensions : patients et temps
  - Nombreuses alternatives de mesures de similarité entre ces séquences



Représentation des trajectoires de soins

- Trajectoires **multivariées** :
  - Structure de données en **trois** dimensions : patients, événements de soins et temps

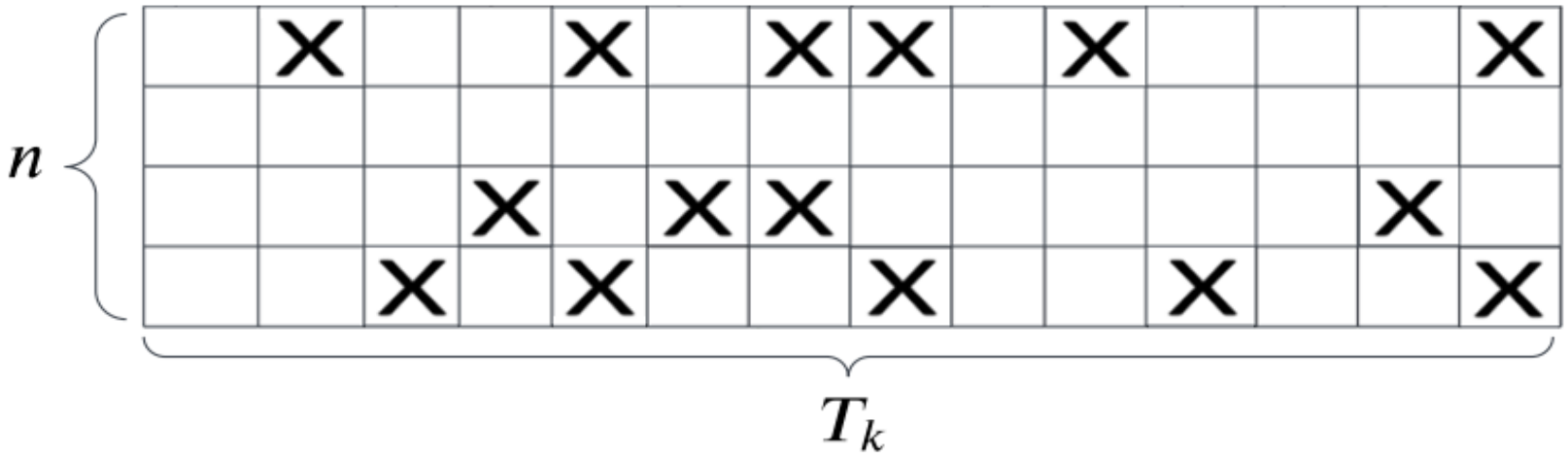
Pour 1 patient :



Représentation des trajectoires de soins

- Trajectoires **multivariées** :
  - Structure de données en **trois** dimensions : patients, événements de soins et temps

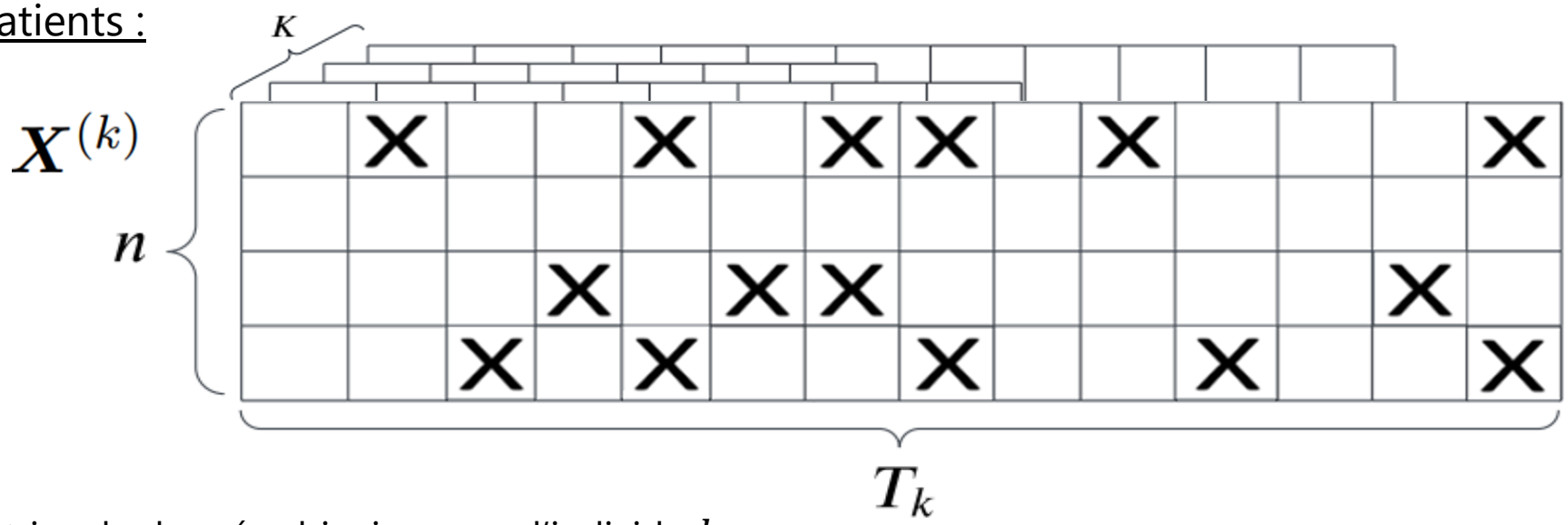
Pour 1 patient :



Représentation des trajectoires de soins

- Trajectoires **multivariées** :
  - Structure de données en **trois** dimensions : patients, événements de soins et temps

Pour K patients :

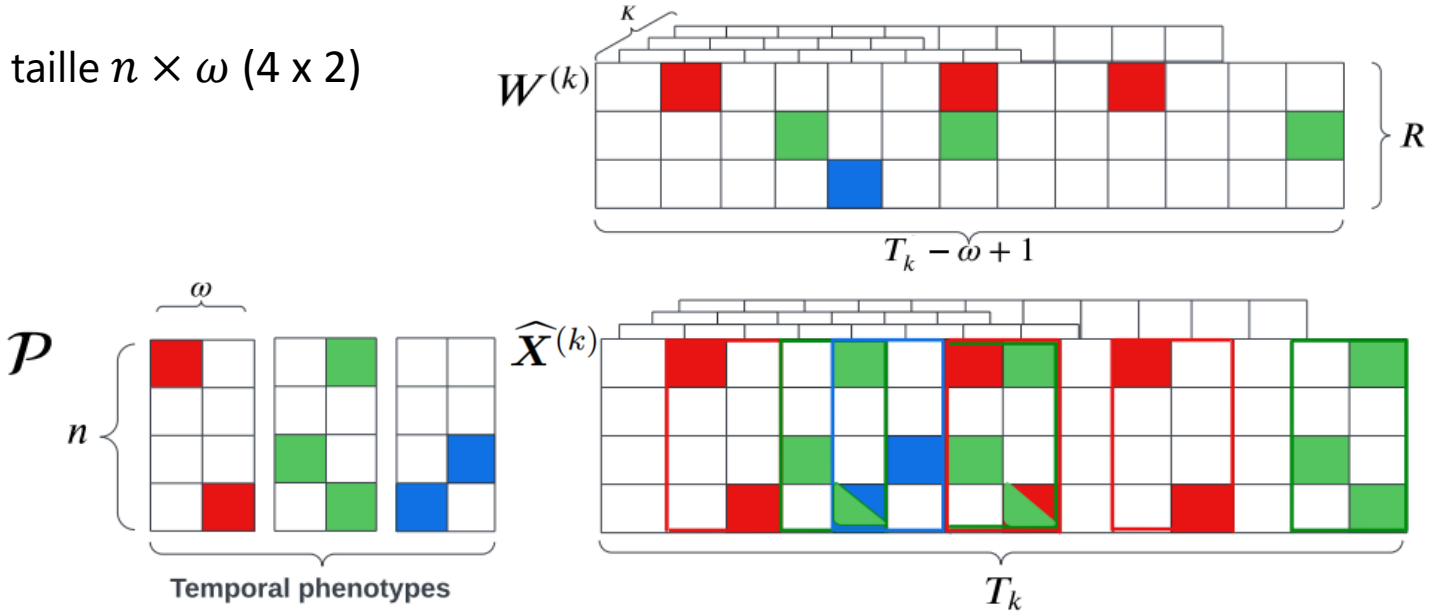


$\boldsymbol{X}^{(k)}$: matrice de données binaires pour l'individu k
K : nombre d'individus (patients)
n : nombre de catégories (événements)
T_k : durée des observations de l'individu k

Décomposition tensorielle

- **SWoTTeD** (Sliding Window for Temporal Tensor Decomposition)
  - Exploite la décomposition de tenseurs
  - Observations récurrentes découvertes $\mathcal{P}$: **phénotypes temporels**

R = 3 phénotypes de taille $n \times \omega$ (4 x 2)
T_k = 14
T'_k = T_k - 2 + 1 = 13



Sebia, H., Guyet, T. & Audureau, E.
SWoTTeD: an extension of tensor decomposition to temporal phenotyping.
Mach Learn 113, 5939–5980 (2024).

Application

- Données : **VICAN5** (2018) + données supplémentaires issues du dossier médical + informations sur la consommation de soins (extraites du SNIIRAM)

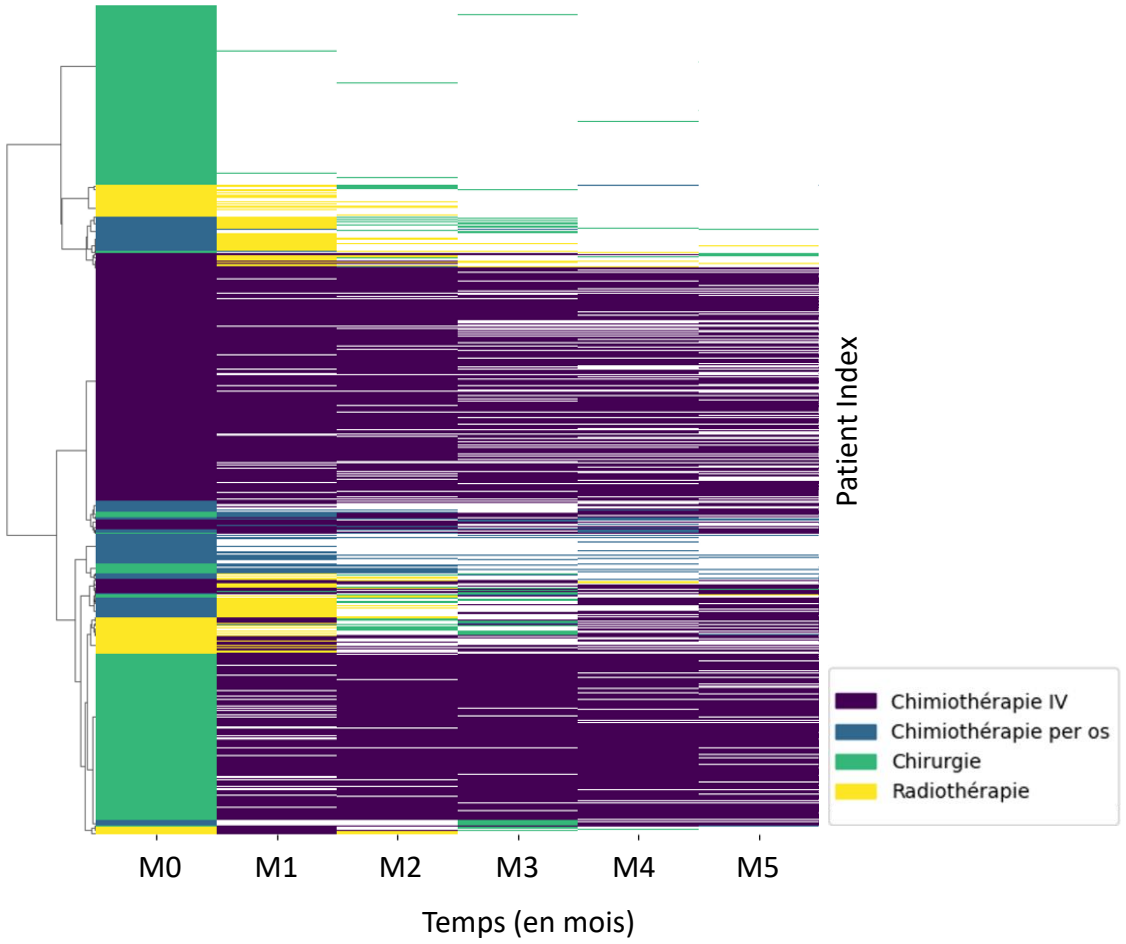
- Critère de sélection : première séquence d'événements thérapeutiques liée au **cancer colorectal**
  - Chirurgie
  - Chimiothérapie IV
  - Chimiothérapie per os
  - Radiothérapie

- Période de suivi : 2009-2016

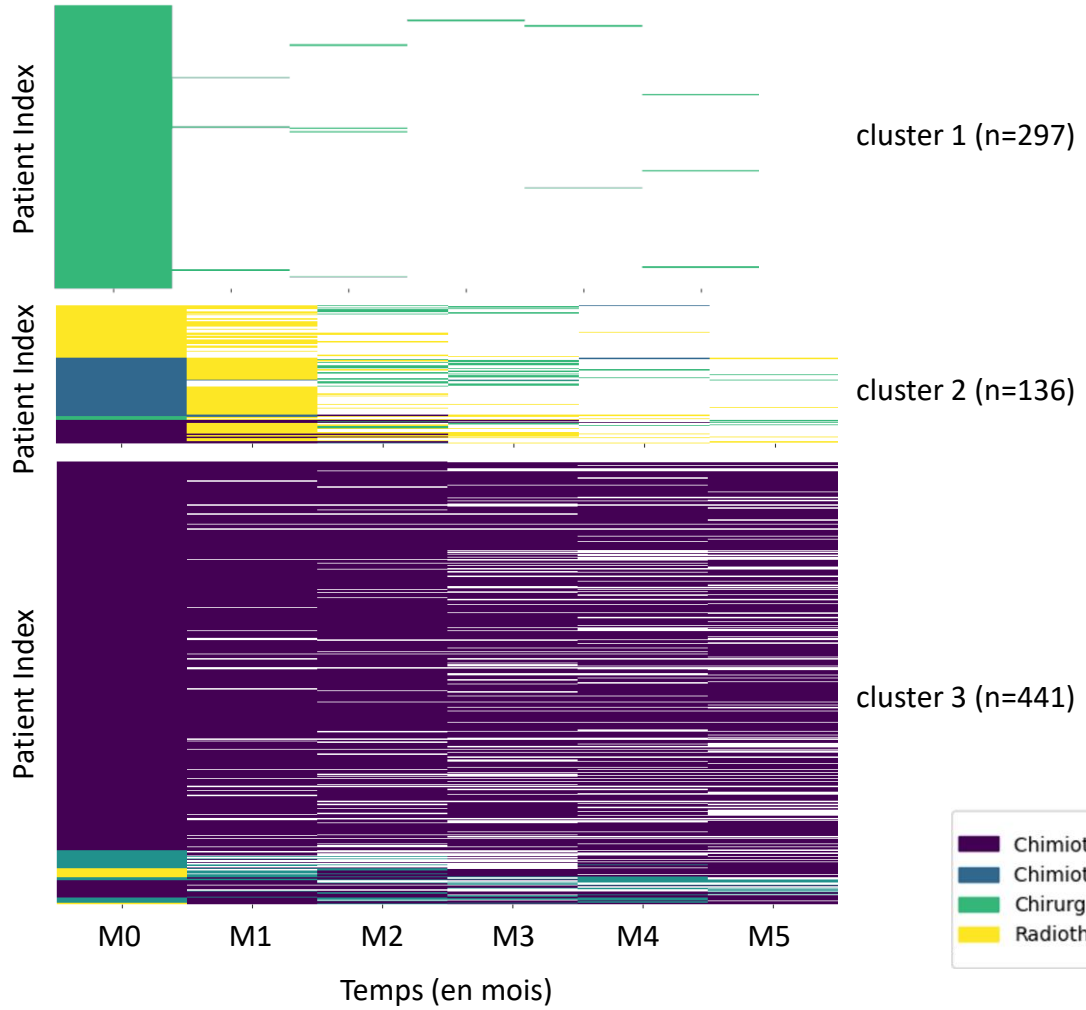
- 1371 patients ont reçu au moins un traitement (identification avec les codes CCAM, CIM-10 et ATC)

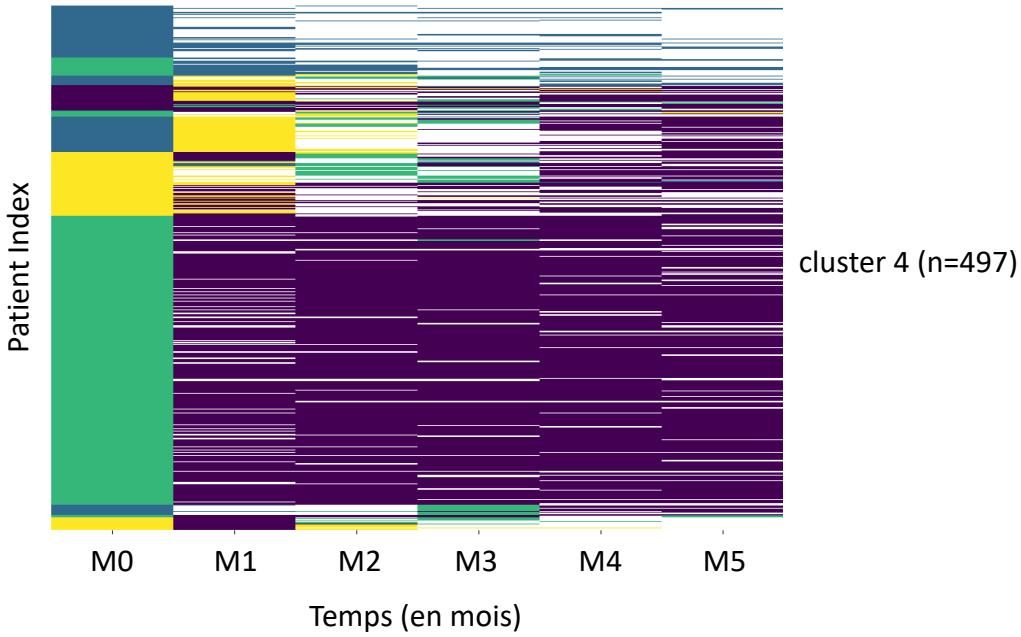
Résultats

- Analyse univariée
  - Métrique de similarité : Dynamic Time Warping
  - Algorithme de clustering : Classification Ascendante Hiérarchique



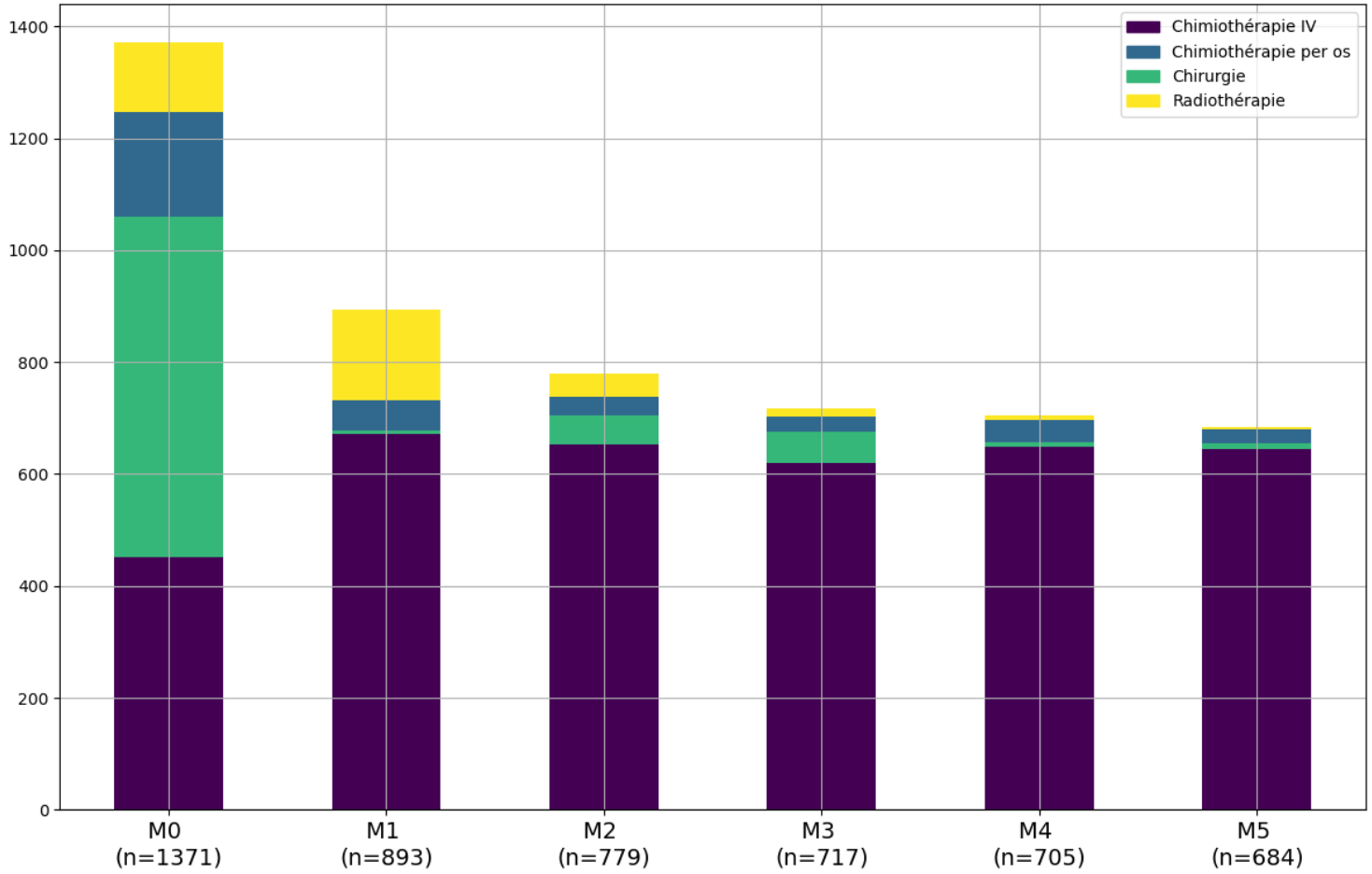
Résultats


Patient Index
Patient Index
Patient Index
cluster 1 (n=297)
cluster 2 (n=136)
cluster 3 (n=441)
M0
M1
M2
M3
M4
M5
Temps (en mois)
Chimiothérapie IV
Chimiothérapie per os
Chirurgie
Radiothérapie

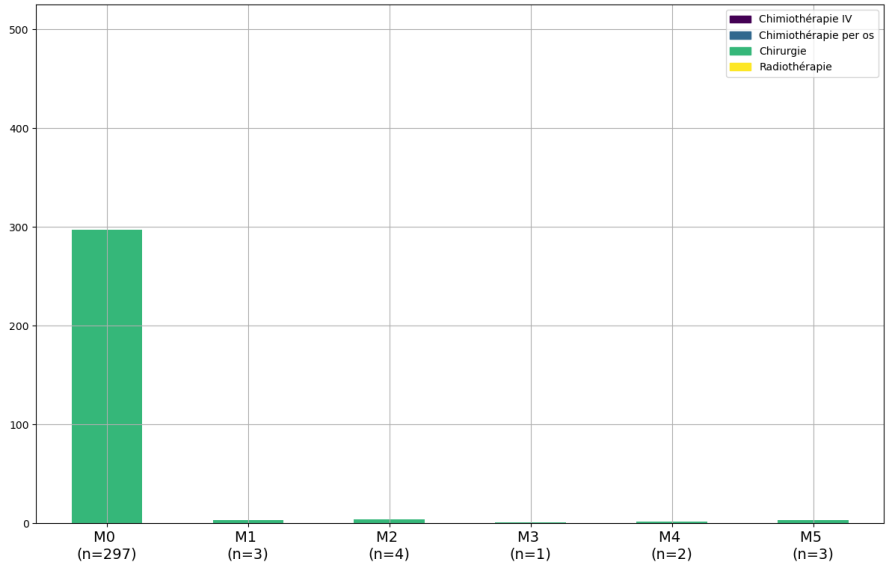

Patient Index
cluster 4 (n=497)
M0
M1
M2
M3
M4
M5
Temps (en mois)

Résultats

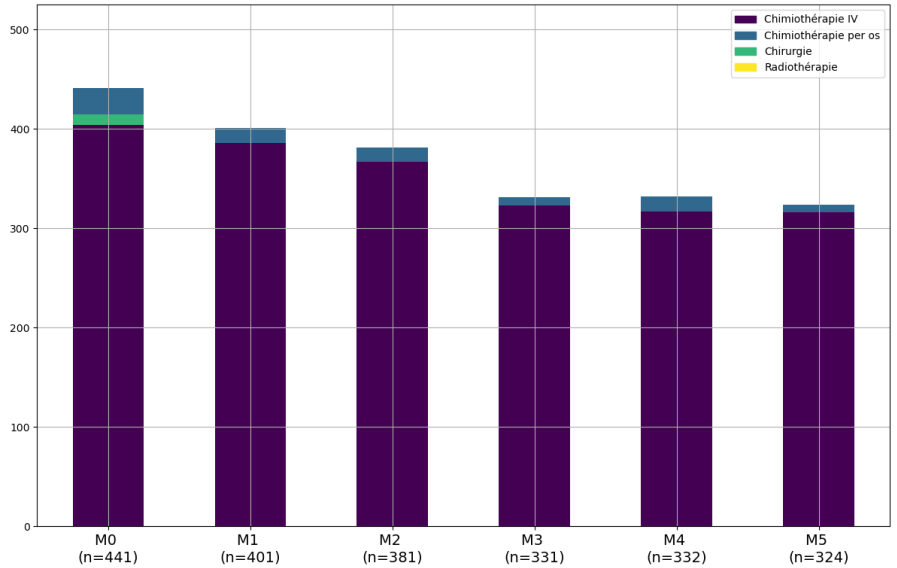
- Analyse univariée



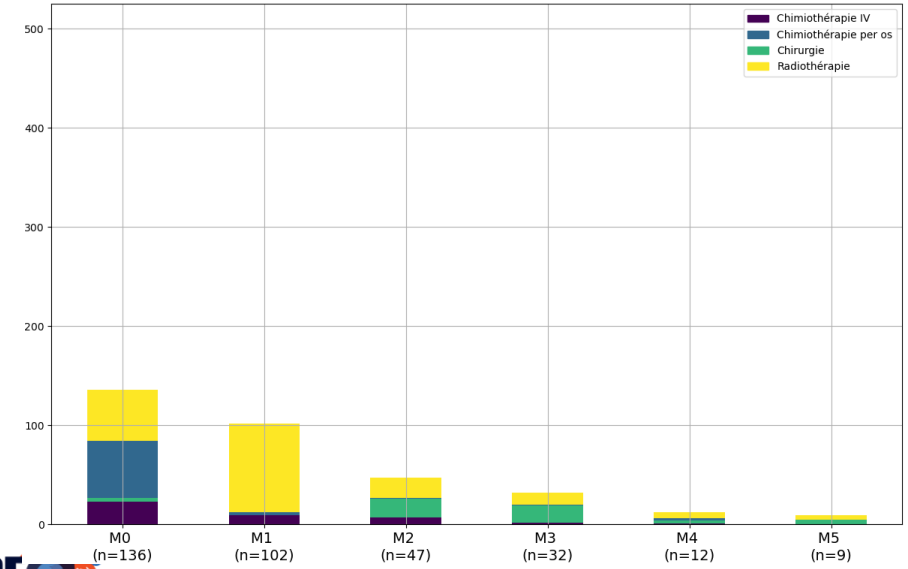
Résultats



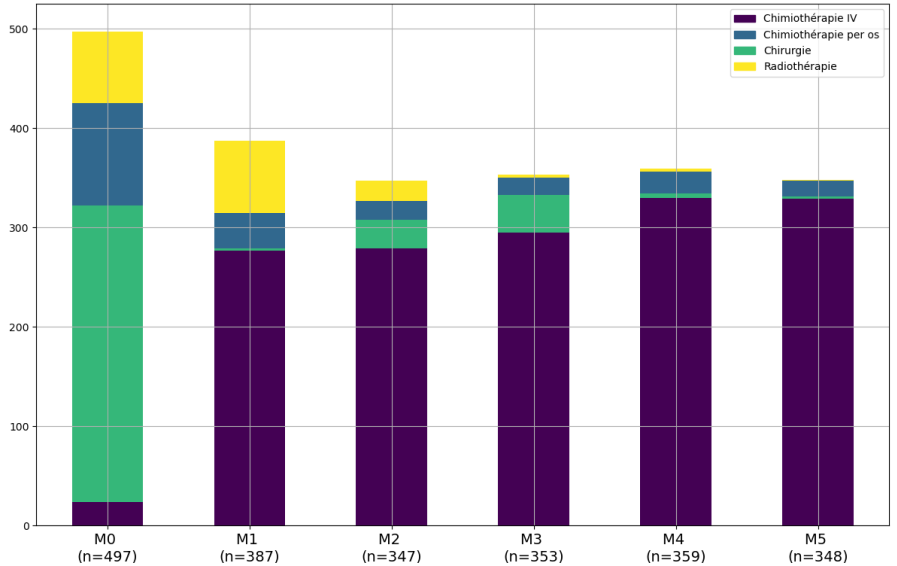
cluster 1 (n=297)



cluster 3 (n=441)



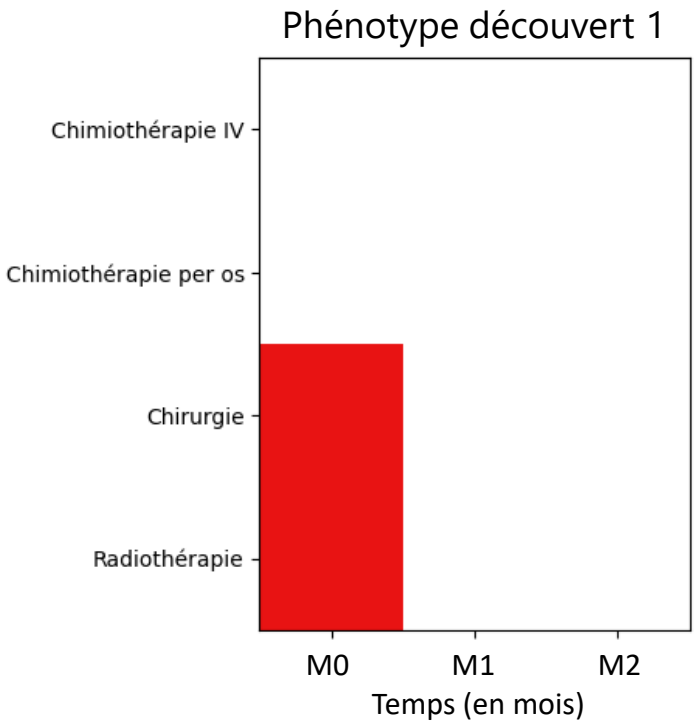
cluster 2 (n=136)

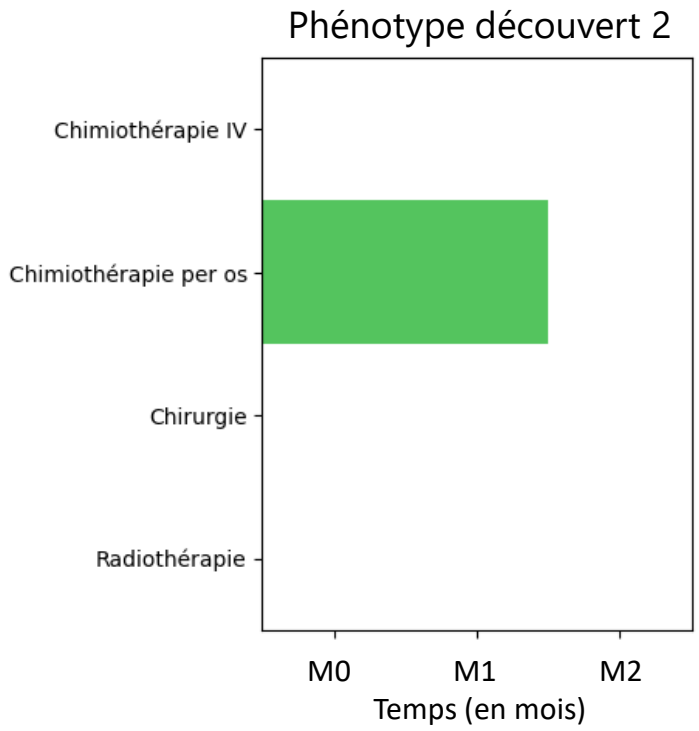


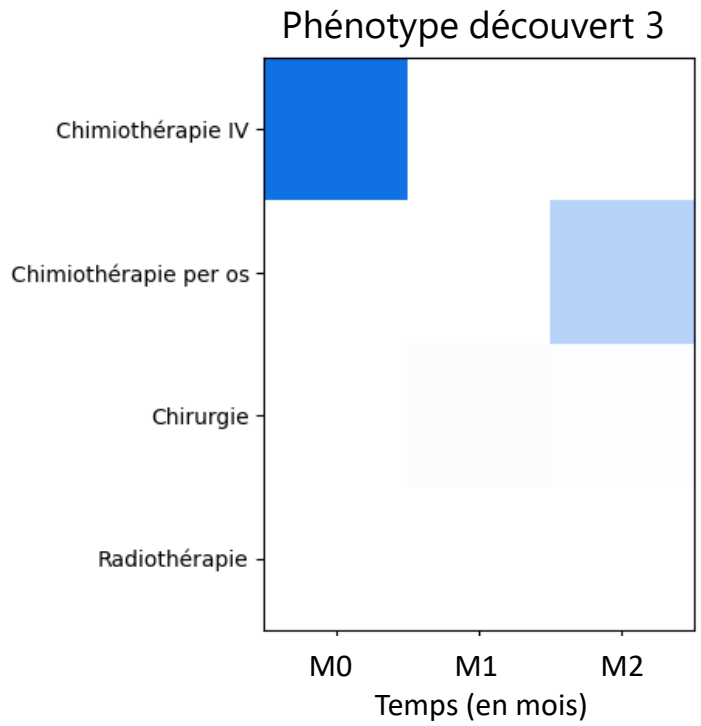
cluster 4 (n=497)

Résultats

- Analyse multivariée
  - Paramètres des phénotypes : nombre $(R) = 3$ de dimensions $(n \times \omega) = 4 \times 3$

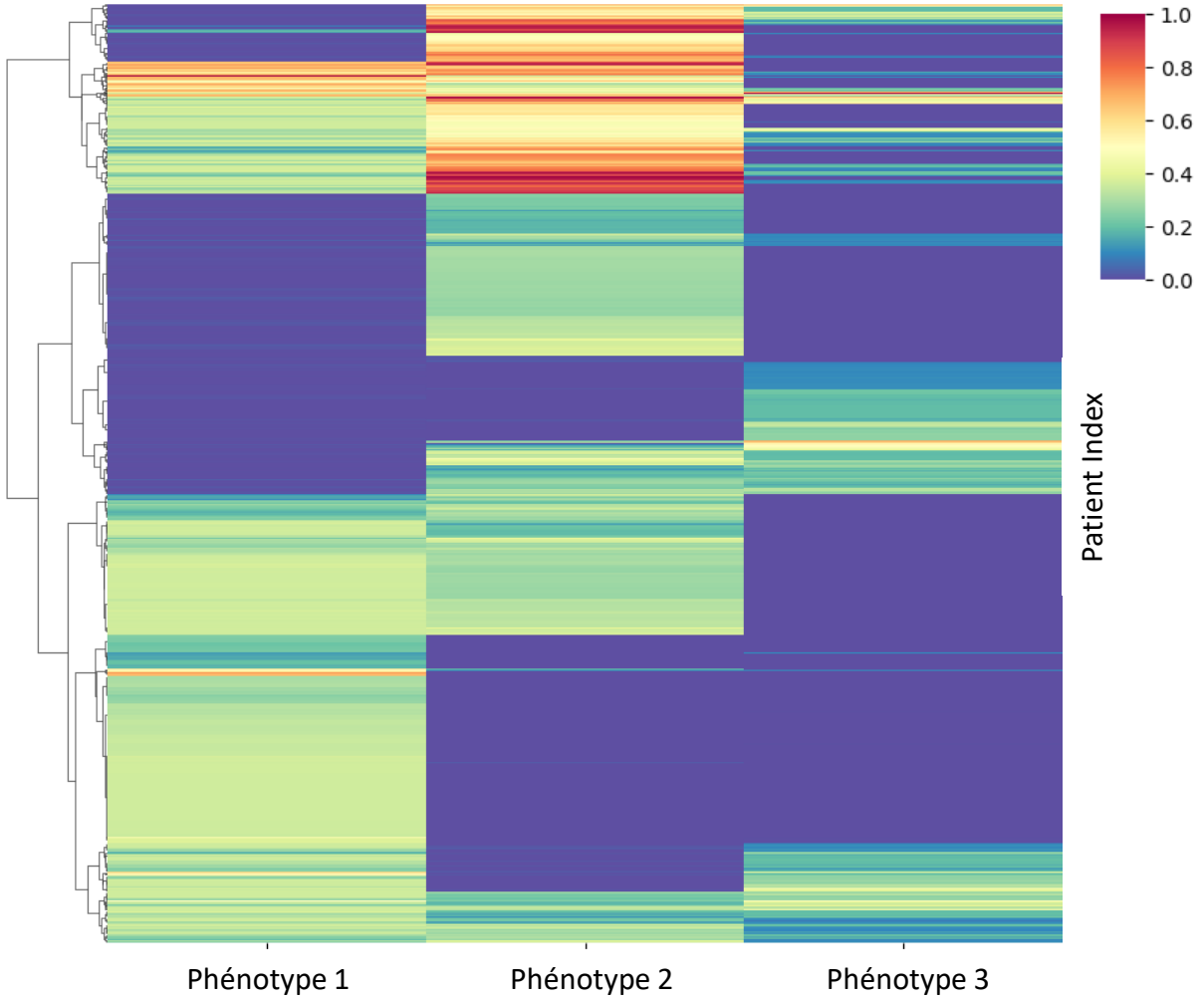






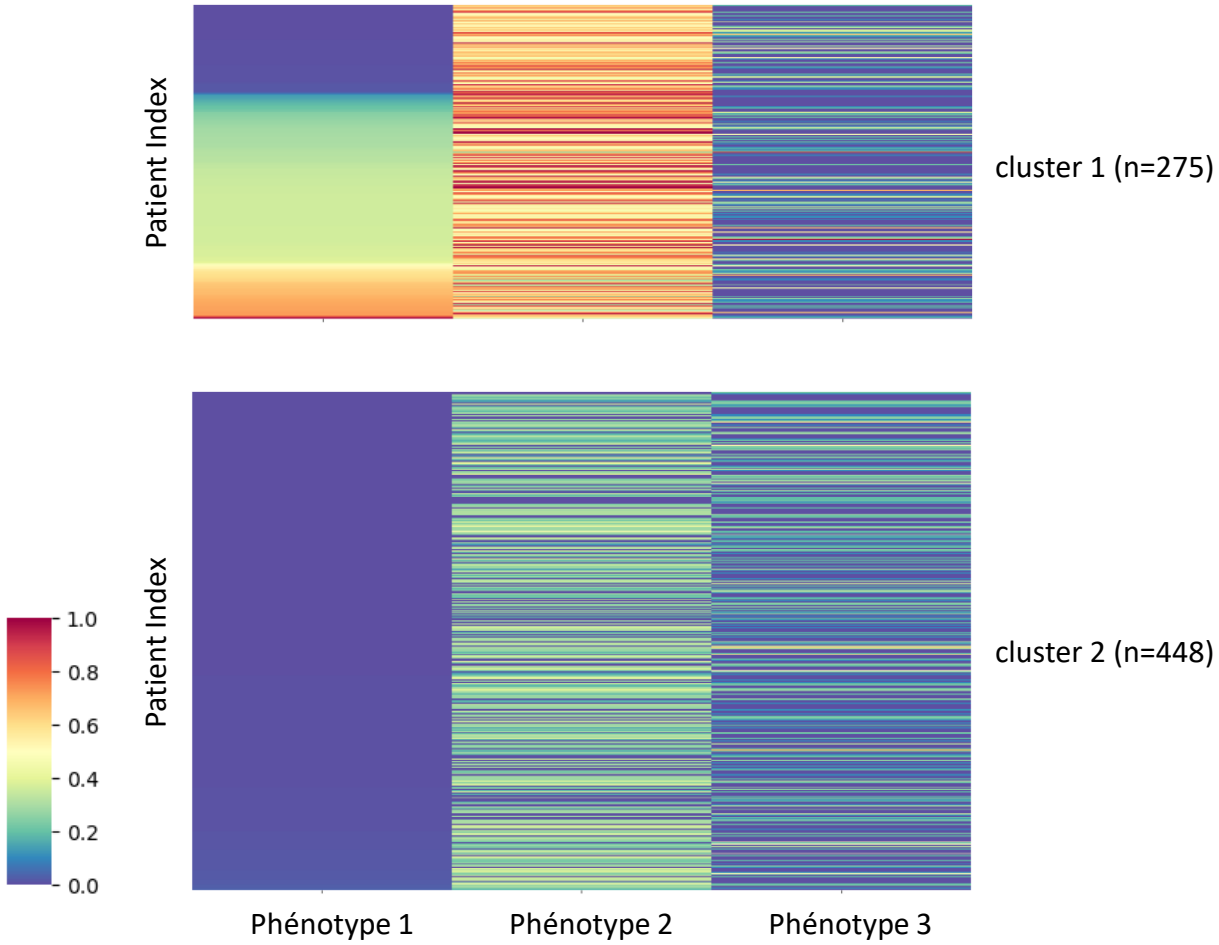
Résultats

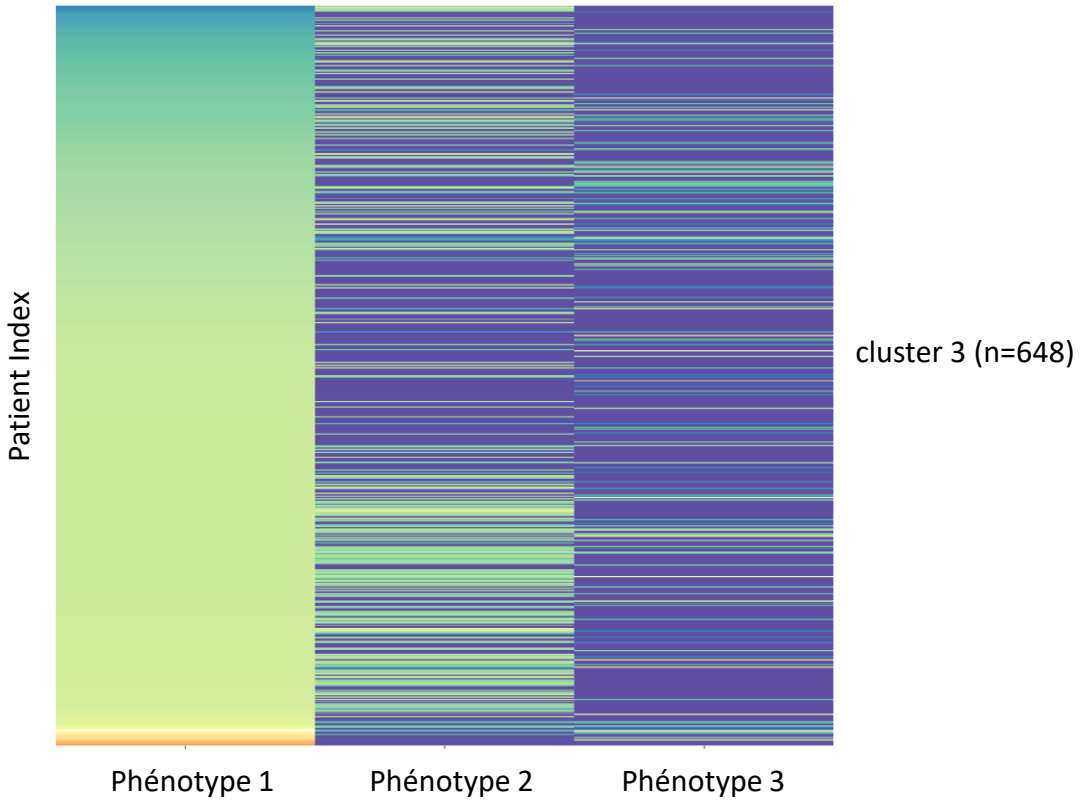
- Analyse multivariée



Résultats

- Analyse multivariée





Conclusion

- Les méthodes de représentation des parcours de soin combinées à une approche de clustering permettent d'identifier des patterns cliniquement pertinents.
- Fonctionnalités disponibles dans TCA : *github.com/ndiaga21/TrajectoryClusteringAnalysis*
  - Mesures de similarités : Euclidienne, Dynamic Time Warping, Levenshtein, Optimal Matching
  - Algorithmes de Clustering :
    - Classification Ascendante Hiérarchique
    - K-Means
    - Partition Around Medoids (en cours de développement)
  - Identification et interprétation des parcours caractéristiques de ces groupes (en cours de développement)

SWOTTED

- Problème d'optimisation non-supervisée avec données binaires
- Définition de la fonction *loss* de reconstruction :

$$\mathcal{L}^{\circledast}(\hat{\mathcal{X}}, \mathcal{X}) = \sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{i=1}^{n} \log(\hat{x}_{i,t}^{(k)} + 1) - x_{i,t}^{(k)} \log(\hat{x}_{i,t}^{(k)}).$$

Hong et al. [2020]

$$\begin{aligned} \underset{\{\boldsymbol{W}^{(k)}\},\mathcal{P}}{\arg\min} \quad & \mathcal{L}^{\circledast}\left(\mathcal{P} \circledast \mathcal{W}, \mathcal{X}\right) + \alpha \, ||\mathcal{P}||_1 + \beta \sum_{k=1}^{K} \mathcal{S}\left(\boldsymbol{W}^{(k)}\right), \\ \text{subject to} \quad & \forall k, \, 0 \leqslant \boldsymbol{W}^{(k)} \leqslant 1, \quad 0 \leqslant \mathcal{P} \leqslant 1. \end{aligned}$$

$$FIT_X = 1 - \frac{\sum_{k=1}^{K} ||\boldsymbol{X}^{(k)} - \widehat{\boldsymbol{X}}^{(k)}||_F}{\sum_{k=1}^{K} ||\boldsymbol{X}^{(k)}||_F}$$